Applied Mathematics and Computational Intelligence

In Press, Accepted Manuscript – Note to users

60-year Research History of Missing Data: A Bibliometric Review on Scopus Database (1960-2019)

Farah Adibah Adnan[1*], Mohd Hafiz Zakaria[1] and Safwati Ibrahim[1]

[1]Universiti Malaysia Perlis

ABSTRACT

Research on missing data was initiated in 1960 and the study on this topic grew exponentially across various subject areas since then. Therefore, this study aims to analyze those studies, specifically journal articles published in the context of missing data. Scopus database and analysis tools were utilized to retrieve all available journal articles related to missing data and its data. Next, due to the large number of articles found in the Scopus database, its information can only be efficiently extracted and combined using Mendeley software. To further obtained insights on the extracted information, VOSviewer was used to obtain network visualization and overlay visualization on authors' keyword and citation metrics was obtained using Harzing Publish or Perish software. Additionally, the growth of publication, languages used, subject area, country involved, and publication activity were also presented using bibliometric analysis. In total, 6227 journal articles were found. The record shows that a drastic increment of research in missing data happened in 2016, with 446 publications compared to 361 in 2015. Most of the research was affiliated with researchers in the United States, and was written mainly in English. In terms of subject area, they were mostly appeared in mathematics, decision sciences, medicine, and computer science. It is expected that the publications on this topic will be more critical in 2020 due to its research trend that is currently blooming in the area of medicine, and therefore lead potential directions for future research.

Keywords: Bibliometric, Missing Value, Imputation, Regression

1. INTRODUCTION

In conducting researches related to the data collection process, the issue of missing data is unavoidable. Normally we may find that the needed information is not available or missing due to many reasons. For instance, improper data entry, network error, machine breakdown, database system problems and many more. The chances of an observational research to encounter with this situation is almost certain and need to be dealt wisely.

Researchers globally realize deals with missing data need to be more effective. However, typically, missing data was ignored and cases that have some missing data in variables included in the analysis were simply deleted. This rule is not suitable for every domain. Hence, careful study should be considered if the missing data contain some important information and really represent the target population. Various methods have been used in dealing with missing data for example, expectation maximization, Gaussian mixture model (GMM), hot deck (HD), linear/logistic regression, least squares, principal component analysis (PCA), multiple imputation and lots more [1, 2, 3, 4, 5]. Failure to properly deal with missing data in analyses may lead to bias in the research outcome[6].

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

*farahadiba@unimap.edu.my

This study focused on analyzing scientific literature published from 1960 to 2019 on missing data using bibliometric analysis. Bibliometrics investigates the formal properties of knowledge domains by using mathematical and statistical methods. Specifically, it is defined as a statistical evaluation of published source type and is an effectual way to measure the influence of publication toward the research community [7]. Hence, by using this analysis, this study identifies articles that conducted studies related to the context of missing data from various countries and covered by various subject area. Surprisingly, to the best of our knowledge, there is no research on missing data using bibliometric analysis.

2. METHODS

This study collected all information from the Scopus database started on 16 December 2019. Scopus is the largest abstract and citation database of peer-reviewed literature [8] and the largest searchable citation and abstract source of searching literature [9].

The collected data was later refined several times as to obtain the best data that is closely related to the context of missing data. This refinement was made using the search strategy and the data retrieval process shown in Table 1. For this study, the focused was on journal article related to context of missing data, which were based on title or author keyword. Ultimately, the following query was conducted: (TITLE ("missing data" OR "missing value" OR "missing values") OR AUTHKEY ("missing data" OR "missing value" OR "missing values")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (SRCTYPE , "j")). This query produced 6227 documents.

Table 1 Search strategy and data process

Search refinement stage	Date	Database	Search String	Result
First stage	16/12/2019	SCOPUS	Title of "missing data", "missing value" or "missing values", and Keyword "missing data", "missing value" or "missing values", and is limited to journal article.	1778 articles
Second stage	17/12/2019	SCOPUS	Title of "missing data", "missing value" or "missing values", or Keyword "missing data", "missing value" or "missing values", and is limited to journal article.	6850 articles
Final stage	18/12/2019	SCOPUS	Title of "missing data", "missing value" or "missing values", or Author keyword "missing data", "missing value" or "missing values", and is limited to journal article.	6227 articles

3. RESULTS AND DISCUSSION

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

The data collected were analyzed to ascertain publication's per year, languages, subject areas, author keywords, geographical distribution, publication activity, and citations. Most of the results presented in frequency and percentage.

3.1 Publication by Year

The first research on missing data was published back in 1960 by G. N. Wilkinson comparing missing value procedures [10]. The progress of the related publications, then increased steadily till 1992. After that, it was exponentially increased. It is anticipated that the number of publications in 2020 will increase even more compared to previous years. It is because based on Figure 1, it can be seen that a publication kick-start was made. Although the year 2020 is still yet to come, some publications have already been indexed in the Scopus database.

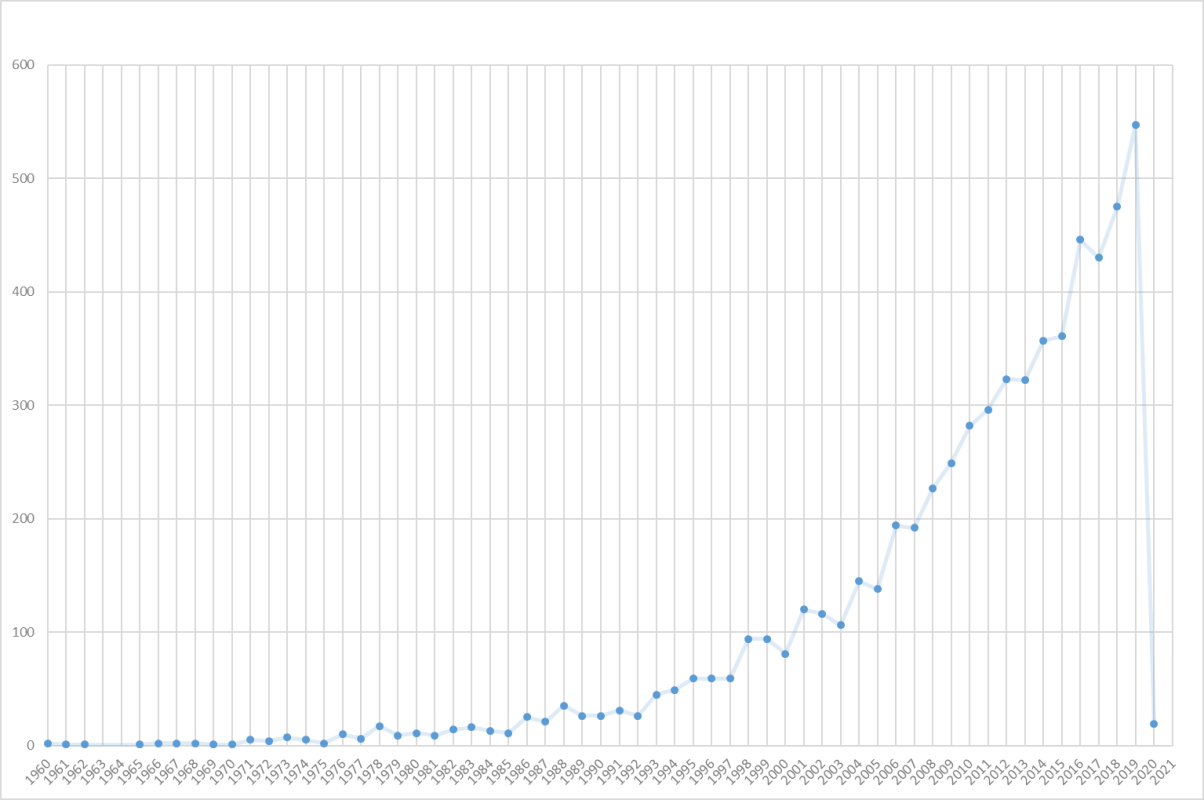


Figure 1. Publication by year.

3.2 Languages of Articles

Table 2 shows that most of the retrieved journals were published in English (96.37%). Out of 6227 journals, 26 journals were published in dual languages (0.4%) and hence it resulting 6253 papers in total. Among all, Croatian, Czech, Greek, Italian, and Korean were the least languages used in the papers studied, with one publication each.

Table 2 Languages

Language	No of Publication	%

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

English	6026	96.37
Chinese	123	1.97
Spanish	25	0.40
German	16	0.26
Japanese	16	0.26
Portuguese	13	0.21
French	12	0.19
Russian	10	0.16
Persian	3	0.05
Dutch	2	0.03
Turkish	2	0.03
Croatian	1	0.02
Czech	1	0.02
Greek	1	0.02
Italian	1	0.02
Korean	1	0.02
Total	6253	100

3.3 Subject area

This study also listed the published journals based on its subject areas. Most of the studies on missing data were in the area of Mathematics, Decision Sciences, Medicine, and Computer Science represents the top 50% percentile of the total articles. The additional subject areas covered in the missing data study are presented in Table 3.

Table 3 Subject area

Subject area	No of Publication	%	Percentile
Mathematics	2967	24.84	100
Decision Sciences	1359	11.38	75
Medicine	1329	11.13	64
Computer Science	1256	10.51	53
Engineering	740	6.19	42
Social Sciences	609	5.10	36
Agricultural and Biological Sciences	596	4.99	31
Biochemistry, Genetics and Molecular Biology	578	4.84	26
Psychology	362	3.03	21
Environmental Science	303	2.54	18
Earth and Planetary Sciences	220	1.84	15
Economics, Econometrics and Finance	209	1.75	14
Pharmacology, Toxicology and Pharmaceutics	206	1.72	12
Immunology and Microbiology	157	1.31	10
Business, Management and Accounting	150	1.26	9
Physics and Astronomy	130	1.09	8
Chemistry	124	1.04	6

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

Health Professions	113	0.95	5
Neuroscience	112	0.94	5
Arts and Humanities	105	0.88	4
Chemical Engineering	81	0.68	3
Nursing	76	0.64	2
Materials Science	64	0.54	1.37
Multidisciplinary	54	0.45	0.84
Energy	35	0.29	0.39
Veterinary	8	0.07	0.09
Dentistry	3	0.03	0.03
Total	11946	100	

3.4 Keyword analysis

The author keywords were diagrammed with VOSviewer, a software tool for constructing and visualizing bibliometric networks. Figure 2 and Figure 3 presents a network visualization and overlay visualization respectively.

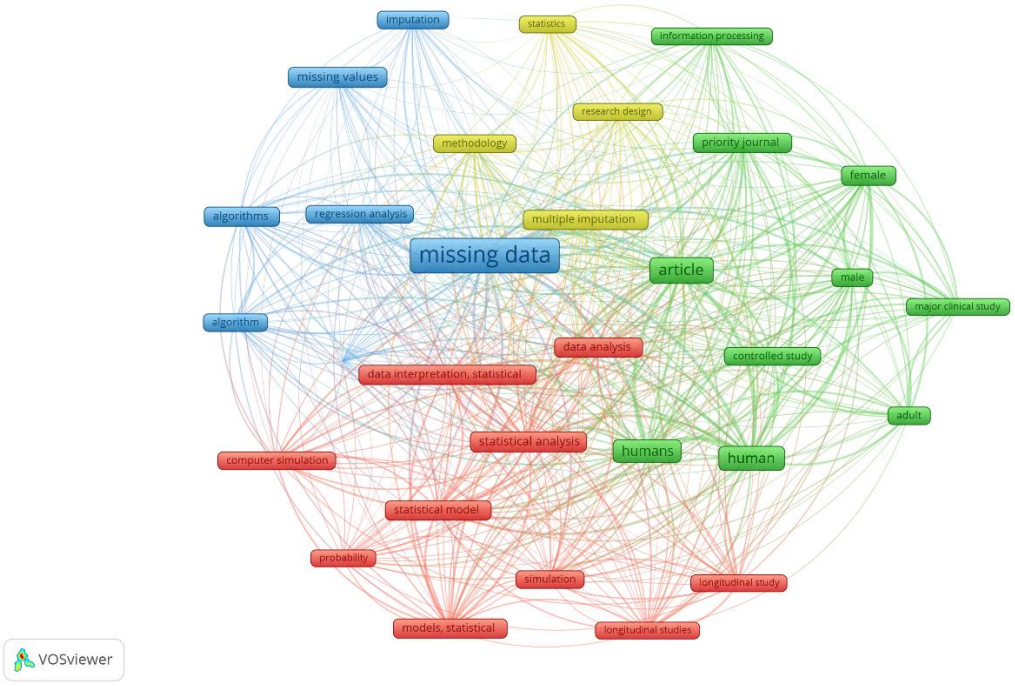


Figure 2. Network visualization.

In Figure 2, keywords with the same colour were usually listed together. So, in this study, for example, missing data, regression analysis, algorithms, missing values, and amputation were coded with the same blue colour, suggesting that these keywords have a correlation and mentioned together.

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

In Press, Accepted Manuscript – Note to users

Even so, this visualization can still be questioned as multiple imputation keyword; which was among a popular keyword used in the context of missing data, were coded with different colours (yellow), suggesting that it is not a common word that co-occur with missing data keyword. Therefore, Figure 3 (overlay visualization) was referred as to making sense of the suggestion.

Figure 3 was meant to show the time series of keywords occurrence. Based on its legend, the keywords were classified into the four series of occurrence, which started in 2009 and below, and ended at 2012 and above. As it can be seen, missing data keyword occurred mostly in 2010 to 2011, as for multiple imputation, it was commonly used as a keyword for missing data researches in 2012 independently. Due to that, VOSviewer suggested that these two keywords did not have a close relationship.

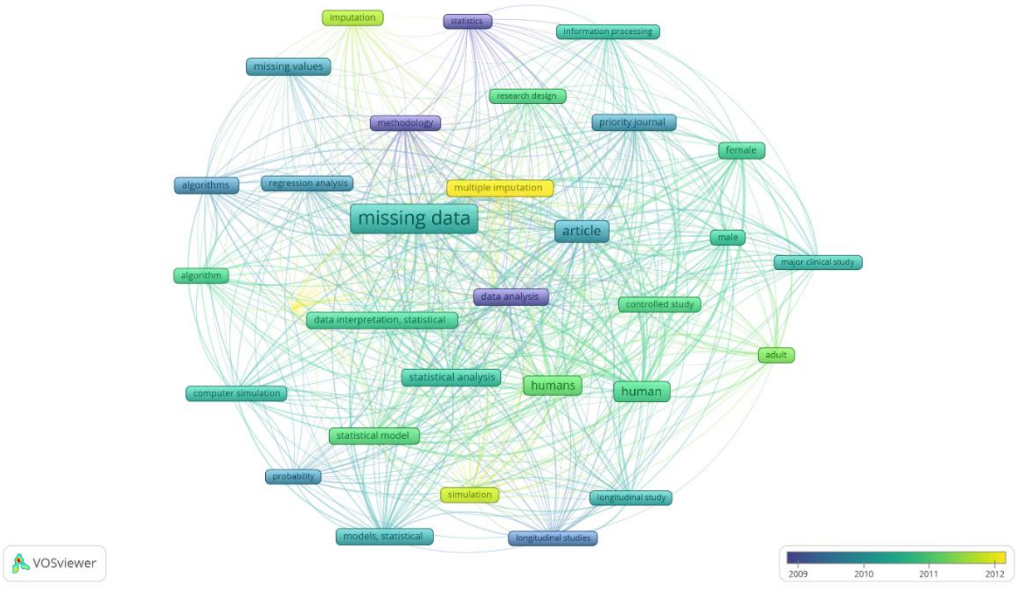


Figure 3. Overlay visualization.

Based on the number of occurrences, the top 20 keywords used in the missing data research are shown in Table 4.

Table 4 Top 20 keywords used by the authors

KEYWORD	Frequency	%
Missing Data	3534	21.53
Article	1680	10.23
Human	1424	8.67
Humans	1207	7.35
Statistical Analysis	712	4.34
Missing Values	665	4.05
Statistical Model	659	4.01
Multiple Imputation	601	3.66
Data Interpretation, Statistical	580	3.53
Female	560	3.41

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

Priority Journal	545	3.32
Models, Statistical	539	3.28
Algorithms	533	3.25
Data Analysis	500	3.05
Male	483	2.94
Computer Simulation	470	2.86
Methodology	456	2.78
Algorithm	435	2.65
Regression Analysis	423	2.58
Controlled Study	411	2.50

3.5 Geographical Distribution

In general, 106 countries involved in the publication of missing data articles in the Scopus database. The top 20 countries are listed in Table 5. United States was tiered first with 2683 documents, followed by the United Kingdom (669), and China (651).

Table 5 The top 20 countries involved in the publication of the missing data article

Country	Frequency	%
United States	2683	39.28
United Kingdom	669	9.80
China	651	9.53
Canada	370	5.42
Australia	247	3.62
Germany	245	3.59
France	227	3.32
Netherlands	225	3.29
Spain	207	3.03
Belgium	176	2.58
India	175	2.56
Italy	164	2.40
Japan	146	2.14
South Korea	124	1.82
Sweden	96	1.41
Taiwan	96	1.41
Hong Kong	88	1.29
Brazil	83	1.22
Switzerland	80	1.17
Finland	78	1.14

3.6 Publishing Activity by Journal

The 6227 articles appeared in 159 journals across 27 different subject areas. Table 6 lists the journals with the most articles on missing data. The leading journals are the Statistics in Medicine, followed by the Journal of the American Statistical Association and then the Biometrics. The

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

In Press, Accepted Manuscript – Note to users

context of the missing data issue belongs to the initial screening process of every data analysis, which generally tapped into the areas of interest of most journals, regardless of their level of quality (in terms of Scimago journal rank).

3.7 Citation Analysis

In conducting citation analysis, Harzing's Publish or Perish software was used to analysed the citation metrics for the extracted data. However, due to the large amount of article filtered from Scopus, data cannot be retrieved directly. Data extraction in RIS format was done several times as at a time, only the data of 2000 articles can be extracted. This process resulted in several RIS files. After that, all of the files were re-combined using Mendeley software.

Table 6 Journals with the most articles on missing data

Source title	Quartile	Publisher	Number of Articles
Statistics In Medicine	1	John Wiley & Sons Inc.	241
Journal Of The American Statistical Association	1	Taylor & Francis	156
Biometrics	1	Blackwell Publishing Inc.	138
Communications In Statistics Theory And Methods	3	Marcel Dekker Inc.	115
Biometrika	1	Oxford University Press	92
Computational Statistics And Data Analysis	1	Elsevier BV	88
Statistical Methods In Medical Research	1	SAGE Publications	65
Biometrical Journal	1	John Wiley & Sons Ltd.	61
Journal Of Biopharmaceutical Statistics	2	Marcel Dekker Inc.	58
Communications In Statistics Simulation And Computation	3	Dekker	51
Psychometrika	1	Springer New York LLC	50
Journal Of Statistical Computation And Simulation	2	Taylor & Francis	49
Statistics And Probability Letters	3	Elsevier BV	49
Journal Of Applied Statistics	3	Routledge	48
Journal Of Statistical Planning And Inference	1	Elsevier BV	47
Journal Of Multivariate Analysis	1	Elsevier Inc.	42
Journal Of The Royal Statistical Society Series C Applied Statistics	1	Blackwell Publishing Inc.	42
Statistica Sinica	1	Academia Sinica	40
Neurocomputing	1	Elsevier BV	38
American Journal Of Epidemiology	1	Oxford University Press	37

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

Next, the combined data has been imported into Harzing's Publish or Perish software to generate the citation metric. Table 7 summaries the citation metrics for the extracted journal articles as at 18 December 2019.

Table 7 Citation metrics

Reference date:	18/12/2019 12:54
Publication years:	1960 - 2020
Citation years:	59 (1960 - 2019)
Papers:	6227
Citations:	180793
Citations/year:	3064.29
Citations/paper:	29.03
Citations/author:	92597.35
Papers/author:	2778.31
Authors/paper:	3.04
Hirsch h-index:	168 (46.8% coverage)
Egghe g-index:	323 (57.7% coverage)
PoP hI, norm:	119
PoP hI, annual:	2.02

The top 20 cited articles in the context of missing data were listed in Table 8. An article titled "Missing data: Our view of the state of the art" by Schafer and Graham [7] obtained the highest number of citation in the Scopus database with 6491 citations (308.65 citations per year).

4. CONCLUSION

This paper presents a bibliometric review to gain a better intuition into the trends, review, contributions, and citation of the research on missing data. The study of this topic started early back in 1960 and increased steadily since then. Throughout the 60-year publication of article about missing data, a drastically increased publication has been witnessed several times, but the most drastic increment happened in 2016, with 446 publications compared to 361 in 2015. It is anticipated that the publications in the context of missing data will be published more in 2020 based on the observation that, as of December 2019, publications indexed for 2020 now stretch to 24 documents.

This study also reveals that one of the most covered subject areas in the missing data research is related to medical and health based on the journal with the most articles on missing data. It shows that this topic of research will become more and more critical in the future, as it will significantly correlate with human health. The variation of involved countries of the extracted data shows that the United States has dominated both publications and number of citations as compared to other developed countries such as the United Kingdom. Hence, we encourage that the research on missing data should be done in other developing countries as well, as in terms of missing data in a local setting, there is still a lot more that can be discovered.

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

Even so, few limitations of this study still need to be considered. It should be highlighted that even though Scopus is one of the largest academic paper databases, some still did not indexed under it. In addition, this study only concentrated on the topic related to missing data based on the title and the author keyword used in the articles. Moreover, the search was limited to journal article only, thus, all the other document and source type that related to missing data was filtered out. It is also important to note that the search query cannot be considered fully accurate in searching for all missing data journal articles. In this case, the search query can be further improved. The citation analysis obtained by this study through Publish and Perish, was based on the data extracted on 18 December 2019 at 12:54 pm. Despite all these limitations, this study is among the first to analyse the detailed bibliometric indicators of the published journal article in the context of missing data.

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

Table 8 Top 20 cited articles in missing data

No	Article title	Authors	Year	Source	Cited by	Cited / year
1	Missing data: Our view of the state of the art	Schafer, J.L., Graham, J.W. [11]	2002	Psychological Methods	6491	380.65
2	Inference and missing data	Rubin, D.B. [12]	1976	Biometrika	4436	102.86
3	Sampling-based approaches to calculating marginal densities	Gelfand, A.E., Smith, A.F.M. [13]	1990	Journal of the American Statistical Association	3872	133.38
4	A test of missing completely at random for multivariate data with missing values	Little, R.J.A. [14]	1988	Journal of the American Statistical Association	2622	84.03
5	Multiple imputation using chained equations: Issues and guidance for practice	White, I.R., Royston, P., Wood, A.M. [15]	2011	Statistics in Medicine	2585	320.38
6	Multiple Imputation after 18+ Years	Rubin, D.B. [16]	1996	Journal of the American Statistical Association	1896	82.26
7	The relative performance of full information maximum likelihood estimation for missing data in structural equation models	Enders, C.K., Bandalos, D.L. [17]	2001	Structural Equation Modeling	1894	104.78
8	Missing value estimation methods for DNA microarrays	Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B. [18]	2001	Bioinformatics	1851	102.67
9	Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering	Browning, S.R., Browning, B.L. [19]	2007	American Journal of Human Genetics	1317	109.25
10	Review: A gentle introduction to imputation of missing values	Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M. [20]	2006	Journal of Clinical Epidemiology	1138	87.23
11	How many imputations are really needed? Some practical clarifications of multiple imputation theory	Graham, J.W., Olchowski, A.E., Gilreath, T.D. [21]	2007	Prevention Science	1107	92.00
12	Estimation of regression coefficients when some regressors are not always observed	Robins, J.M., Rotnitzky, A., Zhao, L.P. [22]	1994	Journal of the American Statistical Association	1105	43.96
13	Modeling the drop-out mechanism in repeated-measures studies	Little, R.J.A. [23]	1995	Journal of the American Statistical Association	1014	42.13

UNIVERSITI MALAYSIA PERLIS

In Press, Accepted Manuscript – Note to users

No	Article title	Authors	Year	Source	Cited by	Cited / year
14	Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation	Stephens, M., Scheet, P. [24]	2005	American Journal of Human Genetics	983	70.07
15	Maximum likelihood estimation via the ecm algorithm: A general framework	Meng, X.L., Rubin, D.B. [25]	1993	Biometrika	932	35.81
16	Working with missing values	Acock, A.C. [26]	2005	Journal of Marriage and Family	899	64.16
17	Analysis of semiparametric regression models for repeated outcomes in the presence of missing data	Robins, J.M., Rotnitzky, A., Zhao, L.P. [27]	1995	Journal of the American Statistical Association	881	36.58
18	An approach to time series smoothing and forecasting using the em algorithm	Shumway, R.H., Stoffer, D.S. [28]	1982	Journal of Time Series Analysis	817	22.05
19	A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes	Pyron, R.A., Burbrink, F.T., Wiens, J.J. [29]	2013	BMC Evolutionary Biology	813	135.17
20	Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data	Yuan, K.-H., Bentler, P.M. [30]	2000	Sociological Methodology	771	40.47

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewer for their constructive comments and suggestions to improve this paper.

REFERENCES

[1] Ghorbani, S., Desmarais, M. C., Appl Artif Intell **31**, 1 (2017) 1–22.

[2] Kang, P., Neurocomputing **118,** (2013) 65–78.

[3] De Souto, M. C. P., Jaskowiak, P. A., Costa, I. G., Bioinformatics **16,** (2015) 64–72.

[4] Pati, S. K., Das A. K., Knowl Inf Syst **52,** 3(2017) 709–750.

[5] Valdiviezo, H.C., Van, A. S., Inf Sci **31,** 1(2015) 163–181.

[6] Pampaka, M., Hutcheson, G., Williams, J., International Journal of Research & Method in Education **39**, 1 (2016) 19-37.

[7] Madani, F., Weber, C., World Patent Information **46,** (2016) 32-48.

[8] Burnham, J. F., Biomed. Digit. Libr. **3**, 1 (2006) 1.

[9] Chadegani, A. A., Asian Soc. Sci. **9**, 5 (2013).

[10] Wilkinson, G. N., Australian Journal of Statistics 2**,** 2 (1960) 53-65.

[11] Schafer, J. L., Graham, J. W., Psychological Methods **7**, 2 (2002) 47-177.

[12] Rubin, D. B., Biometrika **63**, 3 (1976) 581-592.

[13] Gelfand, A. E., Smith, A. F. M., Journal of the American Statistical Association **85**, 410 (1990) 398-409.

[14] Little, R. J. A., Journal of the American Statistical Association **83**, 404 (1988) 1198-1202.

[15] White, I. R., Royston, P., Wood, A. M., Statistics in Medicine **30**, 4 (2011) 377-399.

[16] Rubin, D. B., Journal of the American Statistical Association **91**, 434 (1996) 473-489.

[17] Enders, C. K., Bandalos, D. L., Structural Equation Modeling **8**, 3 (2001) 430-457.

[18] Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R. B., Bioinformatics **17**, 6 (2001) 520-525.

[19] Browning, S. R., Browning, B. L., American Journal of Human Genetics, **81**, 5 (2007) 1084-1097.

[20] Donders, A. R. T., Heijden, G. J. M. G., Van, D., Stijnen, T., Moons, K. G. M., Journal of Clinical Epidemiology **59,** 10 (2006) 1087-1091.

[21] Graham, J. W., Olchowski, A. E., Gilreath, T. D., Prevention Science **8**, 3 (2007) 206-213.

[22] Robins, J. M., Rotnitzky, A., Zhao, L. P., Journal of the American Statistical Association **89**, 427 (1994).

[23] Little, R. J. A., Journal of the American Statistical Association **90**, 431 (1995) 1112-1121.

[24] Stephens, M., Scheet, P., American Journal of Human Genetics **76,** 3 (2005) 449-462.

[25] Meng, X. L., Rubin, D. B., Biometrika **80,** 2 (1993) 267-278.

[26] Acock, A. C., Journal of Marriage and Family **67**, 4 (2005) 1012-1028.

[27] Robins, J. M., Rotnitzky, A., Zhao, L. P., Journal of the American Statistical Association **90**, 429 (1995) 106-121.

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.

In Press, Accepted Manuscript – Note to users

[28] Shumway, R. H., Stoffer, D. S., Journal of Time Series Analysis **3,** 4 (1982) 253-264.

[29] Pyron, R. A., Burbrink, F. T., Wiens, J. J., BMC Evolutionary Biology **13,** 1 (2013).

[30] Yuan, K. H., & Bentler, P. M., Sociological Methodology **30,** 1 (2000) 165-200.

Note: Accepted manuscripts are articles that have been peer-reviewed and accepted for publication by the Editorial Board. These articles have not yet been copyedited and/or formatted in the journal house style.