Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001

JOURNAL OF INFORMATION SYSTEM AND TECHNOLOGY MANAGEMENT
(JISTM)
www.jistm.com

A BAD IDEA OF USING MODE IMPUTATION METHOD

Afiqah Bazlla Md Soom[1*], Aisyah Mat Jasin[2*], Aszila Asmat[2], Roger Canda[2], Juhaida Ismail[2]

[1] College of Computing, Informatic and Media, Universiti Teknologi MARA Cawangan Pahang Kampus Jengka, Pahang, Malaysia
[2] College of Computing, Informatic and Media, Universiti Teknologi MARA Cawangan Pahang Kampus Raub, Pahang, Malaysia
[*] Corresponding Author (afiqahbazlla@uitm.edu.my, aisyahmj@uitm.edu.my)

Article Info:

Article history:
Received date: 15.09.2022
Revised date: 24.10.2022
Accepted date: 17.11.2022
Published date: 01.12.2022

To cite this document:

Md Soom, A. B., Mat Jasin, A., Asmat, A., Canda, R., & Ismail, J. (2022). A Bad Idea Of Using Mode Imputation Method. *Journal of Information System and Technology Management, 7* (29), 01-09.

DOI: 10.35631/JISTM.729001

This work is licensed under CC BY 4.0

Abstract:

Missing data is a recurring issue in psychology questionnaire when a respondent does not respond to questions due to personal reasons. In general, two types of imputation techniques are used to replace missing data: single imputation and multiple imputation (MI). The single imputation technique generates a single value to impute each missing data. The simplest methods of single imputation are mean, mode and median. In contrast, the multiple imputation technique imputes each missing data several times resulting in multiple complete datasets. The most popular method in MI that can deal with numerical and categorical data type is the predictive mean matching (PMM). The aim of this article is to compare and visualize how the mode imputation method in the single imputation technique will lead to a biased data distribution and the PMM method in the MI techniques will reduce this issue. Both methods, mode imputation and PMM are often considered when dealing with categorical data types. The mode imputation replaces a missing data with the most frequent value of an item in a survey. Meanwhile, the predictive mean matching is an extension of regression model that apply donor selection strategy to replace a missing data. Results from bar charts visualize the multiple imputation shows less discrepancy between the original distribution and imputed distribution. Thus, in this research, it can be concluded that the PMM method in MI technique shows a less biased distribution than implementing the mode imputation method. A comparison of imputation methods with different missing rates on a survey dataset should be considered for future work.

Keywords:

Missing Data, Single Imputation, Mode Imputation, Multiple Imputation, PMM

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001

Introduction

Missing data is a common occurrence in a questionnaire-based study in various research domains such as medical, psychology, economics, and numerous other disciplines. Missing data occurs when the value of a variable is recorded only for certain subjects in the sample (Austin & van Buuren, 2022). The presence of missing data is generally due to the reluctance of the respondents to answer certain items in the questionnaire which may lead to the occurrence of missing values in the total score. Besides, its presence may also cause misleading statistical inferences.

There are various methods to deal with missing data issue in questionnaires such as single imputation and multiple imputation (MI) methods (Van Buuren, 2012). Single imputation method generates a single replacement value to impute each missing value. Among the mean, mode and median imputation method, the mode imputation is appropriate for categorical data type. For continuous data type, the mean, median (Little, 1988) or any other imputation techniques using regression model are more relevant. Each of these methods has their limitations such as generate biased and unrealistic results. Yet, it is stated that multiple imputation method is the most frequently used in addressing missing data as it may reduce bias and affect the representative of the results (Austin et al., 2021). According to [4], MI is an approach where multiple plausible values are drawn from a distribution and replaced in the missing element of variables. Hence, multiple completed data sets being created. Multiple imputation has been applied using statistical prediction based on Rubin's rules in many research domains (Schafer, 1997; Rubin and Schenker, 1986). Kabir et al. (2014) applied single imputation and MI in a water distribution network for continuous data type, found that the use of MI outperformed the single imputation in the comparison study. Acuña and Rodriguez (2004) applied the case deletion, mean imputation, median imputation and KNN imputation method on twelve datasets to evaluate the effect of these methods on the classification accuracy. Although the KNN imputation showed as the best performance and reduces bias compared to the other methods, the median imputation is still relevant when missing rate is not high and distribution of features are asymmetric and contains outlier. Nakagawa (2015) stated that, when applied to the datasets of the bird population on Lundy Island, the results of employing single imputation methods generally cause substantial bias in parameter estimates; however, when multiple imputation methods are used, the problem of biased uncertainty is solved. When comparing the multiple imputation method to the single imputation method for the data from digital health technologies (DHTs), especially accelerometers and continuous glucose monitors (CGM), Di et al (2022) concluded that the multiple imputation method is the one that is most commonly used for complex incomplete data. The association that exists between the variables in the data can be maintained using multiple imputation, which also allows for the possibility of taking into consideration the uncertainty.

The focus of this paper is to visualize how the mode imputation method may lead to a biased data distribution and how the PMM method in MI technique can improve this drawback. These missing data imputation methods will be conducted on a set of questionnaires to measure inhumane behaviour towards animal among participants.

The paper is organized as follows. We begin by presenting the three phases of imputation method: imputation phase, analysis phase and pooling phase. The findings of this study enable us to visualize the best imputation method in MI technique based on the comparison conducted.

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001

Methodology

MI was introduced by Donald B. Rubin (Van Buuren, 2012) in the 1970s and it is a robust imputation method compared to a single imputation that uses a Bayesian framework to handle imputation of missing data. This method is designed to assess the uncertainty of missing values by generating several complete datasets from an incomplete dataset by imputing the missing data several times to produce a single set of inference.

Imputation Method

The three phases of MI are shown in Figure 1 which include imputation, analysis, and pooling.

Imputation Phase, all missing values in a dataset are replaced with reasonable values to generate a complete version of the dataset. This process will be repeated M times to produce M complete datasets that has the same missing rates with different imputed values. The recommended value of M is discussed further in (Van Buuren, 2012).

Analysis Phase. Each of the M complete datasets are analysed to generate M sets of parameter estimates such as standard errors and confidence intervals.

Pooling Phase. M set of parameter values are combined using Rubin's rules to generate a single set of unbiased parameter estimates that include both within-imputation variance and between-imputation variance.

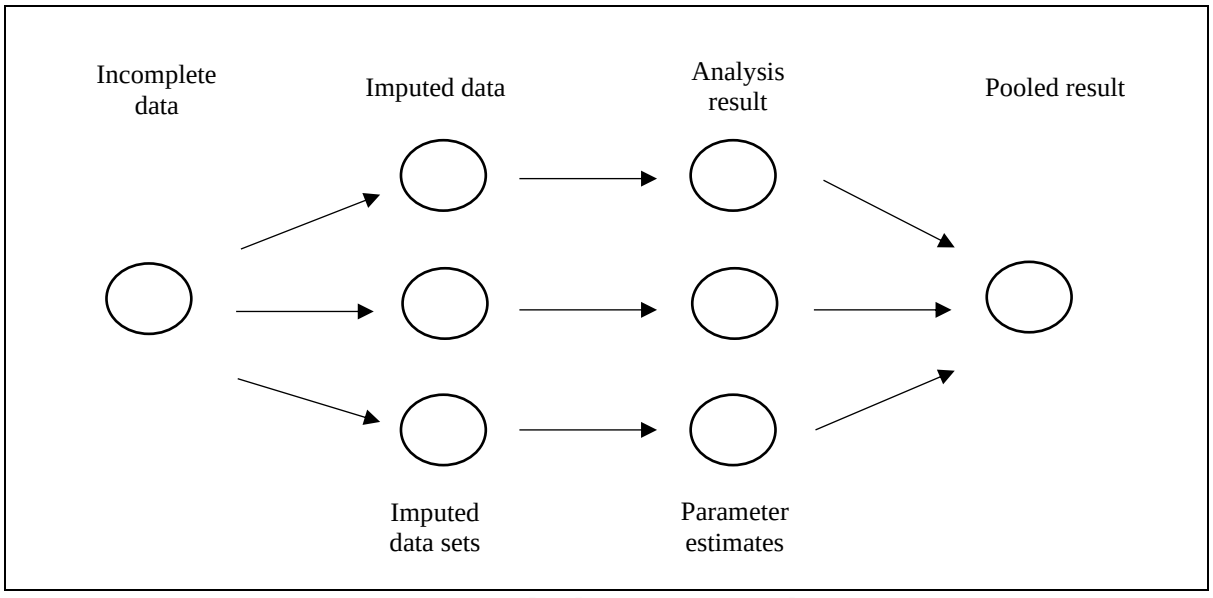

Figure 1. Multiple Imputation Phases

Data in this study was collected using an instrument consisting of 18 questions in three subsections. The subsections of the instrument were as follows: emotion (EM), social negligence (SN), and economy (EC). Each subsection consist of 6 questions and responses are based on a Likert scale consisting of a scale of 1 to 5 where 1 indicates strongly disagree and 5 indicates strongly agree. The instrument was designed to identify consent on inhumane behaviour towards animal among students in Universiti Teknologi MARA Pahang branch and was distributed to a total of 254 students who were randomly selected. Most selected students were male with a percentage of 52% and the rest were female. Each respondent's total score is

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001

calculated by adding the responses to all the questions in the instrument with a maximum score of 90. This value will be converted to a 100%.

Table 1: S Scale of Inhumanity Behaviour towards Animal

No	Label	Statement
1	EM1	I feel releasing anger on pets is helpful. */ Saya rasa melepaskan kemarahan terhadap haiwan peliharaan sangat memuaskan.*
2	EM2	I think it is fine if I am not reporting the suspect of animal abuse. */ Saya rasa tiada apa-apa jika saya tidak melaporkan suspek yang mendera haiwan.*
3	EM3	I am not reporting the suspect of animal abuse due to lack of courage. / *Saya tidak melaporkan suspek penderaan haiwan kerana kekurangan keberanian.*
4	EM4	I think people who abuse their animal are more likely to abuse their children and spouse. */ Saya merasakan pelaku yang mendera haiwan lebih cenderung mendera anak-anak serta pasangan.*
5	EM5	I think people who abuse animals are more likely to suffer from mental illness. / *Saya merasakan pelaku yang mendera haiwan lebih cenderung menghidap penyakit mental.*
6	EM6	I abuse animals to show masculinity. / *Saya mendera haiwan untuk melambangkan kemegahan.*
7	SN1	My parents will bring us to shopping mall rather than going to animal center. / *Ibu bapa saya akan bawa kami ke pusat membeli-belah sebaliknya daripada pergi ke pusat haiwan.*
8	SN2	I think people who use punishment-based animal training methods are more likely to engage in animal abuse. / *Saya merasakan masyarakat yang mengamalkan konsep kaedah latihan haiwan berasaskan hukuman lebih cenderung dalam penderaan haiwan.*
9	SN3	School authorities barely organize animal awareness affairs. / *Pihak sekolah kurang menganjurkan program kesedaran haiwan.*
10	SN4	Local council need to volunteer groups in the veterinary department to deal with animal inhumanity cases. / *Majlis Tempatan perlu menyertai kumpulan sukarelawan di Jabatan Veterinar untuk menangani kes keganasan haiwan.*
11	SN5	My peers will only acknowledge me if I am willing to abuse animals / *Rakan-rakan saya hanya akan mengiktiraf saya jika saya sanggup mendera haiwan.*
12	SN6	Lack of a dedicated organizations specialized for animal welfare. / *Kurangnya organisasi khusus untuk memelihara haiwan.*
13	EC1	I support the use of live animals in medical teaching and research. / *Saya menyokong penggunaan haiwan dalam bidang perubatan dan kajian.*
14	EC2	I support the fashion industry for using animals as a product for business purposes. / *Saya menyokong industri fesyen untuk menggunakan haiwan sebagai satu produk bagi tujuan urusan perniagaan.*
15	EC3	I enjoy riding animal rides provided by the entertainment industry. / *Keseronokan menunggang haiwan untuk tujuan hiburan.*
16	EC4	I am fine with paying more for cosmetic products even though they are tested on animals due to effective advertisement. / *Saya boleh membayar lebih untuk barangan kosmetik walaupun ianya diuji ke atas haiwan oleh kerana pengiklanan yang berkesan.*

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001

17	EC5	I agree with food industries that serve scarce meal such as rabbit, deer and quail eggs for the sake of gaining higher profit and sales. / *Saya bersetuju dengan industri makanan yang meghidangkan sumber haiwan yang sukar didapati seperti arnab, rusa dan telur burung puyuh demi meraih keuntungan yang tinggi.*
18	EC6	I do think zoos exist to entertain humans instead of saving, helping and preserving animals. / *Saya merasakan zoo diwujudkan untuk menghiburkan manusia daripada menyelamatkan, membantu dan memelihara haiwan.*

In this study, there are possible reasons for missing values to occur on the EM6 and SN1 items based on gender variable. Female respondents might ignore to answer the question on item EM6 because of the term masculinity is often associated with the actions and attitudes of boys or men (Eskola and Handroos 2013). Likewise for item SN1 which is to identify perceptions on inhumane behaviour towards animals that is related with relationship with parents among female and male respondents, there is possibility that respondents refuse to answer this item. A study conducted by Yahaya et al. (2016) found that male students need more parental involvement, while female students need more parental support. Therefore, in this study, data for these two items will be deleted randomly at a missing rate of 20%.

Results & Discussion

We carried out this experimental study for the inhumane behaviour towards animal survey dataset using Mice package in R programming tools with 3.13.0 and 4.0.5 version respectively. The MI method in the Mice R package implements Markov Chain Monte Carlo method for sampling from posterior distribution to multiply imputed values through the pair I-Steps and P-Steps.

The results for all imputed data are visualized in separate bar graphs. Basically, there are group of three vertical bars presented in each bar graph. The first red-bar represents the number of original item score while the other two blue-bar and green-bar represent the number of imputed item score using MI method in Mice R package and mode method respectively. Each group of bar graph presents a comparison of imputed data using the MI and mode imputation methods on the item EM6 based on male respondents in Figure 2 and female respondents in Figure 3. Meanwhile the comparison of both MI and mode imputation method implemented on the item SN1 for male and female are shown in Figure 4 and Figure 5 respectively.

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001

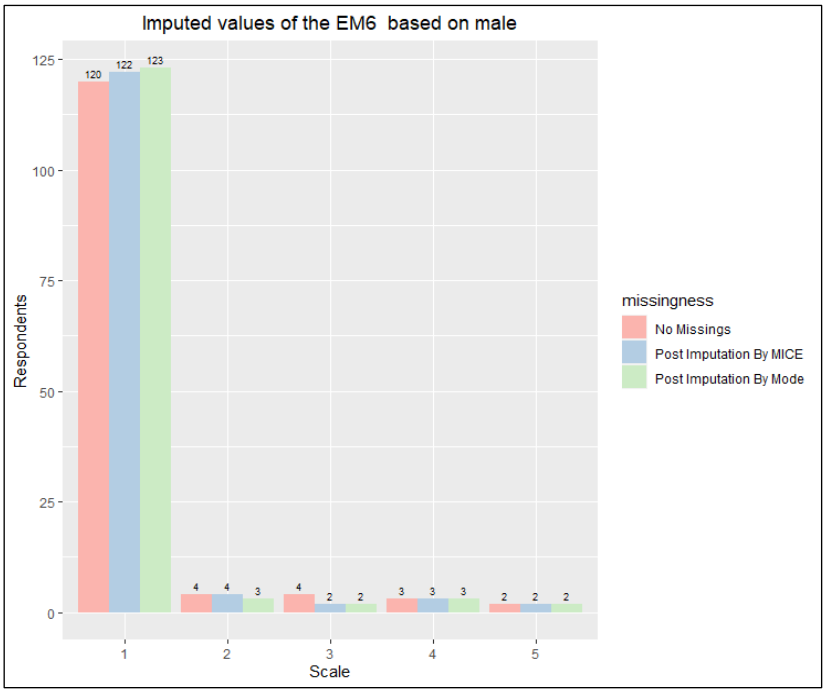

Figure 2. Imputation on EM6 Based on Male Respondents

Figure 2 shows the results using the mode imputation and the PMM methods along with the observed values for item EM6. Three missing values occurs in item EM6 due to missing rate of 20% is created randomly. Based on the observed values, 120 male respondents strongly disagreed (Likert scale at 1), and four male respondents disagreed (Likert scale at 2) that animal abuse shows masculinity. Four were neutral (Likert scale at 3) with the statement, while three and two respondents agreed (Likert scale at 4) and strongly agreed (Likert scale at 5) respectively. From the figure, the highest bar that indicates the most often scale chosen by male respondents is at 1, thus, the mode value for this item is 1. This number is imputed to the missing data for the mode imputation method and the number of input values for male respondents who voted for Likert scale 1 is expected to increase from 120 to 123. However, the value calculated using PMM shows only a slight difference from the original value.

The same pattern can be observed in Figure 3. The number of imputed values for female respondents who voted for Likert scale 1 is expected to increase from 112 to 117 when using the mode imputation although 20% of missing values are also randomly created at another Likert scales such as 2, 3 and 4. The missing values occurred at Likert scale of 4. However, because these missing values are replaced by the mode value, consequently, the Likert scale 4 for the green bar becomes empty while Likert scale 1 is increased.

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved

Journal of Information System and Technology Management
JISTM
EISSN: 0128-1666
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001

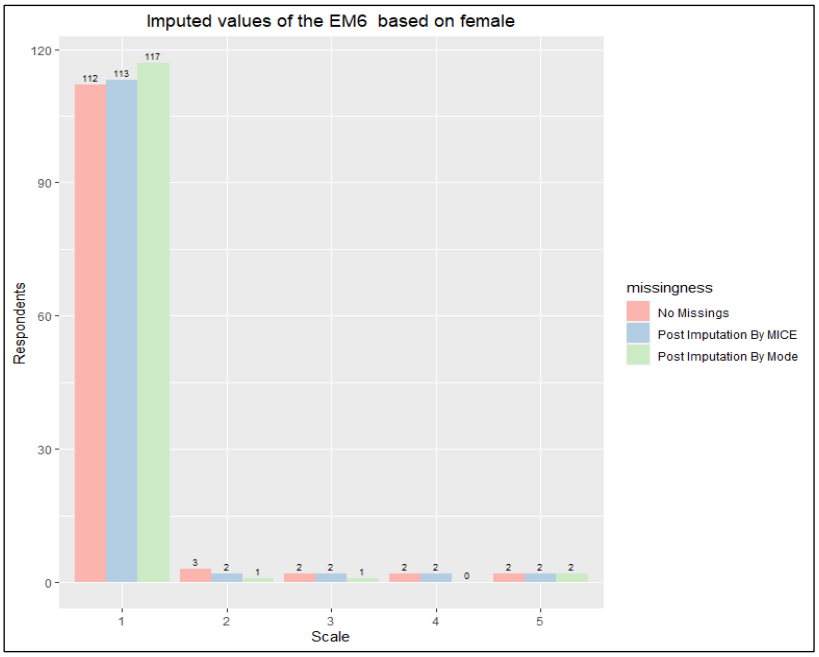

Figure 3. Imputation on EM6 Based on Female Respondent

Figure 4 and Figure 5 show a comparison before and after imputation process on the item SN1 based on male and female respondents respectively. The results in both Figure 4 and Figure 5 show that the imputation values using the mode imputation method have higher discrepancies than the imputation method using PMM model with the multiple imputation method in the Mice R package. In Figure 4, although missing values also occur in Likert scale 1, 2, 3 and 4, all these missing values are imputed by mode value. Therefore, the total Likert scale 5 is expected to increase-from 52 to 68. Similarly, in Figure 5, all the missing values in Likert Scale 1, 2, 4 and 5 are imputed by the Likert scale 3 since that is the most frequent scale selected by respondents. Therefore, the total number of Likert scale 3 is expected to increase dramatically from 35 to 51, while the rest of Likert scale values are decreased.

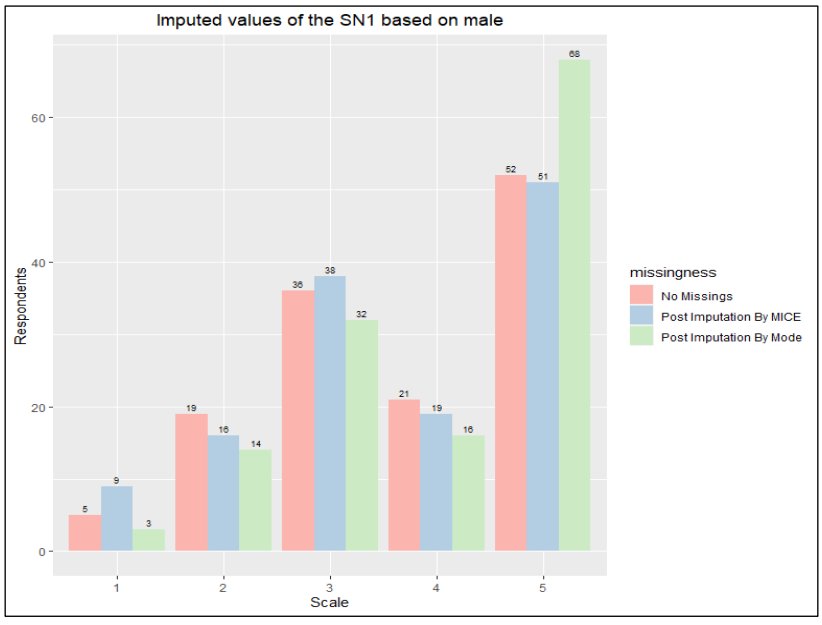

Figure 4. Imputation on SN1 Based on Male Respondent

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved

Journal of Information System and Technology Management
EISSN: 0128-1666
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001

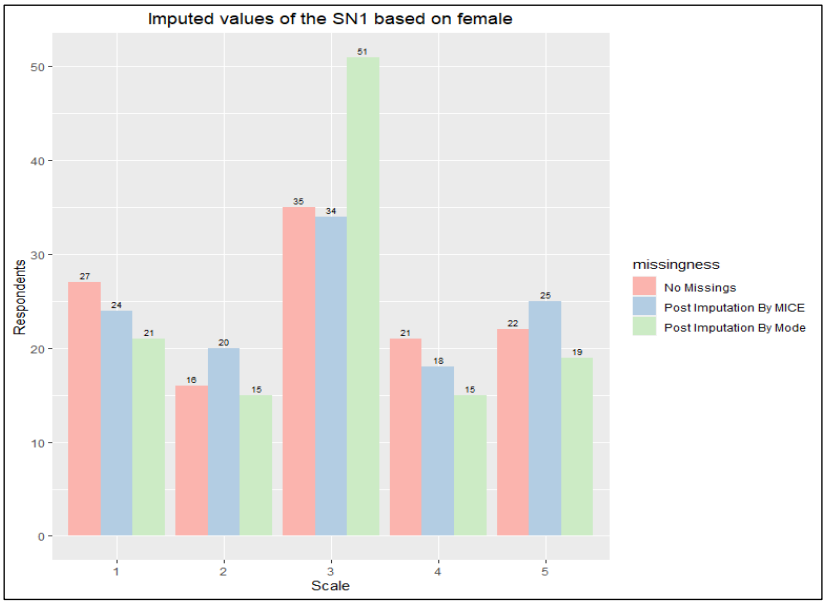

Figure 5. Imputation on SN1 Based on Female Respondent

Conclusion

This research presents a comparison between the mode imputation method as a single imputation and PMM imputation method in MI on the survey dataset. The results presented that using the mode imputation method creates a higher discrepancy between the original data and imputed data. However, using PMM method blended with MI procedure has very slight discrepancy between original data and imputed data. Therefore, it can be concluded that in this research, the mode imputation method is visually lead to a biased distribution of item EM6 and SN1. It contrasts with the PMM imputation method in MI technique, the method appeared to have better performance at all results. In future work, with similar comparison work, our target is to present a comparison of imputation method on a survey dataset based on different missing rate.

References

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications*, 639–647. https://doi.org/10.1007/978-3-642-17103-1_60.

Austin, P. C., & van Buuren, S. (2022). The effect of high prevalence of missing data on estimation of the coefficients of a logistic regression model when using multiple imputation. *BMC Medical Research Methodology*, *22*(1), 1–14. https://doi.org/10.1186/s12874-022-01671-0.

Austin, P. C., White, I. R., Lee, D. S., & van Buuren, S. (2021). Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology*, *37*(9), 1322–1331. https://doi.org/10.1016/j.cjca.2020.11.010.

Di, J., Demanuele, C., Kettermann, A., Karahanoglu, F. I., Cappelleri, J. C., Potter, A., Bury, D., Cedarbaum, J. M., & Byrom, B. (2022). Considerations to address missing data when deriving clinical trial endpoints from Digital Health Technologies. *Contemporary Clinical Trials*, *113*, 106661. https://doi.org/10.1016/j.cct.2021.106661.

Enders, C. K. (2010). *Applied Missing Data Analysis*. The Guilford Press.

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved

Journal of Information System and Technology Management
EISSN: 0128-1666
JISTM
Volume 7 Issue 29 (December 2022) PP. 01-09
DOI: 10.35631/JISTM.729001
Eskola, R., & Handroos, H. (2013). Novel horseback riding simulator based on 6-DOF Motion Measurement, a motion base, and interactive control of Gaits. *Advanced Robotics*, *27*(16), 1249–1257. https://doi.org/10.1080/01691864.2013.824134.

Kabir, G., Tesfamariam, S., Hemsing, J., & Sadiq, R. (2019). Handling incomplete and missing data in water network database using imputation methods. *Sustainable and Resilient Infrastructure*, *5*(6), 365–377. https://doi.org/10.1080/23789689.2019.1600960.

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, *6*(3), 287. https://doi.org/10.2307/1391878.

Nakagawa, S. (2015). Missing data: mechanisms, methods and messages. In *Ecological statistics: Contemporary theory and application* (pp. 81–105). essay, Oxford University Press.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, *81*(394), 366–374. https://doi.org/10.1080/01621459.1986.10478280.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman & Hall/CRC.

Van Buuren, S. (2012). Flexible imputation of missing data. https://doi.org/10.1201/b11826.

Yahaya, A., Ramli, J., Lin, W. M., & Muhamad, Z. (2016). Hubungan Antara Tingkah Laku Keibubapaan dengan Penghargaan Kendiri di Kalangan Remaja. *J. Pendidik. Univ. Teknol. Malaysia*, *11*, 36–46.

Copyright © GLOBAL ACADEMIC EXCELLENCE (M) SDN BHD - All rights reserved