

Identifying Missing Data Handling Methods with Text Mining

Krisztián Boros (✉ soirkorob@gmail.com)

Waseda University

Zoltán Kmetty

Centre for Social Sciences

Research Article

Keywords: missing data, imputation, fasttext, text mining

Posted Date: June 6th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3019294/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Identifying Missing Data Handling Methods with Text Mining

Krisztián Boros^{1,2*} and Zoltán Kmetty^{2,3}

^{1*}Graduate School of Economics, Waseda University, Totsumakachi 1-104, Shinjuku,
169-8050, Tokyo, Japan.

²Centre for Social Sciences, Hungarian Academy of Sciences, Tóth Kálmán str. 4,
Budapest, 1097, Pest, Hungary.

³Faculty of Social Sciences, Eötvös Loránd University, Pázmány Péter sétány 1/A,
Budapest, 1117, Pest, Hungary.

*Corresponding author(s). E-mail(s): soirkorob@gmail.com;
Contributing authors: kmetty.zoltan@tk.hu;

Abstract

Missing data is an inevitable aspect of every empirical research. Researchers developed several techniques to handle missing data to avoid information loss and biases. Over the past 50 years, these methods have become more and more efficient and also more complex. Building on previous review studies, this paper aims to analyze what kind of missing data handling methods are used among various scientific disciplines. For the analysis, we used nearly 50.000 scientific articles that were published between 1999 and 2016. JSTOR provided the data in text format. Furthermore, we utilized a text-mining approach to extract the necessary information from our corpus. Our results show that the usage of advanced missing data handling methods such as Multiple Imputation or Full Information Maximum Likelihood estimation is steadily growing in the examination period. Additionally, simpler methods, like listwise and pairwise deletion, are still in widespread use.

Keywords: missing data, imputation, fasttext, text mining

1 Introduction

Missing data is an immanent part of every empirical research. Every time a patient drops out of a clinical study or a respondent does not answer a question in a survey; we encounter missing data. Researchers have developed various techniques to account for these scenarios to reduce information loss. Such techniques spread from deletion of missing cases to complex algorithms that replace missing information with a predicted value ([1], [2], [3], [4], [5], [6]). During the past five decades, these techniques continuously evolved. Despite the

improvements, many research uses more conservative methods such as listwise deletion (complete case analysis) or mean imputation. Numerous types of research suggest that the more advanced missing data handling techniques, such as Multiple Imputation, are more flexible and reliable than the older and simpler ones ([6], [7]). Of course, there are scenarios where a simple deletion method could perform nearly as well as a more modern approach ([7]), but in general, it is recommended to use an advanced technique [2, p. 39]. Our goal in this study is to identify missing data handling methods in scientific papers during the period 1999-2016 in order to examine the

usage of advanced missing data handling methods. Throughout our research, we utilize a text-mining approach to extract the necessary information from the articles. The start of our examination period coincides with the publication of “Statistical Methods in Psychology Journals: guidelines and explanations” by [8]. This paper discusses the recommended methodologies for missing data in Psychology journals and highlights the importance of proper documentation of data analysis. As [2, p. 39] mentions, several studies support the insights of [8] on missing data handling practices and warn about the disadvantages of deletion methods. Besides Psychology, researchers in other fields preferred simpler techniques to account for missing data, especially listwise- and pairwise-deletion ([9], [10], [11]). Even though a comprehensive survey of missing data handling methods is yet to be made, there were a considerable amount of reviews about techniques for handling missing data. Most of the studies were conducted in the educational, psychological, and medical research areas, probably because of the discipline-specific origins of the missing data paradigm and its applications in survey-type designs ([12]). [13] examined randomly selected articles from two Psychology journals between 1989 and 1991 and concluded that the usage of deletion methods is pervasive in these journals. Later studies on various fields supported the results of [13] ([14]). [15] reviewed Political Science papers in a five-year period between 1993 and 1997 and found similar trends as [13]. [11] compared articles from 1999 and 2003 in Educational Research and, on one hand, concluded that the popularity of deletion methods did not change between 1999 and 2003. On the other hand, however, they noted that the reporting of missing data increased significantly: *“In 1999, 33.75% of the studies that we identified as having missing data explicitly reported the problem, whereas this number more than doubled, to 74.24%, in 2003.”* [11, p. 30]. Additionally, [16] also reviewed the practices of handling missing data in this field between 1998 and 2004. Their findings are consistent with the conclusions of [11]. In Psychology, [17] found similar trends concerning missing data methods between 2000-2006. In the Medical Research field between 2001 and 2002, [18] with the analysis of seven cancer journals, and [19] with the examination of four medical journals found similar results concerning listwise- and

pairwise-deletion. [9] reviewed articles from four medical journals in 2013 and provided a comparison of previous studies which examined missing data methods. They compared their results with the findings of [20], [21], [22], and [19]. The main conclusion was that the usage of deletion methods remained unchanged during the period 1997-2013, but there was a slight increase in the usage of advanced imputation methods. By and large, studies from Psychology, Educational Research, Medical Research show that the usage of deletion methods remained unchanged during 1989-2013, but the usage of advanced missing data handling methods increased slightly. This means that more and more research use an advanced technique when missing data occurs. All the aforementioned studies examined the randomly selected papers by manually reviewing and evaluating them. Our approach, however, is quite different: we train classification models to classify the articles on a large scale. This is a major methodological difference that has its own advantages and disadvantages. Probably the most advantageous aspects of this approach are time efficiency and scalability, since after making an adequate training set, we can use it to classify thousands of papers without even seeing them. This saves time and permits us to apply this approach to an arbitrarily large sample. This time-saving and scalable framework, however, leads to a major disadvantage which stems exactly from the fact that we do not see the majority of the articles. We can not tell whether our model truly found something or not — it is merely a matter of probability. Overall, our analysis could serve as an interdisciplinary overview of trends in the usage of missing data handling methods, and it may facilitate the application of text mining in future research in this area.

2 Data collection

Our data was provided by JSTOR’s Data for Research service (DfR) ([23], [24]) Since we received the data from JSTOR on 2020.07.20, the service (and platform) went through a few changes. This means that the way of our data request is no longer available, but this does not affect our data in any way. Currently, JSTOR offers a sophisticated text-mining platform named “Constellate” in order to help researchers create databases and perform basic text-mining tasks.

This platform was not available at the time of our research, therefore a simple keyword search might give different results now, mostly because of the extended data sources. The original data request and collection procedure was the following. At first, a search query had to be made with the required parameters to access the list of articles. This resulted in a search URL and a search syntax, which can be accessed in our [repository](#). Our search parameters included the time interval of publications (1999.01.01.–2016.12.31.), the keywords/expressions (“missing data”, “missing observations”, “incomplete data”, “imputation”), and language of the articles (English). In the case of keywords/expressions we had to specify which keywords/expressions should the corpus and title of the articles contain. Our request was that the corpus of articles must contain at least one of the following

keywords/expressions: “missing data”, “missing observations”, “incomplete data”, “imputation”; but the title of the articles must not contain any of the following keywords/expressions: “missing data”, “incomplete data”, “imputation”. With these specifications, we wanted to focus on papers that at least mention missing data related terms and to exclude those articles that are about missing data handling methods. Our initial goal was to examine only the usage of missing data handling methods, therefore those articles that did not even mention related terms were ineligible for our analysis. Later, during preprocessing, we make an extra step to ensure that only those papers are in our corpus that could have used some kind of missing data technique and not about missing data handling. As a result, we have collected 49.603 articles with metadata, uni-, bi-, tri-grams, and article content.

3 Methodology and data preparation

The goal of our research was to classify articles based on their use of missing data handling methods using advanced text-mining methods. As far as we know, there are no other research that tried this approach so far, therefore we had to create a feasible methodological framework to work with. In the following section we give an overview

about this process and highlight the most important aspects. In order to identify the usage of missing data handling methods we had to make a proper “sample” to work with. Although the data we have gathered from JSTOR is filtered by the predefined keywords, it does not guarantee that only those articles get into our analysis that are eligible for our research. We had to make sure that only those papers are present in our sample which use missing data handling methods, and are not about missing data handling methods. After selecting our sample, we performed several classifications to distinguish between various missing data handling techniques. Overall, our approach consisted of three main levels regarding classification. On the first level, we separated the papers according to their relation to missing data handling methods: if a given article was about missing data handling methods, then it was classified as “1” and was removed from the analysis. Otherwise, we kept the article in our corpus. The second level was destined to separate the usage of imputation methods from other techniques such as deletion, and from those cases where no technique was used. Accordingly, if an article used any type of imputation technique (multiple imputation, regression imputation, etc.), then it was classified as “1”, otherwise “0”. The third and the last level was divided into two parts. On one hand, we checked whether a paper — that was classified into the “imputation” category on the previous level — used an advanced imputation method or not. On the other hand, if a paper did not use any imputation technique — according to the second level —, then we checked if it has used any deletion method or not. Before heading towards the description of the preprocessing and classification, we must clarify what we mean by “imputation”, “advanced imputation”, and “deletion”. The importance of this clarification lies in the fact that we had to have a solid definition of each method to be able to label our training set correctly. We heavily relied on the taxonomy of [5] since it gives a comprehensive overview of the techniques. Based on this paper, we have treated any form of substitution, replacement, and imputation as “imputation”. For example, “mean substitution”, “hot/cold-deck imputation”, and “regression imputation” were treated as “imputation”. The “advanced imputation” methods were

the “full information maximum likelihood estimation” (FIML), and all variants of “multiple imputation” (MI). The definition of “deletion” is quite straightforward: every technique that includes the deletion of cases/observations, such as listwise/pairwise deletion.

4 Preprocess

To be able to extract the necessary information from the articles, we had to clean the text from meaningless symbols and noises, since the body of each article contained for example \LaTeX markups, numbers, and Optical Character Recognition (OCR) errors. To remove \LaTeX markups and to collect additional keyword lists and functions, we have created a small auxiliary package called *jprep*. All preprocessing and cleaning script can be accessed on GitHub. Every part of the preprocessing was conducted in R, and we used Python for the classification models. We would like to emphasize some of the preprocessing steps because they have significant impact on the results. As a general preprocessing step, we have removed stopwords from our corpus. The standard English stopword list of the *tm* package, however, contains the words “were”, “not”, and “at” [25]. These words were very important for us since — for example — the expressions “data were missing”, “missing at random”, or “observations were discarded” contain these stopwords. To avoid information loss and to be able to analyze our corpus correctly we had to keep these words. The next important preprocessing step was to trim the corpus from long texts. One would assume that since we are querying only articles, there are no outliers in the sense of article length. Unfortunately, there were several documents that got into our query which were not articles. After examining the distribution of the number of tokens in the corpus, we have chosen a reasonable threshold (20000 tokens) for cutting the “tail” of our distribution to remove outliers (407 articles). Another step concerning the length of the documents was the removal of references and bibliography. These sections were unnecessary for our analysis and most likely would have biased the classification results. If an article used some kind of missing data technique, then probably referenced it afterwards. This means that we would have missing data-related keywords outside

the main contents. As we will discuss shortly, our models used a small piece of information from the article bodies, therefore an additional noise — such as the detailed references of other papers — could have shifted the focus of the classifier. As an example, let us suppose that an article mentions only once the EM-Algorithm in the body and cites one of the works of Little and Rubin. In the body, the classifier identifies the context in which the “EM-Algorithm” is mentioned, but since the respective paper is referenced at the end of the article, the classifier gets a further, unnecessary context. Instead of one meaningful context, we end up with two, from which one is absolutely useless. Lastly, we have applied another technique to boost our classification accuracy by further trimming the content of the documents. We “snipped out” the context of some predefined keywords¹ from each document in order to focus only the key parts of texts. During qualitative examination of some randomly selected papers we have seen that only a small fraction of the body of a paper deals with or even mentions the missing data handling method. So trimming down the papers only discards unnecessary noise from the texts — what we do not need for our classification task. For the sake of example, let us assume that there is an article about a clinical experiment where the researchers have decided to remove some observations due to their ineligibility, and documented their decision with the statement “[...] 12 cases were deleted from the analysis due to missingness.”. This is the only part of the article that would contain information about the missing data handling method, but this sentence is only a small piece of text compared to the whole paper. In order to identify the missing data handling method, our model should be able to correctly classify this article based on one sentence. To bypass this difficulty, we snip out the context around “missingness” and discard the remaining part of the article. As our results showed, using a small but meaningful fraction of each paper not only produced better classification performance, but it decreased the time required to train our models.

¹The keywords were: “miss”, “missing”, “imput”, “impute”, “imputation”, “imputed”, “imputing”

5 Classification

Because of the nature of our analyzed corpus, it was quite hard to find a classification model that can handle our special setup i.e. the aim of the classification was to separate papers based on minimal information which — among other things — consisted of rare words. To train a supervised model, we had to make labeled training sets for each level of classification (see 1. Therefore, the authors of this paper hand-coded 200 papers in each level according to the respective goal, and then trained a model with these training sets. The papers for annotation were randomly selected from the corpus. At the end of the annotation, we had $4 \times 200 = 800$ labeled papers for each level of classification. Our very first attempt was to use popular supervised models such as Support Vector Machine (SVM), Kernel Logistic Regression (KLR), or Naïve Bayes (NB). All of these models have failed, most likely because of the small and imbalanced (10-90 ratio) training samples and unusual classification task. After these models, we tried several semi-supervised models ([26], [27], [28], [29]) in order to utilize the unlabeled cases. It was a small step forward in terms of classification performance, but far from ideal. Not only the training times were extremely long, but the accuracy of the semi-supervised models was not that much improvement that we anticipated. Furthermore, the implementation of the models made it tedious to use them effectively. All the aforementioned supervised and semi-supervised models used GloVe embeddings ([30]) for classification, so we assumed that it may have some effect on the performance of the models. Based on this assumption, we changed to an all-in-one fastText model ([31]). FastText is not only extremely efficient and fast, but it has a huge advantage over traditional word-embedding models, since it handles out-of-vocabulary and rare words better ([32]). Like GloVe, most embedding techniques create a word vector for each word in the training corpus, hence ignoring the morphological details. FastText, on the contrary, uses character level vectorization, i.e. creates character n-grams. These character n-grams are then added together to represent the respective word. For example, GloVe gives a 5-dimensional vector representation² for the word

Level	F1-score	Recall	Specificity	MCC
1	0.98	0.98	0.75	0.73
2	0.94	0.93	0.67	0.57
3.1	0.88	0.88	0.81	0.69
3.2	0.95	0.91	0.42	0.61

Table 1 Performance of fastText classification models by classification levels. Levels: 1 - About missing data or not, 2 - Imputation or not, 3.1 - Advanced Imputation or not, 3.2 - Deletion or not

“imputation” such as [0.34, 0.8, -0.12, -0.45, 0.77]. FastText, on the other hand, creates the word vector as the sum of the following character n-grams³: <im, imp, mpu, put, uta, tat, ati, tio, ion, on>. This way, even if “imputation” is not in the model’s training corpus, the character n-grams are. FastText’s main advantage in our research is its character n-gram approach. We found many OCR errors in the article bodies. The fastText model handled these errors better than the GloVe embeddings.

Before discussing the results of our analysis, we briefly present the performance of the classification models at each level. As we have mentioned, we used fastText models in all levels. To measure performance, we used the F1-score, Recall, Specificity, and MCC (Matthews Correlation Coefficient) metrics. It may seem unnecessary to present all of these metrics in order to assess the performance, but our imbalanced training set requires us to consider a more in-depth evaluation.

As we can see from Table 1, the “easiest” task was to decide whether an article was about missing data handling methods or not (Level 1). This coincides with our intuition: if an article discusses missing data handling methods, then it includes a lot of sentences which contain keywords like “imputation” or “missing data”. It gives the model more information to identify and distinguish these articles from others. On Level 2, however, we can see that despite the high F1-score and Recall values, the Specificity dropped to 0.67. It means that it was more difficult for the model in this level to find the articles that used imputation. If we recall our preprocessing steps again, we can conclude that the several meanings of the word “imputation” might affect the performance. Level 3.1 was quite consistent: our model could safely identify if an article used advanced imputation or not. On the contrary, on Level 3.2

²The numbers are arbitrary

³If we take n=3

the model had more trouble finding the papers that used some kind of deletion technique. This result is consistent with the observation of [11], namely that researchers tend to omit the reporting of deletion in their papers. [11] says, moreover, that sometimes only the tables or degrees of freedoms imply that some cases were deleted from the database. Of course, our models are not able to identify such subtle details.

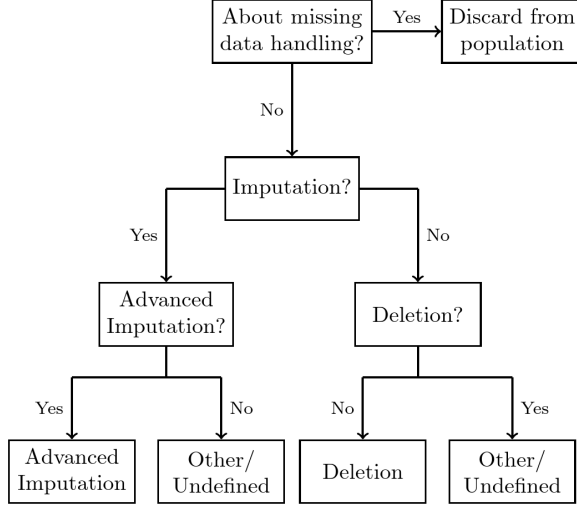


Fig. 1 Levels and stages of classification

6 Corpus description

As a result of the first classification level (about missing data or not), we discarded 1243 papers from our initial corpus (After removing outliers, the corpus consisted of 49.196 papers). The interpretation of imputation is problematic sometimes because of the polysemantic nature of the words “imputation” and “impute”. Besides the statistical meaning of “imputation”, as per the Cambridge dictionary [33]⁴, it has the following meaning: “a suggestion that someone is guilty of something or has a particular bad quality”. This definition implies that in certain disciplines a bias could occur. Indeed, the discipline categories “Criminology & Law” and “Humanities & Arts” presented high relative frequency of imputation,

Discipline	Frequency (Percentage)
Biological Sciences	9169 (20.2)
Business & Economics	8617 (18.9)
Social Sciences	7840 (17.2)
Health Sciences	5577 (12.3)
Science & Mathematics	4099 (9)
Psychology & Education	3583 (7.9)
Public Policy & Administration	2450 (5.4)
Other	1548 (3.4)
Environmental Science	1360 (3)
Political Science	1246 (2.7)
Humanities & Arts	-
Criminology & Law	-
Total	45489 (100)

Table 2 Frequency of discipline categories in the corpus

but low frequency of other missing data handling method. Therefore, we decided to exclude all papers from these two discipline categories from our analysis (2464 papers). Therefore, our actual working corpus consisted of 45.489 articles. We have divided the articles into 12 major discipline categories based on the journal and scientific discipline information from their metadata. The majority of the papers is from Biological Sciences (20.2%), Business & Economics (18.9%), Social Sciences (17.2%), and Health Sciences (12.3%). As we have mentioned before, previous research on missing data handling methods focused mainly on Social- and Educational Sciences, therefore our research may provide a more widespread perception of the applied missing data techniques. There is a caveat, however, since we do not know exactly which paper used empirical data during their respective research. This is a major difference between ours and the previous studies’ approach. Therefore, we need to assume that there are studies in our corpus that used some kind of empirical data and that our model can identify them. Of course, it is not a plausible assumption in the case of for example Humanities & Arts. One must keep in mind that the distribution of disciplines we show in Table 2 is only a description of scientific disciplines in our initial corpus. It does not carry any information about the distribution of missing data handling methods among these academic fields in general. Our results about missing data handling methods can not be generalized to the full set of articles since we do not know the exact distribution of them.

⁴<https://dictionary.cambridge.org/us/dictionary/english/imputation> (last accessed 2023.05.28.)

The time interval of our study was from 1999.01.01. to 2016.12.31. This means that only

those articles could get into our corpus that were published in this period. Table 3 in the appendix shows the distribution of papers by publication years in our corpus. There were 22 articles where no publication year was documented.

7 Results

7.1 Missing data handling methods by year

Overall, the three main categories of our classification were imputation, advanced imputation, and deletion. We focus primarily on advanced imputation, but we also highlight some exciting trends from the other two categories. Figure 3 displays the change in the usage of missing data handling methods over years. Since the amount of articles differs year by year, we did not use the raw frequencies. Instead, we made relative frequencies for missing data handling methods in each year (and later, in each discipline). There is a significant increasing trend in the case of imputation and advanced imputation. The usage⁵ of advanced imputation methods grew from 2.4% to almost 10% over the years. It even surpassed the relative frequency of deletion methods. The turning point between these two techniques was the period 2009-2011. The change in the usage of imputation methods is similar to the trend of advanced imputation. From 10.3%, it almost reaches 19% at the end of the interval. The usage of deletion methods is stagnating with a little fluctuation over the period. It constantly stays between 5.8% and 8.1%.

There are several factors that could have influenced the usage of advanced missing data handling techniques over the years. Maybe the most straightforward to assume would be the spread of modern statistical softwares, packages, and other analytic tools. As the softwares used in data analysis became more advanced, more missing data handling options were implemented. For example, in the case of the R programming language, the packages MICE and Amelia offer sophisticated and easily applicable methods to deal with missing data ([34], [35], [36]). This claim is further supported by the observation of [36, p. 83]:

⁵We do not know whether a missing data handling method was *actually* used – one of the limitations of text mining

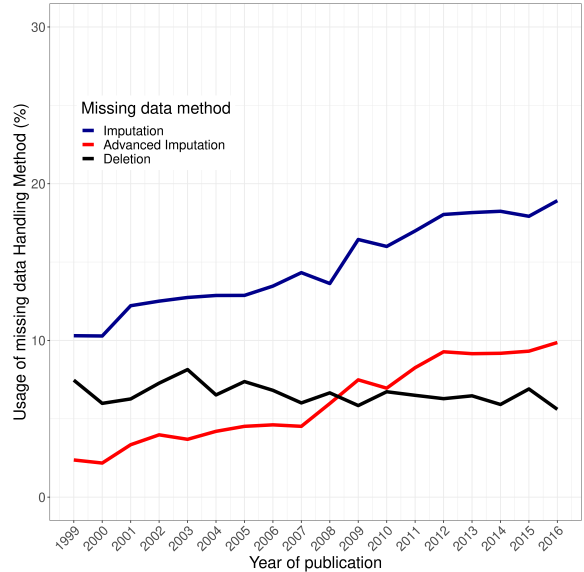


Fig. 2 Usage of missing data handling methods by year (1999-2016). The vertical axis shows the percentage of papers that used the respective missing data handling methods in a given year.

“Both reviews [referring to the articles of [18] and [37] — note by K.B.] indicate that there is a considerable gap between statistical methodologies and methods that are commonly used in practice. Flexible comprehensive implementations of these methods may spur their use.” Our findings imply that advancements in implementations of missing data handling techniques may have increased their usage.

7.2 Logistic models

To present the differences between disciplines, we fit two binomial logistic regression models on our data: one where the outcome variable was whether there was an imputation or not; and one where the outcome variable was whether there was an advanced imputation or not. For explanatory variables, we used publication year and discipline category. In the case of discipline category, we used political science as the reference category. We extended the base models with the interaction of time and discipline category to explore whether we find differences in the temporal variation of imputation in different disciplines.

We start the analysis with the imputation part of the regression. The time variable was significant with a positive value, which confirms the univariate approach, there was a significant increase in the use of imputation method. We plotted the predicted marginal estimates for the disciplines (see Figure 3 and Figure 4 in the appendix). Political Science and Social Science papers used most frequently imputation methods. In the other side of the scale, we could find Biological and Environmental Sciences. The interaction presents the different temporal pattern of imputation trends across disciplines. The reference category is Political Science, where the increase is equal with the estimate of publication year variable. Compared to Political Science, Biological and Health Science, had a steeper increase in the use of imputation (see the positive and significant interaction terms), and Business and Economics and Science and Mathematics differed negatively from this trend. For the latter based on the marginal predications we could observe a decrease in the use of imputation. The second regression analyze the factors behind the variance of advanced imputation level. Here we can observe a positive trend value, as expected, so year by year the advanced imputation was more and more popular. Our results also mean, that within imputation, advanced imputation usage was increased. But we can observe high differences between the disciplines (see Figure 4 in the appendix for marginal predictions). In the disciplines of Psychology and Education, Health Science and Political Science the level of advanced imputation was around 60-70 percent. But in Business and Economics, Biological Science and Environmental Science only 20-30 percent of the imputations used advanced technique. The interaction terms reveal three temporal differences behind the spreading of advanced imputation. Psychology and Education had a more intense increase in the level of advanced imputation compared to the rest of the disciplines. On the other hand, we could observe a decline of these techniques in the field of Science and Mathematics and Biological Science.

8 Discussion

The aim of our study was to identify the trends in the usage of missing data handling methods within various scientific disciplines from 1999 to

2016. Missing data is a pivotal element of many empirical research since it is practically impossible to gather all the data we originally intended. This immanent problem of information loss helped missing data handling methods emerge and evolve. During the past 50 years, more and more techniques were developed for assessing missing data. Researchers began to examine the various methods in disciplines like Educational Science or Psychology; and so the number of surveys and meta-analyses on this topic started to increase. In contrast to previous research in this topic, we utilized a text mining approach to extract the necessary information from the articles. This new methodology, however, comes with several advantages and disadvantages. On one hand, it allows us to work with a much larger corpus than the previous studies. The actual analysis of the articles is less human resource intensive and the whole research is very scalable: it does not matter whether we work with ten thousand papers or with one million – the increase of computational time will be negligible. Additionally, with a larger corpus, we are able to make comparisons of missing data handling methods among various disciplines. On the other hand, text mining makes us researchers more distant from the papers. Since we have not actually seen the contents of each paper, we can never be sure if a paper used a missing data handling method or only mentioned it – we need to trust the classification. The identification of the used methods was a problem even in previous studies: oftentimes researchers neglected the appropriate documentation of methods they used to handle missing data. And if a human cannot decide whether there was any kind of missing data handling, then how would a classification algorithm could. For example, there were several instances, where the authors did not mention which kind of technique they used to handle missing data in their research, only the difference of the samples sizes in the models implied that a deletion technique was used ([11]). Clearly, an algorithm is not able to notice such subtle detail, but a human can. All in all, this approach inevitably results in some kind of information loss and the underestimation of imputation usage. We used a limited number of keywords to detect those papers where missingness could be an issue. Our keyword choice might underrepresent some disciplines, where different

phrases are also used to describe non-response (like item-nonresponse in social science). And there are also differences in data-generating processes between fields. Missingness is usually higher in surveys than in experimental designs. But our analysis could well present the temporal trends of applying imputation and advanced imputation within a field. Our results show that the usage of imputation and advanced imputation methods increased during the period 1999-2016. One plausible argument to explain this increase is that the documentation of missing data handling methods improved, therefore it is much more easier to find them in the papers. We think this claim could be one of the possible causes. The evolution and implementation of these techniques could have boosted their application. More and more statistical software implements complex missing data handling methods which makes these techniques more accessible for researchers. Furthermore, the growing tendency of item non-response in survey-type data collection ([38]) could also facilitate the usage of missing data handling methods. As mentioned above, it is also obvious that the type of data researchers analyze differs through disciplines. It is not evident that missingness appears in the same level. In Political Science and Social Science surveys are the main quantitative methods, and surveys always contains some level of missing data. From this point of view, it is not surprising, that imputation is the most common in these fields. But as we narrowed our initial corpus to those papers which contain words about missing data or observation, we could assume, that this disciplinary difference is lower in our sample, compared to the whole fields. And when we find imputation, we could expect less disciplinary differences between advanced and not advanced techniques. Our result did not support this expectation. The disciplinary difference was huge, and we could also observe differences between Political and Social Science, where the type of data sources is quite similar. This is a clear sign, that not only the data type is important, but disciplines have their own methodological canon which has strong effect on how scholars in different fields handle the missing data problem.

9 Declarations

9.1 Founding Sources

The work of Krisztián Boros was supported by the Japanese Government

(Monbukagakusho: MEXT) Scholarship.

The work of Zoltán Kmetty was supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program.

9.2 Competing interests

The authors have no competing interests to declare that are relevant to the content of this article

References

- [1] Dong, Y., Peng, C.-Y.J.: Principled missing data methods for researchers. Springer-Plus **2**(1), 222 (2013) <https://doi.org/10.1186/2193-1801-2-222>
- [2] Enders, C.K.: Applied Missing Data Analysis. Methodology in the social sciences. Guilford Press, New York (2010). OCLC: ocn456171131
- [3] Graham, J.W., Cumsille, P.E., Shevock, A.E.: Methods for Handling Missing Data. In: Handbook of Psychology, 2nd edn., pp. 109–141. Wiley, ??? (2013)
- [4] Little, T.D., Jorgensen, T.D., Lang, K.M., Moore, E.W.G.: On the Joys of Missing Data. Journal of Pediatric Psychology **39**(2), 151–162 (2014) <https://doi.org/10.1093/jpepsy/jst048>
- [5] Little, T.D., Lang, K.M., Wu, W., Rhemtulla, M.: Statistical Issues: What Happens When Data Go Missing? In: Developmental Psychopathology, Third edition edn., p. 37. Wiley, ??? (2016)
- [6] Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, Third edition edn. Wiley series in probability and statistics. Wiley, Hoboken, NJ (2019)

- [7] Schafer, J.L., Graham, J.W.: Missing data: Our view of the state of the art. *Psychological Methods* **7**(2), 147–177 (2002) <https://doi.org/10.1037/1082-989X.7.2.147>
- [8] Wilkinson, L., Task Force on Statistical Inference: Statistical Methods in Psychology Journals: guidelines and explanations. *American Psychologist* **54**(8), 594–604 (1999)
- [9] Bell, M.L., Fiero, M., Horton, N.J., Hsu, C.-H.: Handling missing data in RCTs; a review of the top medical journals. *BMC Medical Research Methodology* **14**(1), 118 (2014) <https://doi.org/10.1186/1471-2288-14-118>
- [10] Cheema, J.R.: A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research* **84**(4), 487–508 (2014) <https://doi.org/10.3102/0034654314532697>
- [11] Peugh, J.L., Enders, C.K.: Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research* **74**(4), 525–556 (2004) <https://doi.org/10.3102/00346543074004525>
- [12] Rubin, D.B.: Inference and Missing Data. *Biometrika* **63**(3), 581–592 (1976)
- [13] Roth, P.L.: Missing Data: A Conceptual Review for Applied Psychologists. *Personnel Psychology* **47**(3), 537–560 (1994) <https://doi.org/10.1111/j.1744-6570.1994.tb01736.x>
- [14] Bodner, T.E.: Missing Data: Prevalence and Reporting Practices. *Psychological Reports* **99**(3), 675–680 (2006) <https://doi.org/10.2466/PRO.99.3.675-680>
- [15] King, G., Honaker, J., Joseph, A., Scheve, K.: Analyzing Incomplete Political Science Data: An Alternative Algorithm for Multiple Imputation. *American Political Science Review* **95**(1), 49–69 (2001) <https://doi.org/10.1017/S0003055401000235>
- [16] Peng, J., Harwell, M., Liou, S.-M., Ehman, L.H.: Advances in missing data methods and implications for educational research. In: *Real Data Analysis. Quantitative Methods in Education and the Behavioral Sciences: Issues, Research, and Teaching*, pp. 31–78. Information Age Publishing, Charlotte, NC (2006)
- [17] Jeličić, H., Phelps, E., Lerner, R.M.: Use of missing data methods in longitudinal studies: The persistence of bad practices in developmental psychology. *Developmental Psychology* **45**(4), 1195–1199 (2009) <https://doi.org/10.1037/a0015665>
- [18] Burton, A., Altman, D.G.: Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines. *British Journal of Cancer* **91**(1), 4–8 (2004) <https://doi.org/10.1038/sj.bjc.6601907>
- [19] Wood, A.M., White, I.R., Thompson, S.G.: Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clinical Trials: Journal of the Society for Clinical Trials* **1**(4), 368–376 (2004) <https://doi.org/10.1191/1740774504cn032oa>
- [20] Fielding, S., MacLennan, G., Cook, J.A., Ramsay, C.R.: A review of RCTs in four medical journals to assess the use of imputation to overcome missing data in quality of life outcomes. *Trials* **9**(1), 51 (2008) <https://doi.org/10.1186/1745-6215-9-51>
- [21] Gravel, J., Opatrny, L., Shapiro, S.: The intention-to-treat approach in randomized controlled trials: Are authors saying what they do and doing what they say? *Clinical Trials* **4**(4), 350–356 (2007) <https://doi.org/10.1177/1740774507081223>
- [22] Hollis, S., Campbell, F.: What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ (Clinical research ed.)* **319**(7211), 670–674 (1999) <https://doi.org/10.1136/bmj.319.7211.670>
- [23] Burns, J., Brenner, A., Kiser, K., Krot, M., Llewellyn, C., Snyder, R.: JSTOR - Data for Research. In: Agosti, M., Borbinha, J., Kapidakis, S., Papatheodorou,

- C., Tsakonas, G. (eds.) Research and Advanced Technology for Digital Libraries vol. 5714, pp. 416–419. Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-642-04346-8_48. Series Title: Lecture Notes in Computer Science
- [24] Boros, K., Kmetty, Z.: Identifying missing data handling methods with text mining [dataset]. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor] (2023) <https://doi.org/10.3886/E185961V1>
- [25] Feinerer, I., Hornik, K.: tm: Text Mining Package. R package version 0.7-8 (2020)
- [26] Abdel-Hady, M., Schwenker, F., Palm, G.: Semi-supervised Learning for Regression with Co-training by Committee, vol. 5768, pp. 121–130 (2009). https://doi.org/10.1007/978-3-642-04274-4_13
- [27] Bennett, K.P., Demiriz, A.: Semi-Supervised Support Vector Machines. In: Kearns, M.S., Solla, S.A., Cohn, D.A. (eds.) Advances in Neural Information Processing Systems vol. 11, pp. 368–374. MIT Press, London (1999)
- [28] Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-supervised Learning. Adaptive computation and machine learning. MIT Press, Cambridge, Mass (2006). OCLC: ocm64898359
- [29] Cohen, I., Cozman, F.G., Sebe, N., Cirelo, M.C., Huang, T.S.: Semisupervised learning of classifiers: theory, algorithms, and their application to human-computer interaction. IEEE Transactions on Pattern Analysis and Machine Intelligence **26**(12), 1553–1566 (2004) <https://doi.org/10.1109/TPAMI.2004.127>
- [30] Pennington, J., Socher, R., Manning, C.: Glove: Global Vectors for Word Representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. Association for Computational Linguistics, Doha, Qatar (2014). <https://doi.org/10.3115/v1/D14-1162>
- [31] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of Tricks for Efficient Text Classification. arXiv:1607.01759 [cs] (2016). arXiv: 1607.01759
- [32] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching Word Vectors with Subword Information. arXiv:1607.04606 [cs] (2017). arXiv: 1607.04606
- [33] Cambridge Dictionary. Cambridge: Cambridge University Press (2020). <https://dictionary.cambridge.org/dictionary/english/imputation> Accessed 2023-05-28
- [34] Buuren, S.v., Groothuis-Oudshoorn, K.: mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software **45**, 1–67 (2011) <https://doi.org/10.18637/jss.v045.i03>
- [35] Honaker, J., King, G., Blackwell, M.: Amelia II: A Program for Missing Data. Journal of Statistical Software **45**(7) (2011) <https://doi.org/10.18637/jss.v045.i07>
- [36] Horton, N.J., Kleinman, K.P.: Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician **61**(1), 79–90 (2007) <https://doi.org/10.1198/000313007X172556>
- [37] Horton, N.J., Switzer, S.S.: Statistical Methods in the Journal (research letter). New England Journal of Medicine **353**, 1977–1979 (2005)
- [38] Luiten, A., Hox, J., Leeuw, E.: Survey nonresponse trends and fieldwork effort in the 21st century: Results of an international study across countries and surveys. Journal of Official Statistics **36**(3), 469–487 (2020) <https://doi.org/10.2478/jos-2020-0025>

Appendix

A Additional Figures

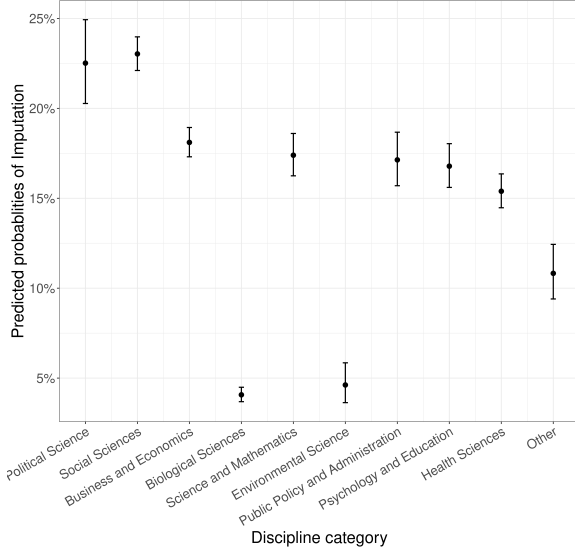


Fig. 3 Predicted marginal probabilities of imputation level per disciplines based on binomial logistic regression model

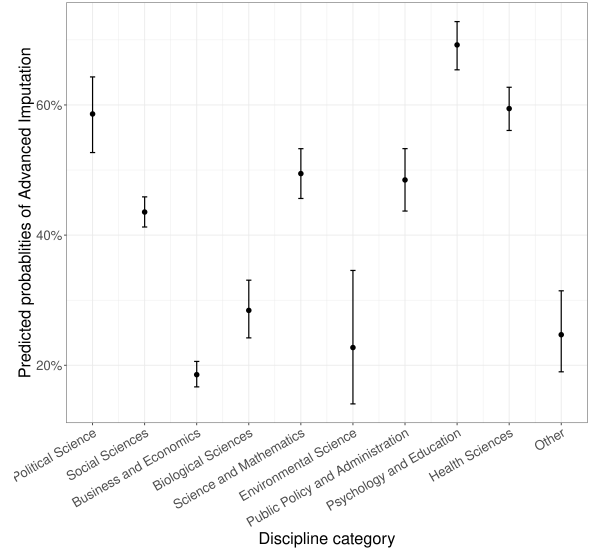


Fig. 4 Predicted marginal probabilities of advanced imputation level per disciplines based on binomial logistic regression model

B Additional Tables

Year of Publication	Frequency (Percentage)
2016	2088 (4.6)
2015	2405 (5.3)
2014	3399 (7.5)
2013	3310 (7.3)
2012	3310 (7.3)
2011	3232 (7.1)
2010	3063 (6.7)
2009	2926 (6.4)
2008	2795 (6.1)
2007	2611 (5.7)
2006	2406 (5.3)
2005	2346 (5.2)
2004	2238 (4.9)
2003	1978 (4.3)
2002	1911 (4.2)
2001	1883 (4.1)
2000	1838 (4)
1999	1728 (3.8)
NA	22 (0)
Total	45489 (100)

Table 3 Frequency of papers in the corpus by year of publication (1999-2016)

<i>Dependent variable:</i>				
	Advanced Imputation	Imputation	Advanced Imputation (with interaction)	Imputation (with interaction)
	(1)	(2)	(3)	(4)
Publication Year	0.080*** (0.006)	0.047*** (0.003)	0.079*** (0.029)	0.051*** (0.015)
Social Sciences	−0.608*** (0.132)	0.029 (0.073)	−86.185 (62.502)	−17.824 (31.461)
Business and Economics	−1.827*** (0.139)	−0.273*** (0.074)	−6.092 (65.078)	85.780*** (31.626)
Biological Sciences	−1.271*** (0.166)	−1.924*** (0.086)	184.183** (79.520)	−188.191*** (39.373)
Science and Mathematics	−0.370** (0.145)	−0.322*** (0.080)	232.169*** (66.942)	111.035*** (34.162)
Environmental Science	−1.572*** (0.323)	−1.792*** (0.144)	110.660 (123.502)	87.222 (58.673)
Public Policy and Administration	−0.409*** (0.157)	−0.340*** (0.087)	73.159 (72.581)	42.407 (36.683)
Psychology and Education	0.463*** (0.152)	−0.365*** (0.081)	−129.815* (71.734)	−23.218 (35.322)
Health Sciences	0.034 (0.142)	−0.468*** (0.077)	−80.501 (67.307)	−86.852** (34.019)
Other	−1.463*** (0.211)	−0.873*** (0.105)	−52.012 (103.886)	17.912 (44.458)
Social Sciences			0.043 (0.031)	0.009 (0.016)
Pub Year* Business and Economics			0.002 (0.032)	−0.043*** (0.016)
Pub Year* Biological Sciences			−0.092**	0.093***

			(0.040)	(0.020)
Pub Year*				
Science and				
Mathematics			−0.116***	−0.055***
			(0.033)	(0.017)
Pub Year*				
Environmental				
Science			−0.056	−0.044
			(0.061)	(0.029)
Pub Year*				
Public Policy and				
Administration			−0.037	−0.021
			(0.036)	(0.018)
Pub Year*				
Psychology and				
Education			0.065*	0.011
			(0.036)	(0.018)
Pub Year*				
Health Sciences			0.040	0.043**
			(0.034)	(0.017)
Pub Year*				
Other			0.025	−0.009
			(0.052)	(0.022)
Constant	−160.953***	−95.380***	−157.780***	−104.178***
	(11.844)	(5.659)	(58.295)	(29.385)
Observations	6,921	45,467	6,921	45,467
Log Likelihood	−4,210.530	−18,351.720	−4,162.216	−18,259.040
Akaike Inf. Crit.	8,443.061	36,725.440	8,364.433	36,558.080

Note:

*p<0.1; **p<0.05; ***p<0.01

Table 4: Logistic models