Розділ 3

МЕТОДОЛОГІЯ ТА МЕТОДИ СОЦІОЛОГІЧНИХ ДОСЛІДЖЕНЬ

UDC 316.4

Kostenko, Y. (2024). Impact of Missing Data on Data Quality in Social Research. *Sociological Studios*, 2 (25), 58–69. https://doi.org/10.29038/2306-3971-2024-02-31-31

Impact of Missing Data on Data Quality in Social Research

Yaroslav Kostenko –
PhD student, Sociology faculty, Taras Shevchenko National University of Kyiv, Ukraine

E-mail: yarosl.kostenko@gmail.com
ORCID: https://orcid.org/0009-0001-7878-5034

DOI: 10.29038/2306-3971-2024-02-31-31

Missing data is a common issue in quantitative social research that negatively affects the data quality. This article explores the consequences of missing data, outlining the potential issues it may pose and emphasizing the importance of properly addressing the missingness. It outlines the patterns of missing data, with a focus on the need to distinguish data that's Missing at Random and data that's Missing Not at Random, explaining how these patterns may affect the choice of handling methods. The article illustrates various approaches to managing missing data through a combination of hypothetical scenarios and case studies from actual research in order to showcase the application and effectiveness of various methods. It showcases the traditional methods of handling missing data, such as complete case analysis and simple imputation methods, and their limitations. Emphasizing the importance of advanced statistical techniques, the article advocates for the use of multiple imputation as a main method of choice when dealing with missing data. By providing a methodological comparison and a strategic framework for social scientists facing missing data challenges, this work provides a strategy to be employed by social scientists when dealing with missing data in order to ensure the proper data quality.

Received: November, 2024
1st Revision: November, 2024
Accepted: December, 2024

Key words: Missing Data, Data Quality, Data Imputation, Multiple Imputation.

Костенко Ярослав. Вплив відсутніх даних на якість результатів у соціологічних дослідженнях. Відсутні дані – це поширена проблема у кількісних соціологічних дослідженнях, що негативно впливає на якість результатів. У статті описано ключові проблеми, що виникають унаслідок пропущених даних, задля розв'язання яких пропущені дані мають бути коректно адресовані. Розглядаються патерни пропущених даних з акцентом на відмінності між патернами «Missing at Random» та «Missing Not at Random», оскільки ці відмінності впливають на методологію роботи з пропущеними даними. У статті використано гіпотетичні приклади й кейси реальних досліджень для ілюстрації застосування методів роботи з пропущеними даними. Розглянуто найбільш поширені методи роботи з пропущеними даними, як-от аналіз повного кейсу та одинарна імпутація, та наголошено на недоліках застосування таких методів. Закликається до використання множинної імпутації як основного методу роботи з пропущеними даними та пропонується загальна стратегія для роботи з ними, що дає можливість дослідникам-соціологам забезпечувати якість даних у кількісних соціологічних дослідженнях.

Ключові слова: пропущені дані, якість даних, імпутація даних, множинна імпутація.

INTRODUCTION

Missing data is a common occurrence in quantitative social research. It is prevalent across all kinds of studies but is particularly frequent in surveys dealing with sensitive questions, longitudinal studies where

participant dropout is an issue, or complex questionnaires that may lead to respondent fatigue and incomplete answers. This prevalence of missing data negatively affects the data quality by distorting true distributions, relationships between variables, and key statistical parameters, which reduces the validity of research findings.

Given these challenges, in this article, we explore the impact of missing data on data quality in social research and emphasize the importance of using the correct methods to handle it. We will begin by examining the limitations of traditional methods such as full case analysis and simple imputation techniques, which include replacing missing values with the mode or median of the observed data. While straightforward, these techniques can introduce biases, distort relationships between variables, and underestimate variability, reducing the overall data quality and potentially leading to incorrect conclusions.

We discuss common challenges dealing with missing data, such as dealing with non-metric scales and recognizing different patterns of missingness. After outlining these challenges and demonstrating the limitations of the traditional methods of handling missing data, we propose a set of approaches aimed at retaining data quality and ensuring that research findings can remain valid despite the missing data. The aim of this article is to provide researchers with an overview of imputation techniques and provide practical guidance on how to address missing data effectively, in order to ensure the research robustness.

1. THE CONCEPT OF DATA QUALITY

There is no uniform, single definition of data quality, as evident from various works that attempt to outline the dimensions of it, ranging from a couple of aspects to more complex structures (Carmines, 1979; Wang, & Strong, 1996; Jesiļevska, 2017; Wang et al., 2024). We'll focus on the three aspects of data quality: representativeness, accuracy, and reliability, as they are of key importance when dealing with empirical social science researches with missing data.

Representativeness refers to the extent to which data accurately reflect the broader population or general group under study. In social sciences, ensuring representativeness means including all relevant subgroups within the population to avoid bias. If certain subgroups are missed or underrepresented, the findings may not be generalizable, leading to skewed conclusions. Researchers aim to design sampling methods that capture the diversity of the population to enhance the validity of their studies. This direct impact of missing data, which often disproportionately affects certain subgroups, can severely compromise the representativeness and, in turn, the overall accuracy of study findings. It is important for social scientists to implement robust data collection strategies that handle the missing data, ensuring that their research accurately represents the population being studied.

Accuracy involves the degree to which data at the individual level (in our case, respondents) accurately represent the real-world conditions or phenomena being studied. This means that each data point should truthfully reflect the participant's experiences, opinions, or behaviors. High accuracy ensures that the collected data provide a reliable basis for analysis and interpretation. Inaccurate data can lead to incorrect conclusions, undermining the study's validity. Some error sources can be impossible to measure, making estimating accuracy difficult (Biemer, & Luberg, 2003). Krejčí (2010) emphasizes that accuracy, while not being the single dimension, is the one that is crucial for the further statistical analysis, and as such has to be of key importance to the researcher. This is the aspect where the impact of missing data manifests in forms such as non-responses or incomplete data entries, especially towards sensitive questions. Such missing information can skew the dataset, leading to inaccuracies that compromise the quality of study's conclusions.

Reliability refers to the consistency and reproducibility of the data collection process. A reliable study yields similar results when repeated under the same conditions, indicating that the data are systematically gathered. This involves standardized procedures and controls to minimize variability caused by external factors. Reliability is important for making the research replicable, allowing other researchers to verify results. Missing data compromises the reliability of research by introducing uncertainties that make the replicability of the study less feasible. Missing data may lead to inconsistent results when the study is repeated, as not all variables are consistently captured across different iterations.

To summarize, missing data impacts three fundamental dimensions of data quality in social research. First, it affects representativeness by disproportionately affecting certain subgroups, which can skew

research findings and reduce the generalizability of the results. Second, it affects accuracy as non-responses or incomplete data entries, particularly on sensitive questions, which can lead to significant inaccuracies in capturing the real-world experiences or behaviors of respondents. Finally, reliability is compromised as the inconsistencies introduced by missing data hinder the reproducibility of the research, making it difficult to replicate findings reliably under the same conditions. Addressing these challenges posed by missing data is important to ensure proper data quality and credibility of the study.

2. MISSING DATA AND ITS FORMS

Classifying the missing data is an important first step to understanding how to further proceed with it. Depending on a situation, approaches towards missing data classification may vary greatly. Classification allows us to simplify these cases and according approaches towards handling them.

The varieties of missing data are most commonly categorized into three types: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). Introduced by Rubin (1976), this classification is commonly called Rubin's Classification System.

Missing Completely at Random (MCAR). When the probability of data being missing is the same for all observations, this missing data is classified as MCAR. The reason for its absence is unrelated to the data itself or any other observed or unobserved data. For example, if a researcher loses a batch of surveys purely by accident, the missing data can be considered MCAR. Therefore, MCAR is the least demanding case of missing data, as it does not affect the distributions, however it still reduces the sample size.

Missing at Random (MAR). Under MAR, the probability of data being missing is related to some of the observed data, but not to the value of the missing data itself. MAR allows for the possibility that missingness is related to variables that you have measured. For instance, if men are less likely to report depression, and depression is the variable with missing values, the missing data is MAR if you have a gender variable in your data.

Missing Not at Random (MNAR). In case of MNAR, its likelihood of being missing is related to its value, whether observed or not. In other words, there is a systematic difference between the missing values and the observed values. For example, individuals with lower income could be less likely to report their earnings. In this case, the missing data on income is MNAR because the likelihood of the data being missing is directly related to the undisclosed income levels themselves. This non-disclosure might stem from factors such as stigma associated with lower income. Ignoring this pattern of missing data can lead to analyses that misrepresent economic behaviors and outcomes, resulting in biased conclusions. MNAR is particularly difficult to address because it often requires extra information not contained in the data itself or assumptions about why data might be missing.

In social sciences, for the most cases, data is either MAR or MNAR, which further complicates the task of working with it. The nature of quantitative social researches (especially in case of sensitive questions) often suggests there's a discrepancy between subset of respondents that did provide and answer and those that did not. A common example would be an employment status. Respondent is more likely to report that they're employed rather than unemployed, which means that we can't project the employment status distribution of those who answered this question onto those that refused to answer it. In cases such as these, a predictive model that'll intelligently impute these values based on others is required. Schouten and Vink (2021) use simulations and showcase that different missing data mechanisms, especially when the correlations are low, can lead to similar statistical inferences. The choice of the method to handle missing data and the assumptions underlying the method needs careful consideration and validation based on the observed data structure and the nature of the missingness.

While Rubin's classification is the most commonly used one, there have been proposals towards its extension. Gomer and Yuan (2021), for instance, propose differentiating between 2 sorts of MNAR data – diffuse MNAR and focused MNAR, with diffuse MNAR being influenced by both missing and available data, whereas focused MNAR is influenced only by the missing data, and suggest different approaches towards these types of missing data. For diffuse MNAR, the approach needs to capture the dependence on both observed and unobserved data, potentially requiring more data or stronger assumptions. Meanwhile, focused MNAR models can often be simpler, as they do not need to account for the influence of observed data on missingness. Graham (2009) states that it's a common misconception to view the three types of

missing data – MCAR, MAR, and MNAR – as completely distinct categories. However, the reality is that strictly categorizing missing data as MCAR, MAR, or MNAR involves assumptions that tend to not hold up in practical scenarios, as real-world data rarely fits perfectly into these 'ideal' categories. Therefore, it's more accurate to perceive missing data as existing on a spectrum between MAR and MNAR. Given that data is rarely purely MAR, the question shouldn't be about whether data is MNAR but rather understanding that all missingness leans towards MNAR to some degree, reflecting a continuum, not a discrete categorization.

Given that data rarely fits perfectly into the MAR category, it is more realistic to view missing data as existing along a continuum, with MAR and MNAR as endpoints, rather than as distinct and separate categories. Let's illustrate this continuum with two examples. In electoral research, data on voter preferences may be missing because respondents choose not to disclose their preferred candidate. However, these same respondents often still provide their opinions on related issues such as taxes, immigration, healthcare, and foreign policy. This availability of related data suggests a scenario closer to MAR, as it allows researchers to use these responses to 'predict' their voting preferences, even if not directly stated. Conversely, in surveys assessing mental health, individuals with severe depression might entirely avoid answering questions about their condition, making it unlikely to infer their mental state from other responses. This places such data closer to MNAR, as the missingness is heavily related to the severity of the condition being measured. The distinction in missing data patterns requires different approaches: for the electoral research scenario, regression model can be a feasible choice, whereas mental health studies may require pattern-mixture models that consider the likelihood of non-response linked to the underlying health issue. Understanding these differences is important because applying inappropriate methods can lead to biased outcomes, undermining the validity of the research.

Graham (2012) continues in his book by stating that while generally differentiating between MAR and MNAR is unfeasible, sometimes such assumptions are possible to be made after assessing the dataset, which, however, requires an individual approach. McKnight et al. (2007) echo this sentiment, noting the difficulty of distinguishing between MAR and MNAR data due to the absence of statistical methods to evaluate the relationship between missingness and unobserved values. They argue that researchers must base their decisions on 'sound logic'. They propose an extension to Rubin's system by outlining the three dimensions in which data can be missing. Within the context of social sciences, these dimensions can be interpreted as follows: At the individual level, data is missing when individuals or specific subgroups refuse to participate. At the variable level, missing values occur for certain questions across the dataset. At the occasion level, missing data refers to absent information in specific waves of a multi-wave sociological study. In social sciences that deal with a single wave, it can be simplified as:

1. Item Nonresponse – occurs when respondents do not provide an answer to some items on a survey or questionnaire but respond to others.

2. Unit Nonresponse – happens when there is no data for an entire survey or questionnaire for a respondent.

Understanding how to classify the missing data allows for better understanding of possible patterns, approaches and solutions towards the process of data imputation. For example, in case of unit nonresponse, the only (if any) knowledge available about the respondents is paradata, based on which they were picked as a respondent. Missing values that lean closer to MNAR type missingness may require constructing more complex predictive models than those that are closer to MAR.

Understanding how to classify missing data allows for better understanding of possible patterns, approaches and solutions towards the process of data imputation. In cases of unit nonresponse, where the only information available might be demographic details used for participant selection, devising an adequate imputation solution becomes challenging. Rather than relying solely on post-hoc techniques like imputation, researchers should proactively seek to minimize missing data during the data collection process. Additionally, strategies like weighting adjustments and model-based approaches can be used to handle the unavoidable missing data, thereby enhancing the robustness of the research findings. This is particularly true for data trending towards MNAR. For data that leans closer to MAR, simpler imputation methods might suffice, as these typically involve fewer biases related to unobserved data. However, when missing data is more likely to be MNAR, relying solely on post-hoc imputation can lead to significant biases. In such cases,

advanced model-based methods or even prevention strategies during data collection become even more important to ensure the integrity and accuracy of research findings. Identification of the nature of missing data on the MAR-MNAR spectrum is highly important for the correct choice of approach towards it, which, in turn, ensures the robustness of the research.

3. ISSUES WITH MISSING DATA IN SOCIAL SCIENCES

There are several issues that arise when working with missing data in Social Sciences. In case of data that's either MAR or MNAR, the distributions with missing data are skewed due to 'missingness' status depending on the other factors. This means that the results might be less accurate. For example, if reporting on employment status, if we calculate the percentage of employed respondents using only the data where they do provide a response, it is likely to be biased towards higher employment rate than in reality, due to higher likelihood of unemployed respondents to refuse this question. Or, in case of questions related to politics, it might lower the deviation, since the extremes might be less socially acceptable answers than the middle ones, and thus respondents might gravitate towards such answers. Overall, this might lead to inaccurate values when conducting further research or presenting the results.

Another issue revolves around sample size. Missing data reduces the effective sample size, which can reduce the statistical power of the study. This reduction in power makes it more difficult to detect a true effect when one exists, potentially leading to Type II errors (failing to reject a false null hypothesis). Therefore, missing data reduces the researcher's capability to prove statistical hypotheses.

Missing values can be an issue for a wide range of analytical procedures, namely those that are based off R-squared. Especially when working with a large set of variables, simple approaches like full case analysis may greatly reduce the sample size and potentially heavily skew the results. Simpler methods like mean imputation may heavily underestimate variance, which may cause issues in factor or cluster analysis, leading to incorrect conclusions. Imputations using regression methods may reinforce correlations, biasing the dataset.

A choice of method of handling the missing data can be important in case of estimating the number of factors in case of factor analysis. In the experiment by Goretzko (2021), multiple imputation with random forest method has shown better results and as a result, author suggests against the default use of pairwise deletion, especially when using the comparison data approach for factor retention. It has been recommended for providing more accurate estimates of dimensionality in exploratory factor analysis.

4. APPROACHES TOWARDS HANDLING OF MISSING DATA

4.1. Full Case Analysis

Full case analysis (also sometimes referred to as "listwise deletion") is the most basic approach toward the analysis of dataset with missing data where fraction of missing values is very low. This approach involves using only entries with no missing values, ignoring all with ones. Graham (2012) notes that in case of MCAR missingness, the distribution is preserved, however this is a rare occurrence in a real-world scenario of missing data. Generally, this approach is not advised for datasets with a notable fraction of missing values due to the reasons discussed in previous chapters – namely, that it ignores the presence of missing data. As such, and as stated in various works (Newman, 2014; Enders, 2010; Graham, 2012; Stavseth et al., 2019; Mirzaei et al., 2022), usage of complete case analysis is deemed viable only when the fraction of missing values is very low.

4.2. Simple Imputation Methods

Single imputation is a common approach towards missing data and involves filling in missing data with a single estimate. There are several common techniques within this approach, and the simplest ones are mean, median, and mode imputation. Mean imputation replaces missing values with the average of the observed values in the dataset. Median imputation, on the other hand, uses the middle value of the observed data, providing a better option in cases where the data is skewed, as it is less affected by outliers than the mean. Mode imputation replaces missing entries with the most frequently occurring value in the dataset, which can be particularly useful for categorical data.

While these methods are quick and easy to implement, they share common drawbacks. Each of these methods introduces a level of bias. They underestimate the variability of the dataset and distort relationships

between variables. Replacing missing values with the mean, median, or mode does not account for the natural variability in data, leading to an underestimation of the actual variance and potential biases in the analysis outcomes. Let's illustrate this on an example: a survey with missing data on respondents' incomes. If we impute missing income data by replacing it with the median income, this method reduces the income variance between respondets. It fails to represent the actual spread of high and low incomes accurately, potentially skewing any analysis related to income levels. This can lead to an underestimation of income inequality or the size of higher or lower income groups, which can affect research findings.

Another common method, particularly in longitudinal studies, is hot-deck imputation. This technique assigns missing values based on responses from 'similar' participants, often using data from the same individuals in previous survey waves. Although hot-deck imputation tends to maintain the overall distribution of data, it can still introduce biases, particularly in estimates of correlations and regression coefficients, as noted by Enders (2010). For example, if in a longitudinal study that studies income of participants, a participant's data is missing in the latest wave, hot-deck imputation might fill this gap using data from a participant with a similar age, education, profession, and work experience profile from a previous wave. While this seems reasonable, this approach may potentially ignore shifts due to various other factors.

Imputation using regression models is one of the most common approaches towards imputation. First, a predictive model for the missing variable should be constructed, one that explains the variation in responses for the missing variables well enough. This should be rooted in both theoretical understanding and available data. To measure the quality of a regression model, researchers tend to employ the R-squared coefficient, which refers to the fraction of variance explained by the model, with higher R-squared referring to more accurate models.

Once the model is constructed, predicted values for the missing data are calculated and inserted into the missing values, creating a complete dataset. The success of this approach heavily depends on the accuracy of the predictive model and the complexity of relationships within the data.

Let's revisit our earlier examples – electoral research and mental health surveys – to illustrate the practical application of regression imputation. In electoral research, missing data on voter preferences can often be accurately predicted through regression models. This is possible because respondents typically provide other related opinions (e.g., on taxes, immigration, healthcare, and foreign policy) that can serve as reliable predictors for their voting behavior, forming a reliable basis for a regression model – whose effectiveness can be evaluated by the R-squared level. However, in mental health surveys, predicting missing data on depression levels is more challenging. Often, there is insufficient related data to construct an effective predictive model. As previously noted, this type of missing data often leans closer to MNAR, making regression imputation less feasible. Therefore, while regression imputation can be effective for scenarios resembling MAR, it may not be the best approach when dealing with missingness that leans closer towards MNAR.

While regression imputation can be effective in the situations closer to MAR, when used as a single imputation technique, it leads to biases, as it tends to overestimate relationships between variables. For example, if we predict income based on age, education, profession, and work experience, it will strengthen the impact these aspects have on income level. To counteract this, a variation known as stochastic regression imputation adds random noise to each of the imputed values. Adding errors to the imputed values restores the lost data variability and reduces the bias. However, Enders (2010) notes that stochastic regression imputation increases the risk of Type I errors – where researchers incorrectly reject a true null hypothesis, which can lead to false positives.

Let's illustrate Type I error on an example. Let's consider a scenario where stochastic regression imputation is used to predict missing voter preferences based on other expressed opinions such as views on taxes, immigration, healthcare, and foreign policy. If these imputed preferences include added random noise to compensate for data variability, it might exaggerate a weak actual preference into a statistically significant finding. For example, if the original data suggested a slight, non-significant trend where individuals concerned about immigration favor a specific candidate, the added noise could amplify this trend, leading to a false conclusion that there is a strong, significant preference for this candidate among all concerned about immigration.

Newman (2014) summarizes the flaws of single imputation as two key aspects. First of all, it leads to biases, even if data is MCAR. For example, regression imputation overestimates the correlations between variables and underestimates variance. This issue can be partially solved by introducing random errors as discussed in the previous paragraph, however, secondarily, single imputation is unable to accurately calculate standard errors for hypothesis testing, as no single sample size n effectively matches all parameter estimates. This issue is compounded by the tendency of researchers to treat the imputed dataset as complete, which results in underestimated standard errors and increases the risk of Type I errors. Multiple imputation addresses this challenge by providing a more reliable method for handling missing data.

Another method worth noting is maximum likelihood imputation. It offers a probabilistic approach to estimating missing values based on observed data distributions, providing more accurate estimates than simpler methods previously mentioned. However, studies, such as those by Little and Rubin (1989), and Cox et al. (2014), have shown that while maximum likelihood can produce less biased results than simplistic imputation methods, multiple imputation tends to provide more accurate estimations for complex relationships between variables. Therefore, we recommend using multiple imputation over maximum likelihood.

4.3. Multiple Imputation

Multiple imputation allows for more accurate and efficient estimation of parameters in the presence of missing data, compared to other methods such as listwise deletion or single imputation, as noted in various works, e.g. Little and Rubin (1989), Newman (2014). Multiple imputation methods tend to provide estimates closer to those derived from complete datasets compared to other methods, in particular, at higher missing data rates (Nartgun, & Sahin Kursad, 2016).

Multiple imputation involves creating multiple plausible values for each missing data point based on the observed data and a set of assumptions about the missing data mechanism. These assumptions typically relate to the reasons why data might be missing – whether missing values depend on observed data, unobserved data, or are entirely random. For example, if missing responses about voting reasons are likely to depend on a respondent's stance on various issues such as economy or healthcare, the model might assume that these characteristics can help estimate the missing values. After multiple datasets with imputed values are created, they are analyzed separately, and the results are then combined to produce a final estimate that accounts for the uncertainty introduced by the missing data.

One of the notable advantages of multiple imputation compared to the simpler methods such as maximum likelihood is that it can work with any type of statistical models, and not just basic ones such as linear and log-linear models (Tufiş, 2008). This can be beneficial when dealing with complex datasets as it allows to use models that can capture the nuances of the missing data more accurately, such as hierarchical models.

Multiple imputation has been widely endorsed and employed effectively across various fields, particularly in studies involving sensitive issues. For instance, Skafida et al. (2022) utilized multiple imputation to provide a more realistic estimate of domestic violence, a typically underreported issue, by incorporating data from prior survey waves. Similarly, Penn (2007, 2009) applied multiple imputation to address missing income data, revealing significant influences on individual economic outcomes such as parental well-being and the impact of household size, which were underrepresented in analyses using complete cases. Additionally, Chen and Fu (2015) used multiple imputation to assess income inequality more accurately and calculate a more representative Gini index, particularly highlighting the economic status of low-income households. In all these cases, multiple imputation has been proven to be an efficient tool both when dealing with distributions and when researching the connections between variables, such as impact of parental well-being on economic outcomes of an individual.

However, some works (e.g. (Gorard, 2020)) criticize multiple imputation for operating under MAR basis, and claim than when data is closer to MCAR it may actually produce greater biases. Little and Rubin (1989) warn that complex multiple imputation models might be difficult to apply with large datasets, and in case of complex relationships between variables, with a large amount of interdependency, simpler models may perform better. As an example, they provide the imputation procedure of income for the Current Population Survey conducted in Britain, which, under the explicit model, may require to model a large number of variables with complex connections and account for correlations between other variables like

family member incomes and income types. The used alternative, hot deck imputation, while generally a simple and unreliable method, is considered to be a reasonable alternative.

The most common classification of approaches towards multiple imputation of missing data are Joint Modeling (JM) and Fully Conditional Specification (FCS).

Joint Modeling is an approach where a single statistical model is used to generate imputations for all variables with missing data simultaneously. This method assumes that there is a joint distribution underlying the variables, allowing for a unified approach to imputation. For example, if there are multiple missing variables such as a preferred election candidate and election candidate voted last time, have missing data, JM would specify one model that captures the relationships among all these variables and uses this model to impute all missing values at once.

In contrast, Fully Conditional Specification approach imputes missing data separately for each variable, taking in account multivariate relationship. This method, also known as Multiple Imputation by Chained Equations (MICE), cycles through each variable missing data, iteratively refining the imputations through multiple rounds until the changes between rounds are minimal.

While FCS may appear like a superior to JM method, in real use cases, both of these methods show comparable results. In the experiment by Grund et. al. (2017), the effectiveness of these two imputation methods was evaluated through an experimental generation of missing data (both MAR and MCAR) and compared to the simpler methods such as listwise deletion. The findings shown that both advanced methods exhibited superior performance in managing multilevel data missingness, surpassing the simpler methods. While the FCS approach was recommended due to its minimal root mean square error (RMSE) in simulations, JM was also recognized as a viable option, with both methods showing lower biases and RMSE values compared to the rest.

Despite multiple imputation being known as an effective method for handling missing data, it is not clear that the method will be efficient when data contain a high percentage of missing observations for a variable. The higher the missing fraction, the more issues could arise during the imputation process. In particular, Lee and Huber (2011) explores the efficiency of multiple imputation for data with 10 % to 80 % missing observations using absolute bias and mean squared error. In his study, the author concludes that while multiple imputation produces less biased estimates), imputation of MAR/MNAR data already becomes a non-trivial process with 20 % missing data, requiring an individual approach. In case of social sciences, for the scenarios of the severe missing data, we can suggest data re-collection, if possible.

5. IMPUTATION OF NON-METRIC SCALES

With the large fraction of data in social sciences being non-metric, it sets limitations on applicable methods. While ordinal scales can be treated as quasimetric, thus allowing use of methods typically reserved for metric data, this approach may not be universally appropriate, particularly in instances where the set of alternatives is limited (for example, only three options). Moreover, employing methods suited for metric scales is completely unfeasible for handling nominal or dichotomous data.

For ordinal data, some of the better-performing methods, according to experiment performed by Wu, Jia, and Enders (2015) are Normal Model Approach: an approach that treats the missing ordinal data as if they were drawn from a normal distribution, and Latent Variable Approach: a method that assumes that each observed ordinal variable is a manifestation of an underlying continuous latent variable. They found that the latent variable model and the normal model approach without rounding generally performed better than other methods in terms of producing smaller standardized biases, lower mean squared errors (MSE), and better confidence interval coverage.

Carpita and Manisera (2011) also explore the imputation of ordinal data on Likert scale, and propose a method by combining Approximate Bayesian Bootstrap technique with Propensity score method. Their method consists of 4 steps: Logistic Regression Step; Propensity Score Step, which computes the probability of being a nonrespondent based on the observed predictors for each respondent; Nearest Neighbour Step, which selects a subset of respondents based on their similarity in response patterns to the nonrespondent, forming a donor pool; and Approximate Bayesian Bootstrap Step, which chooses a donor case from the donor pool.

Regarding the dichotomous data, Ge et al. (2023) compare the performance of various imputation methods for handling missing data in dichotomous variables through simulation and real-data validation. The methods evaluated include traditional statistical methods (mode, logistic regression, and multiple imputation using Bayesian interpolation) and machine learning methods (decision trees, random forest, k-nearest neighbor, support vector machine, and artificial neural network). The study found that machine learning-based methods, particularly the support vector machine, artificial neural network, and decision trees, achieved relatively high accuracy and stable performance across various scenarios, and recommended them for use in case of missing dichotomous data. However, there has been limited practical application of machine learning methods in social sciences, suggesting that this area requires further research.

6. APPLICATION OF WEIGHTING AND PARADATA

Some of the notable techniques that can enhance the data quality when dealing with missing data are application of weighting and paradata variables. As such, we'll provide a quick overview of these strategies.

Weighting is a strategy to address potential non-response bias for unit non-response. With non-response weighting, whether it is a weighting class adjustment method or a response propensity weighting method, survey respondents are assigned a weight to compensate for their differential probability of participation given selection into the sample. Weighting is often used as an additional method to handle the missing data, in case of uneven chance of values to be missing across the dataset. Vandecasteele and Debels (2007) attempt to evaluate the effectiveness of weighting in longitudinal researches due to dropout, and their findings suggest that while weighting does reduce bias, it's a good practice to construct weights based off the powerful dropout predictors, as those that are based mostly about socio-demographic characteristics are not as effective.

In addressing missing data in social science, while weighting is a frequently adopted strategy, it can also increase survey error rates. Peytchev (2012) conducts an analysis contrasting the efficiency of weighting and multiple imputation methodologies in managing missing data. His findings reveal that the application of multiple imputation not only enhances the precision of the estimates but also diminishes their variance when compared with weighting strategies or complete case analysis.

Paradata – data that is collected during the process of conducting a survey – can be also employed when dealing with missing data, both for weighting and as an additional predictor for the imputation procedure. The most common use of paradata at this stage is to employ demographics-based weighting. Although using paradata to correct measurement errors is less developed, there have been conducted some researches that employ item-level paradata in order to understand the question-and-answer process, and potentially improve measurement accuracy and data analysis (Kreuter et al., 2010). There are numerous studies, e.g. (Couper, & Kreuter, 2013), (Da Silva et al., 2016), that highlight the importance of paying attention to the paradata variables that indicate the 'response quality', such as response time, or assessment of the response quality by an interviewer.

Paradata can be particularly helpful for the longitudinal researches, where missing data is a more frequent issue. One of such applications was already previously mentioned while showcasing work of Skafida et al. (2022) by leveraging paradata in order to predict values for the sensitive questions, dealing with domestic violence. Another example is a research by Brunton-Smith and Tarling (2017), which employs multilevel multiple imputation with application of paradata to handle both the item-level nonresponse and unit-level nonresponse when studying the impact of prison education programs on reoffending rates. Chen et al. (2017) explore the role of paradata when working with different survey modes – online and offline – which affect the responses, and propose using paradata related to the past surveys when correcting for unit non-response and sample selection.

While both weighting and paradata can be used when dealing with missing data, their application is more context-specific and less universal. Weighting should be considered specifically for the longitudinal studies with identifiable powerful dropout predictors. We recommend integrating paradata when it covers an important aspect not present in the questionnaire, as a part of a predictive model. An example of such an aspect might be economic conditions, such as the 'living conditions' paradata variable that's recorded in the European Social Survey.

CONCLUSIONS AND DISCUSSION

Missing data is a common problem when dealing with the empirical social science researches. While sometimes it is possible to minimize it at the research design phase, in many cases, especially those that deal with sensitive questions and socially acceptable answers, it is guaranteed to occur. For such cases, there's a need for a robust methodology to handle them. Missing data is not a new problem, but the one, we believe, that does not get enough attention from the researches. This paper highlights the significant impact missing data may have on research validity, affecting data quality. On theoretical level, missing data affects the following aspects of data quality: representativeness, accuracy, and reliability. On practical level, this results into skewed distributions, lower sample sizes, limitations posed on data analysis methods, and, in some cases, a complete invalidation of analysis methods. As such, we call for proper addressing of these issues.

Such procedure consists of a couple of steps. Tackling missing data issues should start with the identification of missing data – the type, the amount, and the potential complexity of connections the missing data may have. In real-world scenarios, most of the missing data ranges on the spectrum between MAR and MNAR, in other words, on the spectrum of being related to the available and unavailable data. As such, the more data leads towards MNAR, the more complex models are required. Therefore, as a first step of handling the missing data cases is the understanding of the connections of the data unavailable, which can be done based off theoretical implications. An extra attention should be paid to the fraction of missing values, as the researches show, with 20 % of missing MAR/MNAR data, imputation already becomes a non-trivial process. If missing data exceeds this threshold or the connections between variables are deemed too complex to accurately predict them, a better approach might be re-collection of data, if possible.

Identification of missing data should be followed by the construction of a predictive model. In general case scenario, it involves construction of a predictive model with a feasible predictive power (which can be evaluated by R-squared when dealing with quasimetric scales). Construction of such model should be rooted both in theoretical implications and available data, and it is important to not overfit the model, as it may introduce bias in terms of variable connections.

The final step involves the method of choice. Our study provides an overview of various methods dealing with missing data. In conclusion, we advise against usage of simple methods of handling missing data such as Full Case Analysis, mode and mean imputation, hot-deck and single imputation using regression when there's a non-negligible amount of missing data. As illustrated with examples, these methods tend to negatively affect data quality by distorting both the distributions and connections between variables. As a robust and reliable method, for the general case, we suggest using multiple imputation using either joint modelling or fully conditional specification, as these methods tend to minimize the biases in introduced values. Nowadays, various software packages (SPSS, Stata) and programming languages (R, Python) allow for implementation of multiple imputation, making the procedure more accessible to the researchers.

However, while we've outlined the general course of action, it is by no means exhaustive. For example, dealing with nominal scales requires a different approach, with some of the better performing methods being machine learning ones such as support vector network and decision trees, which require a different implementation. Some of the potential problems that arise with the missing data have not been solved yet – such as limitations of multiple imputation for the data leaning heavily towards MNAR, dealing with non-numeric data such as graphs, or ensuring the robustness of imputation procedure in longitudinal researches. While we've attempted to highlight the variety of problems that arise from dealing with missing data and importance of dealing with them, we admit that these are not fully solved and invite further discussions and research on how to properly deal with them.

In our paper, we have outlined the prevalent challenges and evolving methodologies to address missing data in social sciences. While we have outlined current approaches, the rapid advancement in technology beckons future researchers to explore innovative solutions. Particularly, we call for the creative approaches that adapt new types of data (such as paradata) and analytical tools (such as machine learning algorithms, which we've touched upon when discussing approaches towards handling missing non-metric data) that become more potent and widespread. We must encourage researchers to not only adopt robust techniques but also to pioneer the evaluation and refinement of these emerging tools. By doing so, they can ensure that social science research remains adaptive and at the cutting edge of methodological advancements.

REFERENCES

Biemer, P. P., & Lyberg, L. (2003). *Introduction to survey quality*. Hoboken, NJ: Wiley.

Brunton-Smith, I., & Tarling, R. (2017). Harnessing paradata and multilevel multiple imputation when analysing survey data: A case study. *International Journal of Social Research Methodology,* 20(6), 709–720. https://doi.org/10.1080/13645579.2017.1287842

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage Publications.

Carpita, M., & Manisera, M. (2011). On the Imputation of Missing Data in Surveys with Likert-Type Scales. *Journal of Classification, 28*(1), 93–112. https://doi.org/10.1007/s00357-011-9074-z

Chen, H., Dunbar, G., & Shen, Q. R. (2017). The Mode is the Message: Using Predata as Exclusion Restrictions to Evaluate Survey Design. *Bank of Canada Staff Working Paper 2017-43*. Retrieved November 10, 2024 from https://www.bankofcanada.ca/wp-content/uploads/2017/10/swp2017-43.pdf

Chen, Y., & Fu, D. (2015). Measuring income inequality using survey data: The case of China. *Journal of Economic Inequality,* 13, 299–307. https://doi.org/10.1007/s10888-014-9283-x

Couper, M. P., & Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 176(1), 271–286. https://doi.org/10.1111/j.1467-985X.2012.01041.x

Cox, B. E., McIntosh, K., Reason, R. D., & Terenzini, P. T. (2014). Working with Missing Data in Higher Education Research: A Primer and Real-World Example. *The Review of Higher Education,* 37(3), 377–402. https://doi.org/10.1353/rhe.2014.0026

Da Silva, D. N., Skinner, C., & Kim, J. K. (2016). Using Binary Paradata to Correct for Measurement Error in Survey Data Analysis. *Journal of the American Statistical Association,* 111(514), 526–537. https://doi.org/10.1080/01621459.2015.1130632

Enders, C. K. (2010). *Applied missing data analysis*. NY: Guilford Press.

Ge, Y., Li, Z., & Zhang, J. (2023). A simulation study on missing data imputation for dichotomous variables using statistical and machine learning methods. *Scientific Reports, 13*, 9432. https://doi.org/10.1038/s41598-023-36509-2

Gomer, B., & Yuan, K.-H. (2021). Subtypes of the missing not at random missing data mechanism. *Psychological Methods, 26*(5), 559-598. https://doi.org/10.1037/met0000377

Gorard, S. (2020). Handling missing data in numeric analyses. *International Journal of Social Research Methodology,* 23(6), 651–660. https://doi.org/10.1080/13645579.2020.1729974

Goretzko, D. (2021). Factor retention in exploratory factor analysis with missing data. *Educational and Psychological Measurement, 82*(3), 444–464. https://doi.org/10.1177/00131644211022031

Graham, J. W. (2009). Missing data analysis: Making it work in the real world. *Annual Review of Psychology,* 60, 549–576. https://doi.org/10.1146/annurev.psych.58.110405.085530

Graham, J. W. (2012). *Missing data: Analysis and design*. Springer.

Grund, S., Lüdtke, O., & Robitzsch, A. (2017). Multiple imputation of missing data for multilevel models: Simulations and recommendations. *Organizational Research Methods,* 21(1), 111–149. https://doi.org/10.1177/1094428117703686

Jesiļevska, S. (2017). Data quality dimensions to ensure optimal data quality. *The Romanian Economic Journal,* 20(63), 89–103. Retrieved November 10, 2024 from https://ideas.repec.org/a/rej/journl/v20y2017i63p89-103.html

Krejčí, J. (2010). Approaching quality in survey research: Towards a comprehensive perspective. *Czech Sociological Review,* 46(5), 1011–1033. Retrieved November 10, 2024 from https://sreview.soc.cas.cz/pdfs/csr/2010/06/06.pdf

Kreuter, F., Couper, M., & Lyberg, L. (2010). The use of paradata to monitor and manage survey data collection. In *Section on Survey Research Methods – JSM 2010*. Retrieved November 10, 2024 from http://sampieuchair.ec.unipi.it/wp-content/uploads/2018/10/Couper-et-al.pdf

Lee, J. H., & Huber Jr., J. (2011). Multiple imputation with large proportions of missing data: How much is too much? In *Proceedings of the 23rd United Kingdom Stata Users' Group Meetings*. Stata Users Group.

Little, R. J. A., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research,* 18(2–3), 292–326. https://doi.org/10.1177/0049124189018002004

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing Data: A Gentle Introduction*. Guilford Press.

Mirzaei, A., Carter, S. R., Patanwala, A. E., & Schneider, C. R. (2022). Missing data in surveys: Key concepts approaches and applications. *Research in Social and Administrative Pharmacy, 18*, 2308–2316. https://doi.org/10.1016/j.sapharm.2021.03.009

Nartgun, Z., & Sahin Kursad, M. (2016). Comparison of the various methods used in solving missing data problems. *The Anthropologist,* 24(1), 380–388. https://doi.org/10.1080/09720073.2016.11892028

Newman, D. A. (2014). Missing data: Five practical guidelines. *Organizational Research Methods,* 17(4), 372–411. https://doi.org/10.1177/1094428114548590

Penn, D. A. (2007). Estimating missing values from the General Social Survey: An application of multiple imputation. *Social Science Quarterly*, 88(2), 573–595. https://doi.org/10.1111/j.1540-6237.2007.00472.x

Penn, D. (2009). Financial well-being in an urban area: An application of multiple imputation. *Applied Economics*, 41(23), 2955–2964. https://doi.org/10.1080/00036840701367507

Peytchev, A. (2012). Multiple Imputation for Unit Nonresponse and Measurement Error. *Public Opinion Quarterly*, *76*(2), 214–237. https://doi.org/10.1093/poq/nfr065

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Schouten, R. M., & Vink, G. (2021). The dance of the mechanisms: How observed information influences the validity of missingness assumptions. *Sociological Methods & Research*, *50*(3), 1243–1258. https://doi.org/10.1177/0049124118799376

Skafida, V., Morrison, F., & Devaney, J. (2022). Answer refused: Exploring item non-response on domestic abuse questions in a social survey affects analysis. *Survey Research Methods*, *16*(2), 227–240. https://doi.org/10.18148/srm/2022.v16i2.7823

Stavseth, M. R., Clausen, T., & Røislien, J. (2019). How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data. *SAGE Open Medicine*, 7, 1–12. https://doi.org/10.1177/2050312118822912

Tufiş, C. D. (2008). Multiple imputation as a solution to the missing data problem in social sciences. *Calitatea vieţii, 1-2*, 199-212. Retrieved November 10, 2024 from https://www.ceeol.com/search/article-detail?id=80322

Vandecasteele, L., & Debels, A. (2007). Attrition in panel data: The effectiveness of weighting. *European Sociological Review*, *23*(1), 81–97. https://doi.org/10.1093/esr/jcl021

Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S., & Aggarwal, S. (2024). Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. *Journal of the Knowledge Economy*, 15(1159–1178). https://doi.org/10.1007/s13132-022-01096-6

Wang, R. Y., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–34. Retrieved November 10, 2024 from http://mitiq.mit.edu/Documents/Publications/TDQMpub/14_Beyond_Accuracy.pdf

Wu, W., Jia, F., & Enders, C. (2015). A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables. *Multivariate Behavioral Research*, 50(5), 484–503. https://doi.org/10.1080/00273171.2015.1022644