AMERICAN ACADEMIC PUBLISHER

INTERNATIONAL JOURNAL OF MATHEMATICS AND STATISTICS (ISSN: 2693-3594)

Volume 05, Issue 01, 2025, pages 09-14
Published Date: - 01-08-2025

Integrating Survey Data through Matched Mass Imputation: A Comprehensive Approach

Dr. Sofia I. Petrov
PhD, Department of Applied Mathematics and Statistics, Moscow State University, Russia

Abstract

The increasing availability of non-probability samples, often collected rapidly and cost-effectively (e.g., through web surveys), presents both opportunities and challenges for statistical inference. While probability samples remain the gold standard for unbiased estimation, their cost and declining response rates necessitate innovative methods for integrating data from diverse sources. This article explores Matched Mass Imputation (MMI) as a robust and efficient approach for combining information from a traditional probability sample with a larger, auxiliary non-probability sample. We detail the methodological framework of MMI, which leverages matching techniques to identify suitable donors from the non-probability sample for recipients in the probability sample, followed by mass imputation of unobserved variables. This approach aims to mitigate biases inherent in non-probability samples and enhance the precision of estimates by effectively utilizing the larger sample size. We discuss the theoretical underpinnings, practical implementation considerations, and the conditions under which MMI can yield reliable inferences, including the crucial common support assumption and the role of statistical learning methods. By synthesizing recent advancements, this paper demonstrates MMI's potential to provide a powerful and flexible solution for modern survey data integration, balancing the need for accuracy with the realities of data collection in an evolving landscape.

Keywords

Survey data integration, matched mass imputation, data harmonization, missing data handling, statistical imputation methods, data fusion, survey methodology, comprehensive data analysis.

INTRODUCTION

In contemporary statistical practice, the landscape of data collection is rapidly evolving. Traditional probability samples, characterized by known inclusion probabilities for each unit, are considered the gold standard for producing unbiased and statistically rigorous inferences about a target population [16, 21]. However, these surveys are increasingly expensive to conduct, often suffer from declining response rates, and can be time-consuming [1, 18]. Concurrently, the proliferation of digital platforms and online panels has led to the widespread availability of non-probability samples. These samples, while often large, cost-effective, and quickly accessible [18], inherently lack the probabilistic foundation of traditional surveys, making direct inference from them susceptible to significant selection bias [1, 2].

The challenge and opportunity lie in effectively integrating these disparate data sources to leverage the strengths of each: the inferential rigor of probability samples and the size and cost-efficiency of non-probability samples [1, 12, 28]. Data integration aims to combine information from multiple sources to produce more accurate, precise, or comprehensive estimates than could be achieved from any single source alone [29]. Various methods have been proposed for this purpose, including weighting adjustments (e.g., propensity score weighting) [11, 14, 19], calibration, and direct imputation [4, 5, 12, 30, 31].

Among these methods, mass imputation has emerged as a promising technique for data integration [4, 12, 30, 31]. In mass imputation, a non-probability sample (often termed the "source" or "donor" sample) is used to impute missing values or unobserved variables into a probability sample (the "target" or "recipient" sample). This approach effectively "transfers"

information from the larger non-probability sample to the smaller, more representative probability sample, thereby enriching the latter and potentially improving the precision of estimates [4, 12].

This article focuses on a specific and robust variant: Matched Mass Imputation (MMI). MMI combines the principles of mass imputation with advanced matching techniques. The core idea is to identify individuals in the non-probability sample who are "similar" to individuals in the probability sample based on a set of common auxiliary variables. Once matched, the unobserved variables from the non-probability sample are imputed onto their matched counterparts in the probability sample. This matching step is crucial for addressing the selection bias inherent in non-probability samples by ensuring that the imputed data comes from donors who are comparable to the recipients in the probability sample [2, 12, 23].

The objective of this article is to provide a comprehensive overview of Matched Mass Imputation for survey data integration. We will delve into its methodological underpinnings, discuss the critical assumptions required for valid inference, explore the role of statistical learning techniques in its implementation, and highlight its advantages and limitations. By doing so, we aim to demonstrate MMI as a powerful and flexible approach for navigating the complexities of modern survey data, offering a pathway to more reliable and efficient statistical estimates in an era of diverse and often biased data sources.

METHODS

The Matched Mass Imputation (MMI) framework for survey data integration involves several key steps, combining principles from imputation, matching, and survey sampling theory. This section details the conceptual and methodological components of MMI.

1 Data Sources and Notation

We consider two primary data sources:

Probability Sample (P-sample): This is a traditional survey sample drawn from the target population using a known probability sampling design (e.g., simple random sampling, stratified sampling). It provides unbiased estimates of population parameters, but may be small or have limited variables [16, 21]. Let SP denote the P-sample, with nP units. For each unit $i \in SP$, we observe a set of auxiliary variables Xi and a variable of interest Yi. However, there might be other variables Zi that are not observed in SP but are available in the non-probability sample.

Non-Probability Sample (NP-sample): This is a larger, auxiliary sample collected without a known probability sampling mechanism (e.g., online panel, administrative data, big data sources). It typically contains the auxiliary variables Xj and the variables of interest Yj, as well as the unobserved variables Zj that we wish to impute into the P-sample [1, 28]. Let SNP denote the NP-sample, with nNP units.

The goal of MMI is to estimate population parameters (e.g., means, totals, regression coefficients) for variables that are only available in the NP-sample, by imputing them into the P-sample.

2 Core Principle: Mass Imputation

Mass imputation is a general strategy where a variable (or set of variables) observed in a donor dataset is imputed into a recipient dataset where it is unobserved [4, 12, 30]. Unlike traditional missing data imputation (where values are missing within a single dataset), mass imputation deals with variables that are entirely unobserved in one dataset but fully observed in another. The fundamental assumption for valid mass imputation is that, conditional on the observed auxiliary variables X, the distribution of the unobserved variable Z is the same in both the P-sample and the NP-sample. This is often referred to as the "missing at random" (MAR) assumption, adapted for data integration [15, 20].

3 The Role of Matching

The key innovation of MMI lies in the matching step, which precedes or is integrated with the imputation process [12, 23]. The purpose of matching is to reduce the selection bias inherent in the NP-sample by ensuring that for each unit in the P-sample (recipient), a "similar" unit (donor) is found in the NP-sample based on common auxiliary variables X. This creates a pseudo-probability sample from the NP-sample that is more comparable to the P-sample.

Common matching techniques include:

Propensity Score Matching: This is a widely used method where the propensity score, $e(Xi)=P(\text{unit } i \in SP|Xi)$, is estimated for each unit using logistic regression or other statistical learning methods [11, 23]. Units from the NP-sample are then matched to units in the P-sample based on similar propensity scores. This helps to balance the distributions of X between the two samples [11, 14].

Nearest Neighbor Matching: For each unit in the P-sample, the closest unit(s) in the NP-sample are identified based on a distance metric computed from the auxiliary variables X [23, 24].

Kernel Weighting: A related approach that assigns weights to NP-sample units based on their similarity to P-sample units, often using kernel functions [11, 26]. This can be seen as a form of "soft matching."

The matching step is crucial for satisfying the conditional independence assumption required for valid imputation. It attempts to achieve the "common support" assumption, meaning that for every unit in the P-sample, there exists a similar unit in the NP-sample [6].

4 Matched Mass Imputation Procedure

The MMI procedure can be summarized as follows:

Identify Common Auxiliary Variables (X): Both the P-sample and NP-sample must share a set of common auxiliary variables X that are predictive of the variables of interest Y and Z, and also predictive of sample membership (i.e., whether a unit

belongs to the P-sample or NP-sample).
Matching (or Weighting) Step:
Propensity Score Estimation: Estimate the propensity score for each unit, representing the probability of being in the P-sample given X. Statistical learning methods like boosted kernel weighting [11] or generalized additive models (GAMs) [29] can be employed to model this relationship, especially with high-dimensional X [14].
Matching/Weighting: Match each unit in the P-sample to one or more units in the NP-sample based on their propensity scores or other similarity measures. Alternatively, assign weights to NP-sample units based on their propensity scores to make them representative of the P-sample.
Common Support Check: Verify that there is sufficient overlap in the distribution of X (or propensity scores) between the P-sample and the NP-sample [6]. Units outside the common support region in the NP-sample may need to be excluded.
Imputation Step: For each unit i in the P-sample, identify its matched donor(s) from the NP-sample. The unobserved variable Zi in the P-sample is then imputed using the observed Z values from its matched donor(s).
Imputation Method: This can range from simple mean imputation (using the mean of Z from matched donors) to more sophisticated methods like regression imputation (predicting Z based on X from matched donors) or hot-deck imputation (randomly selecting a Z value from a matched donor) [4]. Nonparametric mass imputation methods can also be used [4].
Inference: Once the P-sample is augmented with the imputed variables, standard survey estimation techniques (e.g., Horvitz-Thompson estimator [8], generalized regression estimator [21]) can be applied, taking into account the original survey weights of the P-sample and incorporating variance estimation methods that account for the imputation process [4, 12, 30, 31].
5 Assumptions for Valid Inference
The validity of MMI relies on several critical assumptions:
Conditional Independence (MAR-like): The distribution of the unobserved variable Z in the NP-sample is the same as in the P-sample, conditional on the observed auxiliary variables X. That is, $Z \perp \text{sample type} | X$. This is a strong assumption, and the matching step aims to make it more plausible by creating comparable groups [15, 20].
Common Support: For every unit in the P-sample, there exists a comparable unit in the NP-sample based on the auxiliary variables X [6]. If this assumption is violated, extrapolation may be required, which can introduce bias.
Correct Model Specification: If propensity scores or imputation models are used, they must be correctly specified. Statistical learning methods like random forests [17] or generalized additive models [29] can help in modeling complex relationships in X to improve propensity score estimation and imputation accuracy [11, 14].
2.6 Variance Estimation
A crucial aspect of MMI is proper variance estimation. Simply treating the imputed values as observed data will underestimate the true variance. Methods for variance estimation in mass imputation include:
Bootstrap or Jackknife: Resampling techniques that account for both sampling variability and imputation variability [4].
Analytic Variance Formulas: Deriving specific formulas that incorporate the uncertainty introduced by the imputation process [4, 12, 30].
By carefully implementing these steps and considering the underlying assumptions, MMI offers a powerful framework for integrating diverse survey data.

RESULTS (Performance and Applications)

The application of Matched Mass Imputation (MMI) has demonstrated significant potential in enhancing the quality of inferences drawn from integrated survey data, particularly when combining probability and non-probability samples. Empirical studies and theoretical analyses have highlighted its effectiveness in bias reduction and precision improvement.
1 Bias Reduction in Non-Probability Samples
A primary objective of MMI is to mitigate the selection bias inherent in non-probability samples [1, 2]. By matching units from the non-probability sample to those in the probability sample based on common auxiliary variables, MMI effectively creates a subset of the non-probability sample that is more representative of the target population.
Propensity Score Effectiveness: Research consistently shows that propensity score matching, a core component of MMI, is effective in balancing covariates between the probability and non-probability samples [11, 14, 23]. This balancing reduces the bias in estimates of variables of interest that are transferred from the non-probability sample [5, 12]. For instance, studies comparing different data integration methods have found that approaches incorporating matching or weighting based on propensity scores yield less biased estimates than simpler methods that ignore the selection bias [5, 12, 28, 30].
Doubly Robust Properties: Some implementations of mass imputation, particularly those that combine matching/weighting with a robust imputation model, exhibit "doubly robust" properties [5, 31]. This means that the estimator remains consistent if either the propensity score model or the imputation model is correctly specified, providing a safeguard against model misspecification and further contributing to bias reduction.
2 Precision Enhancement
Beyond bias reduction, MMI leverages the larger size of the non-probability sample to improve the precision of estimates.
Increased Effective Sample Size: By imputing unobserved variables from a large NP-sample into a smaller P-sample, MMI effectively increases the amount of information available for analysis, leading to estimates with smaller variances [4, 12, 30]. This is particularly beneficial when the P-sample is small or when estimating parameters for rare subgroups.

Comparison with Other Methods: Studies comparing MMI with methods like direct weighting (where only the NP-sample is weighted to match the P-sample) show that MMI can achieve comparable or even superior precision, especially when the imputation model is strong [4, 12, 30]. The ability to impute a full set of variables allows for more flexible downstream analyses than just weighting.

3 Role of Statistical Learning Methods

The effectiveness of MMI is often enhanced by the judicious use of statistical learning methods in both the matching and imputation stages.

Improved Propensity Score Estimation: When dealing with a large number of auxiliary variables or complex relationships, traditional logistic regression for propensity score estimation may be insufficient. Methods like boosted kernel weighting [11], random forests [17], or generalized additive models (GAMs) [29] can capture non-linear relationships and interactions among covariates more effectively, leading to better-estimated propensity scores and, consequently, more accurate matching [11, 14]. This is particularly relevant for high-dimensional data [14, 31].

Nonparametric Imputation: Nonparametric mass imputation methods, which do not rely on strong parametric assumptions about the relationship between X and Z, can be more robust to model misspecification [4]. These methods often leverage concepts from machine learning to find suitable donors.

4 Practical Applications and Case Studies

MMI and related data integration techniques have found practical application in various domains:

Public Health Surveys: For instance, in the context of the National Health and Nutrition Examination Survey (NHANES) [3], where detailed health data from a probability sample could be augmented with information from large administrative datasets or non-probability health panels to estimate disease prevalence or risk factors more precisely [14].

Social Science Research: Integrating data from traditional demographic surveys with online panel data to study social trends or public opinion [2, 28].

Official Statistics: National statistical offices are increasingly exploring data integration techniques to supplement traditional surveys with administrative records or big data sources, aiming for more timely and granular statistics while maintaining data quality [18].

While MMI offers significant advantages, its success hinges on the careful selection of auxiliary variables, validation of the common support assumption, and appropriate variance estimation [1, 6, 12].

DISCUSSION

The growing demand for timely and cost-effective statistical insights, coupled with the challenges facing traditional probability surveys, underscores the critical importance of robust data integration methodologies. Matched Mass Imputation (MMI) stands out as a powerful and flexible approach for combining the inferential strength of probability samples with the scale and efficiency of non-probability samples.

1 Strengths of Matched Mass Imputation

MMI offers several compelling advantages for survey data integration:

Bias Reduction: By employing matching techniques (e.g., propensity score matching) based on common auxiliary variables, MMI directly addresses the selection bias inherent in non-probability samples [2, 11, 12, 23]. This is crucial for making valid inferences about the target population, as simply weighting or imputing without careful matching can perpetuate bias [1, 6].

Precision Enhancement: The ability to impute variables from a large non-probability sample into a smaller, representative probability sample effectively increases the information content available for analysis [4, 12, 30]. This leads to more precise estimates, particularly for variables that are rare or for analyses requiring fine-grained demographic breakdowns.

Flexibility in Analysis: Once the probability sample is augmented with imputed variables, analysts can perform a wide range of statistical analyses (e.g., regression, subgroup analysis) as if the data were fully observed, using standard statistical software [4]. This is often more flexible than methods that only provide adjusted weights.

Doubly Robust Properties: As noted in the results, some MMI variants can achieve double robustness, providing a safeguard against misspecification of either the matching model or the imputation model [5, 31]. This makes the method more resilient in practice.

Utilization of Statistical Learning: The framework naturally accommodates advanced statistical learning methods for propensity score estimation and imputation, allowing for the capture of complex relationships in high-dimensional data and potentially improving the accuracy of both matching and imputation [11, 14, 17, 29].

2 Challenges and Limitations

Despite its strengths, the successful implementation of MMI is contingent upon addressing several methodological and practical challenges:

Strong Assumptions: The core assumption of conditional independence (or MAR-like assumption) is fundamental. This means that, after controlling for common auxiliary variables X, the non-probability sample must be conditionally representative of the target population with respect to the unobserved variables Z [15, 20]. If important confounding variables are not available in X, residual bias may persist.

Common Support: The common support assumption is critical [6]. If there are segments of the probability sample that have no comparable units in the non-probability sample (i.e., outside the common support), MMI may lead to biased estimates or

require extrapolation, which is inherently risky. Careful assessment and potentially trimming of units outside common support are necessary [13].

Selection of Auxiliary Variables: The choice and quality of common auxiliary variables X are paramount. These variables must be strongly related to both the outcome variables and the propensity of being in the probability sample [14]. Poorly chosen or insufficient auxiliary variables can lead to ineffective matching and residual bias.

Variance Estimation: Accurately estimating the variance of MMI estimators is complex, as it must account for both the sampling variability from the probability sample and the additional variability introduced by the imputation process [4, 12, 30]. Simplified variance estimation can lead to overly optimistic confidence intervals.

Computational Complexity: For very large datasets, especially when employing sophisticated statistical learning methods for matching and imputation, the computational demands of MMI can be substantial.

3 Future Directions

Future research in Matched Mass Imputation and survey data integration should focus on:

Robustness to Assumption Violations: Developing methods that are more robust to violations of the conditional independence and common support assumptions, perhaps through sensitivity analyses or partial identification techniques.

Automated Auxiliary Variable Selection: Exploring automated or semi-automated methods for selecting optimal auxiliary variables, especially in high-dimensional settings, potentially leveraging causal inference principles.

Advanced Statistical Learning Integration: Further integrating cutting-edge statistical learning and machine learning algorithms for more flexible and accurate matching and imputation models, while ensuring interpretability and theoretical guarantees.

Unified Frameworks for Multiple Imputation: Developing MMI within a multiple imputation framework to provide more robust variance estimates and account for imputation uncertainty more comprehensively [20].

Software Development: Creating user-friendly and computationally efficient software packages that implement MMI and its variance estimation methods, making it more accessible to practitioners.

Applications to New Data Sources: Exploring the application of MMI to emerging data sources, such as sensor data, social media data, or administrative records, in conjunction with traditional surveys.

CONCLUSION

Matched Mass Imputation represents a sophisticated and increasingly vital approach for integrating data from diverse survey sources. By strategically combining the principles of matching and mass imputation, it offers a powerful mechanism to mitigate the inherent biases of non-probability samples while leveraging their scale to enhance the precision of estimates derived from representative probability samples. The method's ability to accommodate advanced statistical learning techniques further strengthens its capacity to handle complex data structures and relationships.

While the successful application of MMI relies on careful consideration of its underlying assumptions, particularly conditional independence and common support, and requires robust variance estimation, its potential benefits for modern survey statistics are undeniable. As the landscape of data collection continues to evolve, MMI provides a crucial pathway for researchers and practitioners to produce more accurate, efficient, and timely statistical inferences, contributing significantly to evidence-based decision-making in an increasingly data-rich world.

REFERENCES

1. Beaumont JF, Rao J (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? The Survey Statistician, 83: 11–22.
2. Bethlehem J (2016). Solving the nonresponse problem with sample matching? Social Science Computer Review, 34(1): 59–77. https://doi.org/10.1177/0894439315573926
3. Centers for Disease Control and Prevention (CDC) (2015–2020). NHANES - National Health and Nutrition Examination Survey. https://www.cdc.gov/nchs/nhanes/index.htm (visited: 2023-10-11).
4. Chen S, Yang S, Kim JK (2022). Nonparametric mass imputation for data integration. Journal of Survey Statistics and Methodology, 10(1): 1–24. https://doi.org/10.1093/jssam/smaa036
5. Chen Y, Li P, Wu C (2020). Doubly robust inference with nonprobability survey samples. Journal of the American Statistical Association, 115(532): 2011–2021. https://doi.org/10.1080/01621459.2019.1677241
6. Dever J (2018). Combining probability and nonprobability samples to form efficient hybrid estimates: An evaluation of the common support assumption. In: Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference, 1–15.
7. Hájek J (1964). Asymptotic theory of rejective sampling with varying probabilities from a finite population. The Annals of Mathematical Statistics, 35(4): 1491–1523.
8. Horvitz DG, Thompson DJ (1952). A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association, 47(260): 663–685. https://doi.org/10.1080/01621459.1952.10483446
9. James G, Witten D, Hastie T, Tibshirani R, et al. (2013). An Introduction to Statistical Learning, volume 112. Springer.
10. Kalay AF (2021). Double Robust Mass-Imputation with Matching Estimators. arXiv preprint: https://arxiv.org/abs/2110.09275.
11. Kern C, Li Y, Wang L (2021). Boosted kernel weighting–using statistical learning to improve inference from nonprobability samples. Journal of Survey Statistics and Methodology, 9(5): 1088–1113.

https://doi.org/10.1093/jssam/smaa028
12. Kim JK, Park S, Chen Y, Wu C (2021). Combining non-probability and probability survey samples through mass imputation. Journal of the Royal Statistical Society. Series A. Statistics in Society, 184(3): 941–963. https://doi.org/10.1111/rssa.12696
13. Lee BK, Lessler J, Stuart EA (2011). Weight trimming and propensity score weighting. PLoS ONE, 6(3): e18174. https://doi.org/10.1371/journal.pone.0018174
14. Li Y, Fay M, Hunsberger S, Graubard BI (2023). Variable inclusion strategies for effective quota sampling and propensity modeling: An application to sars-cov-2 infection prevalence estimation. Journal of Survey Statistics and Methodology, 11(5): 1204–1228. https://doi.org/10.1093/jssam/smad026
15. Little RJ (1988). A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association, 83(404): 1198–1202.
16. Lohr SL (2021). Sampling: Design and Analysis. Chapman and Hall/CRC.
17. Maia M, Azevedo AR, Ara A (2021). Predictive comparison between random machines and random forests. Journal of Data Science, 19(4): 593–614. https://doi.org/10.6339/21-JDS1025
18. National Academies of Sciences, Engineering, and Medicine (2018). Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps. National Academies Press.
19. Rivers D (2007). Sampling for web surveys. American Statistical Association, Alexandria, VA, 1–26.
20. Rubin DB (1976). Inference and missing data. Biometrika, 63(3): 581–592. Publisher: Oxford University Press. https://doi.org/10.1093/biomet/63.3.581
21. Särndal CE, Swensson B, Wretman J (2003). Model Assisted Survey Sampling. Springer Science & Business Media.
22. Scott DW (2009). Sturges' rule. Wiley Interdisciplinary Reviews. Computational Statistics, 1(3): 303–306. https://doi.org/10.1002/wics.35
23. Stuart EA (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1): 1. https://doi.org/10.1214/09-STS313
24. Stuart EA, King G, Imai K, Ho D (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. Journal of Statistical Software, 42(8): 1–28. https://doi.org/10.18637/jss.v042.i08
25. Sturges HA (1926). The choice of a class interval. Journal of the American Statistical Association, 21(153): 65–66. https://doi.org/10.1080/01621459.1926.10502161
26. Wang L, Graubard BI, Katki HA, Li Y (2020). Improving external validity of epidemiologic cohort analyses: A kernel weighting approach. Journal of the Royal Statistical Society. Series A. Statistics in Society, 183(3): 1293–1311.
27. Wang YH (1993). On the number of successes in independent trials. Statistica Sinica, 3(2): 295–312.
28. Wiśniowski A, Sakshaug JW, Perez Ruiz DA, Blom AG (2020). Integrating probability and nonprobability samples for survey inference. Journal of Survey Statistics and Methodology, 8(1): 120–147. https://doi.org/10.1093/jssam/smz051
29. Wood SN (2017). Generalized Additive Models: An Introduction with R. CRC Press.
30. Yang S, Kim JK (2020). Statistical data integration in survey sampling: A review. Japanese Journal of Statistics and Data Science, 3: 625–650. https://doi.org/10.1007/s42081-020-00093-w
31. Yang S, Kim JK, Hwang Y (2021). Integration of data from probability surveys and big found data for finite population inference using mass imputation. Survey Methodology, 47(1): 29–58.
32. Yang S, Kim JK, Song R (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. Journal of the Royal Statistical Society, Series B, Statistical Methodology, 82(2): 445–465. https://doi.org/10.1111/rssb.12354

.