Benchmarking Missing Data Imputation Methods in Socioeconomic Surveys

Anonymous authors
Paper under double-blind review

Abstract

Missing data imputation is a core challenge in socioeconomic surveys, where data is often longitudinal, hierarchical, high-dimensional, not independent and identically distributed, and missing under complex mechanisms. Socioeconomic datasets like the Consumer Pyramids Household Survey (CPHS)—the largest continuous household survey in India since 2014, covering 174,000 households—highlight the importance of robust imputation, which can reduce survey costs, preserve statistical power, and enable timely policy analysis. This paper systematically evaluates these methods under three missingness mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), across five missingness ratios ranging from 10% to 50%. We evaluate imputation performance on both continuous and categorical variables, assess the impact on downstream tasks, and compare the computational efficiency of each method. Our results indicate that classical machine learning methods such as MissForest and HyperImpute remain strong baselines with favorable trade-offs between accuracy and efficiency, while deep learning methods perform better under complex missingness patterns and higher missingness ratios, but face scalability challenges. We ran experiments on CPHS and multiple synthetic survey datasets, and found consistent patterns across them. Our framework aims to provide a reliable benchmark for structured socioeconomic surveys, and addresses the critical gap in reproducible, domain-specific evaluation of imputation methods. The open-source code is provided in Appendix A.2.

1 Introduction

Missing data is a pervasive challenge in data science and machine learning, especially in real-world socioeconomic survey datasets (Silva-Ramírez et al., 2015; Wang et al., 2021). Data is often incomplete due to nonresponse or privacy concerns (Rubin, 2004). Imputation mitigates nonresponse bias and supports policy evaluation (Chen & Shao, 2000; Little & Rubin, 2019; Yang et al., 2024; Abdelnaby et al., 2024).

Despite the proliferation of imputation methods, there is a conspicuous lack of benchmarks to evaluate them on publicly available, large-scale, realistic datasets that capture the complexity of real-world socioeconomic survey data while allowing controlled introduction of missingness. Most empirical studies on missing data rely on relatively small datasets—such as the UCI machine learning repository (Zhang et al., 2025; Du et al., 2024; Miao et al., 2023; Bertsimas et al., 2018b) or limited clinical datasets (Zheng & Charoenphakdee, 2022)—or on synthetically generated data with simplistic assumptions (e.g., features drawn from a standard normal distribution) (Sun et al., 2023). Missingness is often simulated by randomly masking data under MCAR or MAR assumptions, which fail to reflect complex real-world patterns. In practice, missingness often follows the more challenging MNAR mechanism, where whether a value is missing depends directly on its unobserved value. Furthermore, numerous current benchmarks emphasize exclusively the accuracy of imputation, specifically evaluating the proximity of the imputed values to the actual values, while neglecting the consequences on downstream tasks (Zhang et al., 2025; Jarrett et al., 2022; Hastie et al., 2015; Biessmann et al., 2019). In practical applications, the goal of imputation is usually to enable reliable analysis or predictive modeling; thus, evaluating how different imputation methods affect the performance of subsequent tasks is crucial.

Table 1: Comparison of imputation benchmark datasets cited in ≥ 3 studies. #Con = number of continuous features, #Cat = number of categorical features. MCAR/MAR/MNAR = number of missingness ratios per mechanism. "Train/Test Evaluation" indicates evaluation with a train/test split.

Dataset	# Rows	# Con	# Cat	MCAR	MAR	MNAR	Train/Test Evaluation	Downstream Task	Hierarchical	Longitudinal
Housing	20k	9	-	1	1	1	✓	-	-	-
Letter	20k	16	-	1	1	1	✓	-	-	-
Credit	30k	14	9	1	1	1	✓	-	-	-
News	40k	58	-	1	1	1	✓	-	-	-
Concrete	1k	8	-	4	4	4	-	-	-	-
Wine	5k	11	-	4	4	4	-	-	-	-
Diabetes	20k	7	14	4	4	4	-	-	-	-
Spam	4k	56	-	4	4	4	-	-	-	-
CPHS	1.4M	16	8	5	5	5	✓	✓	✓	✓
SynthCPHS	1M	16	8	5	5	5	✓	✓	✓	✓
SubSDIC	500k	6	12	5	5	5	✓	✓	✓	-

Although benchmarks exist for imputation on socioeconomic survey data (Wang et al., 2021; Li et al., 2024; Kalton & Kasprzyk, 1982), they typically suffer from several limitations, such as excluding MNAR scenarios, relying on a small set of missingness ratios (Bertsimas et al., 2018b), and lacking a systematic evaluation framework. To address these gaps, our work bridges the divide between restricted real-world data and reproducible experimentation by introducing a comprehensive and open benchmark for missing data imputation in socioeconomic surveys. To the best of our knowledge, we are the first to provide a large-scale, publicly shareable benchmark that integrates real, synthetic, and open socioeconomic datasets under diverse missingness scenarios and systematic evaluation metrics. Our contributions include the following:

- **Evaluations on Real, Synthetic, and Public Datasets:** We benchmark imputation methods on three datasets: the real-world CPHS (Pais & Rawal, 2021), its high-fidelity synthetic counterpart SynthCPHS, and the publicly shareable SubSDIC derived from the World Bank's SDIC.
- **Comprehensive Missingness Scenarios:** We evaluate 14 imputation methods under three missingness mechanisms (MCAR, MAR, MNAR) and five missingness ratios, offering a broad and realistic spectrum of evaluation conditions.
- **Multi-metric Analysis & Downstream Task Evaluation:** In addition to the imputation accuracy on both continuous and categorical variables, we assess performance on downstream classification and regression tasks using multiple models to ensure robustness. We also systematically compare the computational efficiency of each method.

The remainder of this paper is organized as follows. Section 2 reviews related work on benchmark datasets, imputation methodologies, and synthetic data. Section 3 introduces the datasets used in our study. Section 4 defines the problem and missingness mechanisms. Section 5 describes our experimental setup and evaluation protocols. Section 6 presents results and analysis. Finally, Section 7 concludes the paper.

2 Related Work

2.1 Benchmark Datasets for Tabular Imputation

Research on imputation for tabular data often uses small, flat datasets like those from UCI machine learning repository (Kelly et al., 2025). As shown in Table 2.1, these datasets usually contain a few thousand to 100k samples with dozens of features, lacking clear hierarchical or temporal dependencies between variables. Details of these datasets are in Appendix A.4. Even more limiting, most studies simulate missing data, focusing on simplified MCAR or MAR scenarios with a single missingness level, which limits generalizability, since real-world data have more complex patterns.

A notable exception is the work of Jäger et al. (2021), who conducted a large-scale benchmark of imputation methods across 69 heterogeneous datasets from OpenML. Although their study offers valuable insight into

general-purpose imputation performance, the datasets used are not drawn from the socioeconomic survey domain and lack the hierarchical and longitudinal structures typically present in national household surveys.

Recently, synthetic data benchmarks (Sun et al., 2023) have gained traction to test imputation algorithms under controlled conditions (e.g., varying missing rates or mechanisms); but most synthetic setups do not capture the complexity of real-world structured data. In particular, few, if any, existing benchmarks replicate the large-scale, multi-level characteristics of national socioeconomic surveys, which span millions of entries with state or regional hierarchies and repeated observations over time.

2.2 Imputation Methodologies

Approaches to imputing missing values can be grouped by their modeling philosophy. Statistical and iterative methods use repeated estimation cycles, including mean or mode filling, and classic algorithms like MICE (Van Buuren & Groothuis-Oudshoorn, 2011) and MissForest (Stekhoven & Bühlmann, 2012) that iteratively train predictors for each feature, as well as matrix-completion methods like SoftImpute (Hastie et al., 2015). These methods assume linear or low-rank structures and often struggle with complex nonlinear feature interactions. Recent methods such as MIRACLE (Kyono et al., 2021) introduce a causally aware regularization that models the missingness mechanism jointly with the data, encouraging imputations consistent with the underlying causal structure. Distribution-matching methods align observed and imputed distributions: MOT (Missing-data Optimal Transport) (Muzellec et al., 2020) formulates imputation as finding the allocation of missing values that minimizes the optimal transport distance between batches of incomplete data, while its successor TDM (Transformed Distribution Matching) (Zhao et al., 2023) learns a nonlinear mapping before applying optimal transport to better capture the data's intrinsic geometry. These methods achieve state-of-the-art accuracy on many benchmark tasks and are particularly effective for in-sample imputation but generalize poorly to new records since they treat missing entries as learned model parameters.

Deep generative models (VAE (Mattei & Frellsen, 2019), GAN (Yoon et al., 2018), diffusion (Zheng & Charoenphakdee, 2022)) capture joint distributions of observed and missing data. While these models can capture complex nonlinear dependencies, they often face challenges in estimating distributions from incomplete data and in performing conditional inference, especially under high missingness. To overcome these issues, recent methods combine generation with iterative refinement. For example, **DiffPuter** (Zhang et al., 2025) integrates diffusion models into an EM framework, using iterative E- and M-steps to improve imputation quality.

Hybrid deep learning methods blends machine learning pipelines with automated model selection or specialized architectures. **HyperImpute** (Jarrett et al., 2022) employs an AutoML-style pipeline that selects the best model for each variable and updates imputations iteratively. Other architectures leverage advanced designs: **DSAN** (Lee & Kim, 2023) applies self-attention to learn feature and sample dependencies via masked reconstruction, while **ReMasker** (Du et al., 2024) extends masked autoencoding by re-masking observed entries during training, promoting robustness across different missingness patterns. Deep learning models have demonstrated robust performance on challenging imputation tasks, especially when missing rates are high or feature types are heterogeneous (Zhang et al., 2025).

Some recent studies have empirically evaluated the "impute-then-predict" pipeline. For example, Bertsimas et al. (2018a) and Poulos & Valle (2018) proposed frameworks that highlight the importance of integrating imputation with supervised learning tasks. More recent work by Paterakis et al. (2024) questions whether explicit imputation is always necessary in predictive pipelines, particularly within the context of AutoML. However, our work places equal emphasis on two complementary objectives: restoring incomplete datasets to reduce the need for costly follow-up surveys, and ensuring robust performance on downstream tasks.

2.3 Synthetic Data for Imputation

When real-world socioeconomic data is private or lacks ground truth, synthetic data provides a practical alternative for benchmarking imputation methods. Prior work has used synthetic simulations to evaluate methods under controlled conditions (Kyono et al., 2021; Sun et al., 2023), but most rely on simple i.i.d. data or low-dimensional toy settings (Muzellec et al., 2020; Bertsimas et al., 2024), lacking the structural and

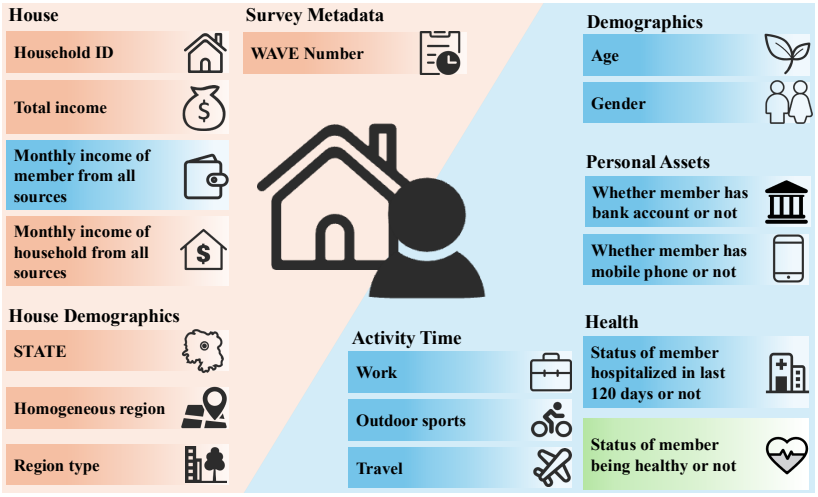

Figure 1: Overview of selected features in the CPHS and SynthCPHS datasets. Household-level (orange) and individual-level (blue) features form a hierarchical structure. The target variable (green) indicates individual health status for downstream classification. The wave number indexes each survey round, capturing longitudinal structure.

statistical complexity of real surveys. National household surveys like CPHS feature multi-level hierarchies, repeated measurements, and non-random missingness, but their proprietary nature limits open evaluation. To address this, we introduce a synthetic benchmark dataset, **SynthCPHS**, which we design to replicate the structure and distribution of CPHS. We validate the similarity using the Kolmogorov–Smirnov (KS) test and Jensen–Shannon (JS) divergence, and the synthetic construction enables GPU-supported evaluation beyond CPHS's secure CPU-only server environment. We also present **SubSDIC**, a public subset derived from World Bank data (World Bank, 2023), which supports systematic and reproducible benchmarking in the socioeconomic domain.

3 Datasets

3.1 CPHS & The SynthCPHS Dataset

3.1.1 CPHS

The Consumer Pyramids Household Survey (CPHS) by CMIE is the largest continuous household survey in India, running since 2014 and covering a panel of over 174,000 sample houses (about 111,000 rural and 63,400 urban) spread across most states in India surveyed thrice yearly (Pais & Rawal, 2021; Somanchi, 2021). It captures a broad array of household attributes including labor supply, income, consumption (expenditure on various needs), borrowing, and asset ownership (Pais & Rawal, 2021). This breadth makes it highly valuable for socioeconomic analysis (Chatterjee & Dev, 2023; Kathuria & Dev, 2024; Jagannarayan & Prasuna, 2024). In this paper, we select 25 socioeconomic features, as shown in Figure 1). Because some variables were missing in earlier waves, we restrict the analysis to complete cases from waves 18–30 (each wave is a survey round), preserving the longitudinal and hierarchical structure and yielding 1,341,651 records.

3.1.2 SynthCPHS

As CPHS is proprietary, available only by subscription, and restricted to a secure CPU-only server, benchmarking is severely constrained, especially for GPU-based imputation. To address this, we construct SynthCPHS, a synthetic dataset that mirrors the statistical properties and structure of CPHS but can be used on external GPU systems, enabling full-scale evaluation of imputation accuracy and efficiency. SynthCPHS includes 1,000,000 records and shares the same feature sets with the CPHS dataset as shown in Figure 1. The dataset was generated using the `synthpop` package in R (Nowok et al., 2016), a widely used tool to produce

artificial microdata that preserve the statistical properties of the original survey while protecting individual confidentiality (Nowok et al., 2017). To support its validity, we compared the marginal and joint distributions of key features in SynthCPHS and CPHS using the Kolmogorov–Smirnov (KS) test and Jensen–Shannon (JS) divergence, finding no significant distributional differences.

3.2 SubSDIC

To ensure reproducibility, we introduce **SubSDIC**, a public subset dataset derived from the World Bank's Synthetic Data for an Imaginary Country (SDIC)-a fully synthetic census dataset representing an imaginary middle-income country. SDIC was generated using REaLTabFormer (Solatorio & Dupriez, 2023), a deep generative model trained on global household survey data, including IPUMS International, DHS, and the World Bank Global Consumption Database. SDIC consists of two flat tables for household- and individual-level attributes. We join these tables via household ID, select 19 mixed-type variables spanning both levels, and randomly sample 500k records from the full 10 million to construct SubSDIC. We designate the individual's highest educational attainment (`cat_educ_attain`) as the target variable for downstream classification and years of schooling (`con_yrs_school`) as the target variable for downstream regression. SubSDIC preserves realistic socioeconomic frameworks, including hierarchical relationships among households, individuals, provinces, and districts, allowing for regulated evaluation of imputation accuracy, efficiency, and downstream performance across different missing data scenarios. Kolmogorov-Smirnov tests and Jensen-Shannon divergence analysis confirm that SubSDIC closely mimics the marginal distributions of the original SDIC dataset. Detailed feature descriptions are provided in Appendix A.5.

4 Problem Definition and Missing Mechanism

4.1 Problem Definition

Let $\mathbf{X}$ denote the matrix $n \times d$ that contains the complete data values in the variables d for all n units in the sample. Define the mask variable $\mathbf{M}$ as an $n \times d$ 0-1 matrix indicating whether a data point of $\mathbf{X}$ is observed (1) or missing (0). The elements of $\mathbf{X}$ and $\mathbf{M}$ are denoted by x_{ij} and m_{ij}, respectively, where $i = 1, ..., n$ and $j = 1, ..., d$. We further define the partially observed data matrix as $\widetilde{\mathbf{X}}$, and its elements $\tilde{x}_{ij}$, such that

$$\tilde{x}_{ij} = \begin{cases} x_{ij}, & \text{if} \quad m_{ij} = 1 \\ \emptyset, & \text{if} \quad m_{ij} = 0 \end{cases},$$

Here $\emptyset$ represents an unobserved value. In the missing data imputation problem, the task is imputing data matrix $\widehat{\mathbf{X}}$ from the observed data matrix $\widetilde{\mathbf{X}}$ and make it as similar as possible to the complete data matrix $\mathbf{X}$.

4.2 Missingness Mechanism

Little & Rubin (2019) find it helpful to differentiate between the missingness mechanisms, which refers to the relationship between the occurrence of missing data and the values of the variables in the data matrix. The missing mechanism indicates whether the occurrence of missingness is connected to the underlying values of the variables in the dataset. The importance of missingness mechanisms lies in the fact that the effectiveness of data imputation methods is highly influenced by the specific dependencies present in these mechanisms. Therefore, we introduce the three missingness mechanisms as defined by Rubin (1976) here. Let $\mathbf{X}$ and $\mathbf{M}$ be defined as in Section 4.1. Assume the rows $(\mathbf{x}_i, \mathbf{m}_i)$ are i.i.d. across i. The missingness mechanism is specified by the conditional distribution $p_{\mathbf{M}|\mathbf{X}}(\mathbf{m}_i \mid \mathbf{x}_i, \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes unknown parameters.

- **MCAR**: If the missingness is independent of the data values, either missing or observed, this means for all i and any distinct values $\mathbf{x}_i, \mathbf{x}_i^*$ in the sample space of $\mathbf{X}$, the conditional distributions are equal:

  $$f_{\mathbf{M}|\mathbf{X}}(\mathbf{m}_i \mid \mathbf{x}_i, \boldsymbol{\theta}) = f_{\mathbf{M}|\mathbf{X}}(\mathbf{m}_i \mid \mathbf{x}_i^*, \boldsymbol{\theta})$$

  where $\mathbf{x}_i^*$ serves as a placeholder for a distinct, hypothetical value that could take the place of $\mathbf{x}_i$ in the sample space of $\mathbf{X}$.

- **MAR**: Let $\mathbf{x}_{(1)i}$ represent the observed components of $\mathbf{x}_i$ and $\mathbf{x}_{(0)i}$ represent the missing components of $\mathbf{x}_i$. A less restrictive assumption than MCAR is that the missingness depends on $\mathbf{x}_i$ only through the observed components $\mathbf{x}_{(1)i}$. This implies that for any distinct values $\mathbf{x}_{(0)i}, \mathbf{x}_{(0)i}$ of the missing components within the sample space of $\mathbf{x}_{(0)i}$, the probability of missingness remains the same. In mathematical terms, the conditional distribution of the missingness mechanism can be expressed as:

$$f_{\mathbf{M}|\mathbf{X}}(\mathbf{m}_i \mid \mathbf{x}_{(0)i}, \mathbf{x}_{(1)i}, \boldsymbol{\theta}) = f_{\mathbf{M}|\mathbf{X}}(\mathbf{m}_i \mid \mathbf{x}^*_{(0)i}, \mathbf{x}_{(1)i}, \boldsymbol{\theta})$$

- **MNAR**: Unlike the MAR mechanism, where missingness is related to the observed values, if missingness is dependent on the unobserved (missing) values, the mechanism is classified as MNAR. The distribution of $\mathbf{m}_i$ depends on the missing components of $\mathbf{x}_i$, which means that equation 4.2 is not valid for some units i and some values $\mathbf{x}_{(0)i}, \mathbf{x}_{(0)i}$ of the missing components.

5 Experimental Setup and Evaluation

5.1 Dataset Distribution Comparison

We evaluate how closely the synthetic dataset (SynthCPHS) reproduces the marginal distributions of the real CPHS across continuous and categorical variables.

- **Method for continuous variables: two-sample KS test.** Let $\{x_i\}_{i=1}^n$ and $\{y_j\}_{j=1}^m$ be two i.i.d. samples with empirical CDFs F_n and G_m. The two-sample Kolmogorov–Smirnov statistic is

$$D_{n,m} \;=\; \sup_{x\in\mathbb{R}} \big|F_n(x) - G_m(x)\big|. \tag{1}$$

  Under the null hypothesis that both samples come from the same continuous distribution, $\sqrt{\frac{nm}{n+m}}\, D_{n,m}$ converges in distribution to the Kolmogorov distribution, which yields exact or asymptotic p-values (Massey Jr, 1951). In all our results we report the unscaled statistic $D_{n,m}$ in Eq.(1), denoted as "KS" in the figures, together with its two-sided p-value.

- **Method for categorical variables: JS divergence.** For two discrete distributions P and Q over the same support $\mathcal{X}$, the JS divergence is the symmetrised, smoothed version of the Kullback-Leibler (KL) divergence:

$$\mathrm{JS}(P\|Q) = \tfrac{1}{2}\,\mathrm{KL}\big(P\big\|M\big) + \tfrac{1}{2}\,\mathrm{KL}\big(Q\big\|M\big), M = \tfrac{1}{2}\,(P+Q), \tag{2}$$

  where $\mathrm{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}$. JS is always finite, bounded between 0 (identical distributions) and $\log 2$ (base-e), and admits a metric square root (Lin, 2002; Endres & Schindelin, 2003). We report values in $[0, 1]$ by dividing by $\log 2$.

5.2 Benchmark Experimental Settings

5.2.1 Missingness Mechanism and Ratio

The effectiveness of missing data imputation methods is strongly influenced by factors such as the missingness mechanism and ratio. To rigorously evaluate imputation methods, we introduce missing values under 3 missingness mechanisms: MCAR, MAR, and MNAR. The missingness implementation details are provided in Appendix A.6. We create versions of the dataset with missingness ratios of 10%, 20%, 30%, 40%, and 50%. The missingness ratio is calculated as the fraction of all entries that are masked, and each feature gets roughly the same fraction of its values missing, though in MAR/MNAR this can vary slightly due to the conditioning. Each missing scenario is generated with five samples and then fixed, so all methods are evaluated on the exact same missing data patterns for fairness.

5.2.2 Imputation Methods

We provide a comprehensive benchmark of 14 widely used imputation methods across four categories: (1) a statistical baseline — Mean/Mode imputation; (2) distribution-matching methods such as **MOT** (Muzellec et al., 2020), which uses optimal transport to align observed and imputed distributions; (3) iterative machine learning methods including **MICE** (Van Buuren & Groothuis-Oudshoorn, 2011), **MIRACLE** (Kyono et al., 2021), **SoftImpute** (Hastie et al., 2015), and **MissForest** (Stekhoven & Bühlmann, 2012); and (4) deep generative models — **MIWAE** (VAE) (Mattei & Frellsen, 2019), **GAIN** (GAN) (Yoon et al., 2018), **DSAN** (self-attention) (Lee & Kim, 2023), and **TabCSDI** (diffusion) (Zheng & Charoenphakdee, 2022). We also include three recent state-of-the-art approaches: **ReMasker** (Du et al., 2024), **HyperImputer** (Jarrett et al., 2022), and **DiffPuter** (Zhang et al., 2025), along with a DSAN variant (**DSN**) without attention to assess its contribution. Implementation details and hyperparameters are in Appendix A.7.

5.2.3 Imputation Performance

For each dataset with missingness, 80% of samples are used for training and 20% for testing. All methods are trained on the training set and then used to impute both in-sample and out-of-sample data. Imputation performance is measured using **RMSE** for continuous and **F1 score** for categorical variables to provide a balanced evaluation given class imbalance. The RMSE is computed on standardized inputs (zero mean, unit variance) based on training-set statistics, and **accuracy** for categorical variables is also reported as a supplementary metric.

5.2.4 Downstream Task Performance

To robustly assess the downstream impact of imputation, we test two task types: **classification** and **regression**. For classification, all three datasets are evaluated with Random Forest (RF) and XGBoost models; for regression, only SubSDIC is used, with the same model pair. Using multiple models reduces model-specific bias. Models are trained on complete training data and evaluated on imputed test sets. For classification, we report the **ROC-AUC degradation**—the drop in ROC-AUC from the fully observed test set—as the main metric, while **accuracy** is provided in the supplement. For regression, we report the **RMSE increase**, the percentage rise in RMSE relative to the fully observed test set. Smaller ROC-AUC degradation or RMSE increase indicates better imputation quality and stronger preservation of predictive signal.

5.2.5 Runtime and Efficiency

To provide practical insight for real-world deployment under resource constraints, we report the total wall-clock time for each method, including both training and imputation. For iterative algorithms such as MICE and MissForest, time covers all iterations; for deep learning models, it includes all training epochs. Appendix A.3 details the experimental setup. Each experiment is repeated five times, and we report mean values and standard deviations as final metrics.

5.3 Ranking Consistency Across Datasets

We assess cross-dataset performance consistency using Kendall's coefficient of concordance W as described by Abdi (2007), computed over the 13 methods common to all datasets. Although the full benchmark includes 14 methods, DiffPuter could not be executed on the proprietary CPHS dataset hosted on a CPU-only server and is therefore excluded from the consistency analysis. For k datasets and N methods, let R_{ij} be the rank of method i on dataset j, $R_i = \sum_{j=1}^{k} R_{ij}$ the aggregate rank, and $\overline{R} = \frac{k(N+1)}{2}$. Then

$$W = \frac{13}{k^2(N^3 - N)} \sum_{i=1}^{N} \big(R_i - \overline{R}\big)^2. \tag{3}$$

$W \in [0, 1]$ (0 = no agreement; 1 = perfect concordance). For $k > 2$, $k(N-1)W$ is asymptotically χ^2_{N-1} under independence (Abdi, 2007). In our study, $k = 3$ datasets and $N = 13$ methods. We use common thresholds: strong ($W > 0.70$), moderate ($0.50 \leq W \leq 0.70$), and weak ($W < 0.50$) agreement (De Maere et al., 2022).

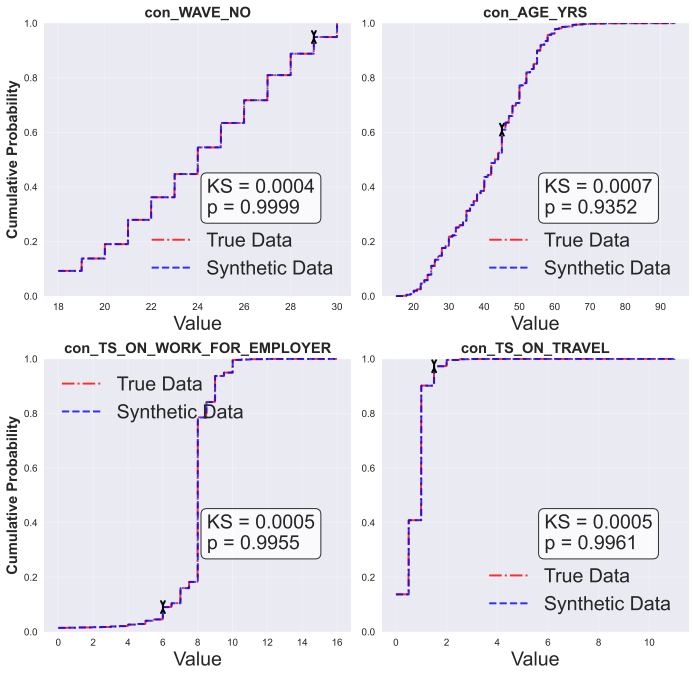

Figure 2: Empirical CDF comparison between CPHS (red) and SynthCPHS (blue) for four representative continuous variables. The double-headed arrow indicates the KS gap $D_{n,m}$.

6 Results and Analysis

This section begins with distribution comparison results. We then present a multi-metric benchmark of 14 imputation methods, covering imputation performance, downstream task performance, and computational efficiency. In all figures, the numbers after method names in the legend indicate the average rank of that metric across missingness ratios (shown for the top nine methods only). Full hyperparameter settings and complete results for all datasets and metrics are provided in Appendix A.7 and in the supplementary material. Finally, we quantify the performance consistency between datasets using the Kendall coefficient of concordance computed on the 13 methods available on all datasets.

6.1 Distribution Comparison Results

- **Continuous variables (KS).** For each variable we compute the two-sample KS statistic $D_{n,m}$ in Eq.(1), comparing CPHS and SynthCPHS. The *KS* values displayed in Fig. 2 are exactly this maximum CDF gap $D_{n,m}$. In most cases $D_{n,m} < 10^{-3}$ and the corresponding two-sided p-values are close to 1, providing no evidence against the null of identical distributions.

- **Categorical variables (JS).** For each categorical variable we compute the normalized JS divergence, defined as $\mathrm{JS}_{[0,1]} = \frac{\mathrm{JS}(\hat{P}\|\hat{Q})}{\log 2}$, where $\mathrm{JS}(\hat{P}\|\hat{Q})$ is defined in Eq.(2) and we simply rescale it to $[0, 1]$. Across all examined variables, $\mathrm{JS}_{[0,1]} = 0$ (to numerical precision), indicating identical empirical distributions between CPHS and SynthCPHS; therefore, we omit plots.

Detailed per-variable values for both continuous and categorical features are provided in the supplementary material.

6.2 Imputation Performance

Figures 3 and 4 compare in-sample and out-of-sample performance across methods for continuous (RMSE) and categorical (F1 score) variables, respectively. We observe four key findings:

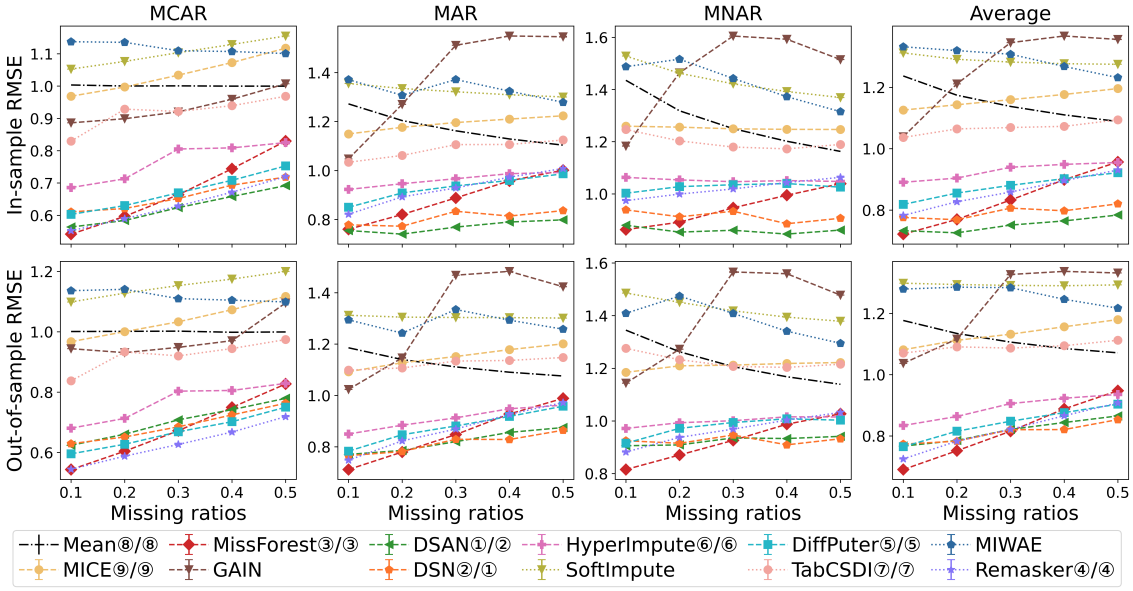

Figure 3: Continuous variable imputation performance. Top row: in-sample; bottom row: out-of-sample. Lower RMSE is better. MIRACLE and MOT are omitted due to excessively high RMSE. Legend entries show in-sample/out-of-sample ranks (first/second number).

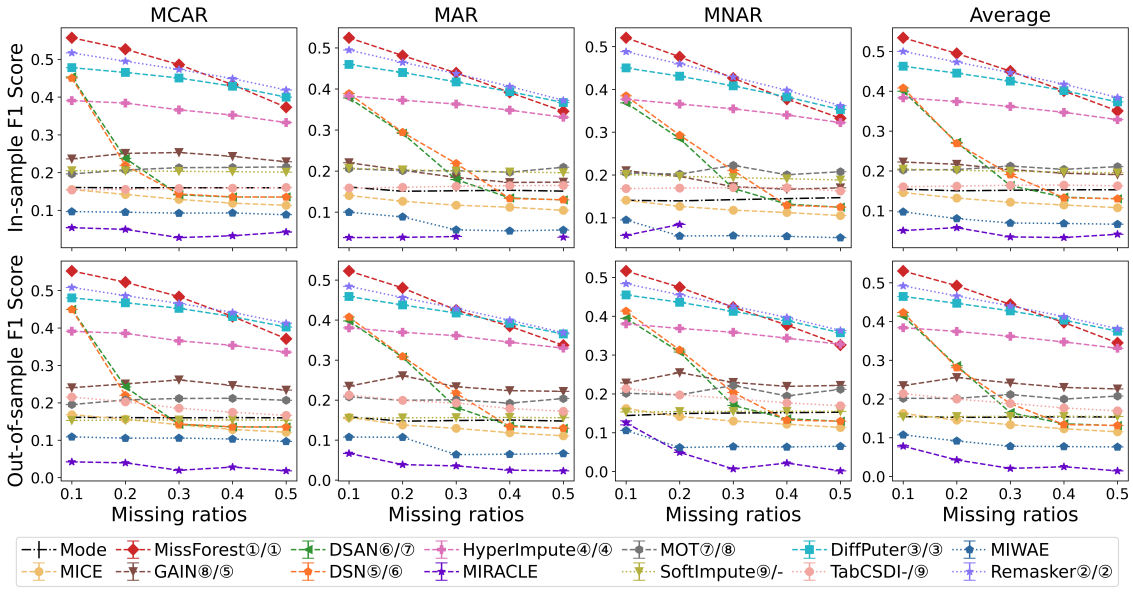

Figure 4: Categorical variable imputation performance. Top row: in-sample; bottom row: out-of-sample. Higher F1 is better. MIRACLE in-sample points are missing at some missingness ratios due to failure. Legend entries show in-sample/out-of-sample ranks (first/second number).

- **Missingness Mechanism Impact.** MCAR yields lower error (lower RMSE and higher F1) than MAR or MNAR, reflecting the challenge of imputing structured missingness. For MAR and MNAR, **RMSE** does not always increase with higher missingness ratios as one might expect. In some methods, such as Mean, SoftImpute, and MIWAE—RMSE slightly decrease as missingness increases, because these models revert to imputing global averages, which limits variability and reduces error compared to poorly estimated values under lower missingness. **F1 scores** for categorical data typically exhibit two trends: either remaining nearly constant across missingness levels, indicating a failure to learn meaningful patterns, or decreasing linearly with higher missingness ratios. An exception is observed in DSAN and its variant DSN, where F1 scores drop sharply under high missingness, suggesting these models are ineffective and unstable in such settings.

- **Method Comparison.** For **continuous variables** (Figure 3), MissForest typically performs best at low missingness ratios (e.g., 10%) but its performance steadily declines as the missingness ratio increases. In contrast, deep learning methods such as ReMasker, DSAN, and DSN outperform MissForest under MCAR, MAR, and MNAR when the missingness ratio is high. For **categorical variables** (Figure 4), MissForest continues to outperform other models at low missingness ratios regardless of the missingness mechanism. However, as the missingness increases, ReMasker and DiffPuter begin to achieve the highest F1 scores. DSAN and DSN, while strong performers for continuous variables, show noticeably worse performance on categorical data, with F1 scores degrading significantly as missingness increases. In summary, MissForest is highly sensitive to the missingness ratio rather than to the type of missingness or variable; it is among the top performers under low missingness (up to 20%), but deep learning methods often become more effective as missingness becomes more severe.

- **Self-Attention Analysis.** As shown in Figures 3 and 4, DSN outperforms DSAN in average ranking across most cases, suggesting limited benefit from the self-attention mechanism. In particular, as shown in Figure 3, while DSAN achieves the best in-sample performance for continuous imputation, DSN outperforms it on average in out-of-sample evaluations, indicating that the attention layer in DSAN introduces a higher risk of overfitting compared to DSN, a phenomenon also observed in a previous study by Dehimi & Tolba (2024).

- **Overfitting Assessment.** Out-of-sample RMSE and F1 scores closely match in-sample results, indicating minimal overfitting for most methods.

6.3 Downstream Task Performance

As shown in Figure 5, we report the downstream impact of imputation on both classification and regression tasks, using Random Forest models as examples.

- **Missingness Mechanism.** The results show that missingness under MCAR and MAR leads to slightly lower performance degradation than MNAR. ROC-AUC scores decrease and RMSE increases with higher missingness ratios, highlighting the adverse effect of missing data on downstream performance.

- **Robustness Analysis.** A clear alignment is observed between raw imputation quality (RMSE/F1) and downstream results: methods with lower degradation also achieve top imputation scores. For instance, DSAN and DSN perform best on continuous variables, while ReMasker, MissForest, and DiffPuter excel in categorical imputation, all showing minimal downstream impact. Those top-performing imputation methods maintain their rankings in both regression and classification settings. In two other datasets, while some methods' rankings vary, ReMasker and HyperImpute consistently perform well across all datasets. Figure 5 shows results using Random Forest models; similar patterns are observed with XGBoost (see the "Experiment_result" folder in the supplementary material). This consistency suggests that the influence of imputation methods on downstream performance is stable across task types and model choices.

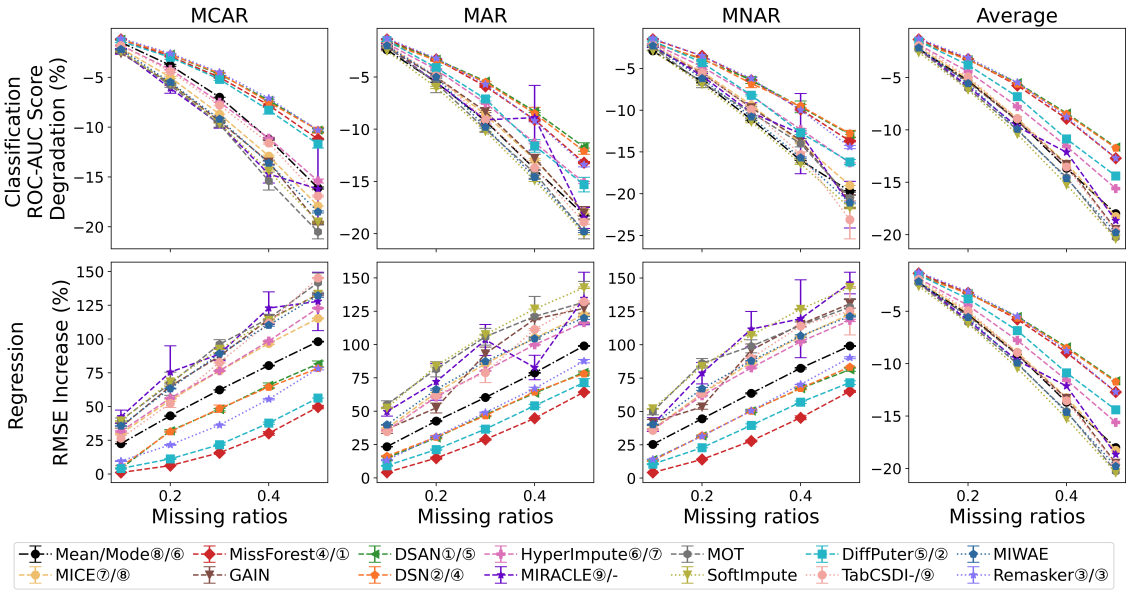

Figure 5: Top: ROC-AUC degradation for downstream classification using Random Forest. Bottom: RMSE increase for downstream regression using Random Forest. Lower ROC-AUC degradation and smaller RMSE increase indicate better preservation of predictive performance.

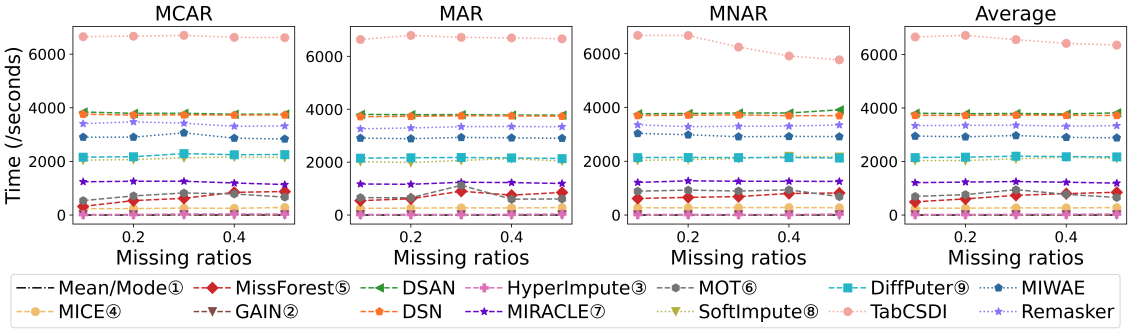

Figure 6: Comparison of imputation runtime.

6.4 Computational Efficiency

As shown in Figure 6, the runtime efficiency analysis indicates that the imputation time is relatively stable across varying missingness ratios and missing mechanisms for most methods. Statistical methods (e.g., Mean/Mode, MICE) and traditional machine learning methods (e.g., HyperImpute, MissForest) demonstrate significantly faster performance, approximately an order of magnitude faster than deep learning-based methods. Among these methods, MissForest stands out for its strong performance on continuous variables at low missingness ratios, solid categorical imputation accuracy, and exceptional time efficiency, making it well suited for large-scale practical applications. For more complex missingness scenarios, ReMasker proves to be a powerful imputation method for both continuous and categorical data, offering competitive efficiency compared to other deep learning approaches and resulting in minimal degradation in downstream task performance.

6.5 Consistency Analysis

We evaluate the consistency of the rankings across the three datasets using Kendall's coefficient of concordance (W) described in Eq.(3). As shown in Table 2, agreement is strong in scenarios A and D, moderate in B, and lower in C. Considering the generally consistent performance across scenarios, we center our discussion

Table 2: Ranking agreement across 3 datasets measured by Kendall's W (higher is better).

Scenario	W	Consensus
A: RMSE (numeric, out-of-sample)	0.857	Strong
B: F1 (categorical, out-of-sample)	0.629	Moderate
C: ROC-AUC degradation (RF)	0.411	Weak
D: Time efficiency	0.904	Strong

above on **SubSDIC**. This dataset is publicly accessible, facilitates both classification and regression tasks, represents the other datasets well, and prevents repetitive reporting across all three.

6.6 Analysis

Overall, by aggregating the average rankings of 14 imputation methods (13 on the CPHS dataset) on three missingness mechanisms and five missingness ratios, across three datasets and four representative tasks (1) Imputation RMSE rank, (2) Imputation F1 rank, (3) downstream regression rank, and (4) downstream classification rank using Random Forest, we find that HyperImpute and MissForest consistently achieve the best overall performance with a clear margin over all other methods. This suggests that although deep learning–based models such as DSAN, DSN, and ReMasker perform competitively, traditional methods like HyperImpute and MissForest deliver more consistent and superior overall results. Our findings therefore reinforce the strong practical competitiveness of traditional machine learning–based methods for missing data imputation, consistent with prior studies (Lalande & Doya, 2022; Zhang et al., 2025; Suh & Song, 2023; Jolicoeur-Martineau et al., 2024; Jäger et al., 2021). Furthermore, our benchmark also suggests that incorporating attention layers may increase the risk of overfitting, in line with the observations of Dehimi & Tolba (2024).

7 Conclusion

Conclusion: This work presents a comprehensive benchmark study across three large-scale socioeconomic survey datasets—both real and synthetic—that reflect key characteristics of the domain: longitudinal, hierarchical, large-scale, and non-i.i.d. Using these datasets, we systematically evaluate 14 diverse imputation methods under controlled missingness mechanisms, varying missingness ratios, and across continuous and categorical variables. Beyond imputation accuracy and downstream task performance, we also assess computational efficiency, providing a well-rounded evaluation of each method's practicality.

Our results confirm the strong performance of classical approaches observed in prior studies, while emphasizing the value of multimetric evaluation, including downstream task impact and efficiency, for understanding real-world applicability. The proposed benchmark offers a realistic, robust testbed for missing data research in structured socioeconomic contexts. By releasing the SubSDIC dataset and evaluation framework, we support reproducible research and foster progress in addressing complex missingness patterns in the survey domain.

Limitations & Future Work: While CPHS and SynthCPHS provide robust validation of our conclusions, third-party licensing restrictions prevent public dataset release. We note that certain baselines were evaluated using default hyperparameters, though we acknowledge this may conservatively estimate their potential. Graph-based methods (e.g., GRAPE, IGRM) were intentionally excluded from comparison, as they generate non-standard output representations incompatible with our evaluation framework. Finally, we recognize the substantial computational demands of comprehensive benchmarking, particularly for modern deep architectures. Future work will (1) release this framework as an open-source Python package for standardized evaluation, (2) integrate graph-based methods, and (3) continuously incorporate emerging imputation techniques.

References

Taher Abdelnaby, Tingyu Feng, Zhang Tiantian, Xiaoming Jiang, Wang Yuming, Zhaojie Li, and Changhu Xue. Impact of frozen storage on physicochemical parameters and quality changes in cooked crayfish. *Heliyon*, 10(11), 2024.

Hervé Abdi. The kendall rank correlation coefficient. *Encyclopedia of measurement and statistics*, 2:508–510, 2007.

Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. From predictive methods to missing data imputation: An optimization approach. *Journal of Machine Learning Research*, 18(196):1–39, 2018a. URL `http://jmlr.org/papers/v18/17-073.html`.

Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. From predictive methods to missing data imputation: an optimization approach. *Journal of Machine Learning Research*, 18(196):1–39, 2018b.

Dimitris Bertsimas, Arthur Delarue, and Jean Pauphilet. Simple imputation rules for prediction with missing data: Theoretical guarantees vs. empirical performance. *Transactions on Machine Learning Research*, 2024.

Felix Biessmann, Tammo Rukat, Phillipp Schmidt, Prathik Naidu, Sebastian Schelter, Andrey Taptunov, Dustin Lange, and David Salinas. Datawig: Missing value imputation for tables. *Journal of Machine Learning Research*, 20(175):1–6, 2019.

Partha Chatterjee and Aakash Dev. Labour market dynamics and worker flows in india: Impact of covid-19. *The Indian Journal of Labour Economics*, 66(1):299–327, 2023.

Jiahua Chen and Jun Shao. Nearest neighbor imputation for survey data. *Journal of official statistics*, 16(2): 113, 2000.

Koen De Maere, Steven De Haes, Michael von Kutzchenbach, and Tim Huygh. Identifying the enablers and inhibitors of organizational learning in the context of it governance: an exploratory delphi study. *Information Systems Management*, 39(3):241–268, 2022.

Nour El Houda Dehimi and Zakaria Tolba. Attention mechanisms in deep learning: Towards explainable artificial intelligence. In *2024 6th International Conference on Pattern Analysis and Intelligent Systems (PAIS)*, pp. 1–7. IEEE, 2024.

Tianyu Du, Luca Melis, and Ting Wang. Remasker: Imputing tabular data with masked autoencoding. In *The Twelfth International Conference on Learning Representations*, 2024. URL `https://openreview.net/forum?id=KI9NqjLVDT`.

Mark Elliot. Final report on the disclosure risk associated with the synthetic data produced by the sylls team. *Report 2015*, 2, 2015.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

Trevor Hastie, Rahul Mazumder, Jason D Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *The Journal of Machine Learning Research*, 16(1):3367–3402, 2015.

Nandini Jagannarayan and Asha Prasuna. An empirical analyses of determinants of health expenditure in rural amravati in march 2019, 2020 and 2021. *Available at SSRN 4975312*, 2024.

Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A benchmark for data imputation methods. *Frontiers in big Data*, 4:693674, 2021.

Daniel Jarrett, Bogdan C Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. In *International Conference on Machine Learning*, pp. 9916–9937. PMLR, 2022.

Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *International Conference on Artificial Intelligence and Statistics*, pp. 1288–1296. PMLR, 2024.

Graham Kalton and Daniel Kasprzyk. Imputing for missing survey responses. In *Proceedings of the section on survey research methods, American Statistical Association*, volume 22, pp. 31. American Statistical Association Cincinnati, 1982.

Rajat Kathuria and Aakash Dev. Technological advancement and employment changes: Recent trends in the indian economy. *The Indian Journal of Labour Economics*, 67(3):637–660, 2024.

Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. The UCI machine learning repository. `https://archive.ics.uci.edu`, 2025. Accessed 2025-10-22.

Trent Kyono, Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Miracle: Causally-aware imputation via learning missing data mechanisms. *Advances in Neural Information Processing Systems*, 34:23806–23817, 2021.

Florian Lalande and Kenji Doya. Numerical data imputation: Choose knn over deep learning. In *International Conference on Similarity Search and Applications*, pp. 3–10. Springer, 2022.

Do-Hoon Lee and Han-joon Kim. A self-attention-based imputation technique for enhancing tabular data quality. *Data*, 8(6):102, 2023.

JiaHang Li, ShuXia Guo, RuLin Ma, Jia He, XiangHui Zhang, DongSheng Rui, YuSong Ding, Yu Li, LeYao Jian, Jing Cheng, et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. *BMC Medical Research Methodology*, 24(1):41, 2024.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 2002.

Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

Frank J Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951.

Pierre-Alexandre Mattei and Jes Frellsen. Miwae: Deep generative modelling and imputation of incomplete data sets. In *International conference on machine learning*, pp. 4413–4423. PMLR, 2019.

Imke Mayer, Aude Sportisse, Julie Josse, Nicholas Tierney, and Nathalie Vialaneix. R-miss-tastic: a unified platform for missing values methods and workflows. *arXiv preprint arXiv:1908.04822*, 2019.

Xiaoye Miao, Yangyang Wu, Lu Chen, Yunjun Gao, and Jianwei Yin. An experimental survey of missing data imputation algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 35(7):6630–6650, 2023. doi: 10.1109/TKDE.2022.3186498.

Boris Muzellec, Julie Josse, Claire Boyer, and Marco Cuturi. Missing data imputation using optimal transport. In *International Conference on Machine Learning*, pp. 7130–7140. PMLR, 2020.

Beata Nowok, Gillian M Raab, and Chris Dibben. synthpop: Bespoke creation of synthetic data in r. *Journal of statistical software*, 74:1–26, 2016.

Beata Nowok, Gillian M Raab, and Chris Dibben. Providing bespoke synthetic data for the uk longitudinal studies and other sensitive data with the synthpop package for r. *Statistical Journal of the IAOS*, 33(3): 785–796, 2017.

Jesim Pais and Vikas Rawal. Cmie's consumer pyramids household surveys: An assessment. In *The Indian Forum*, pp. 16, 2021.

A Paszke. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703*, 2019.

George Paterakis, Stefanos Fafalios, Paulos Charonyktakis, Vassilis Christophides, and Ioannis Tsamardinos. Do we really need imputation in automl predictive modeling? *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–64, 2024.

Jason Poulos and Rafael Valle. Missing data imputation for supervised learning. *Applied Artificial Intelligence*, 32(2):186–196, March 2018. ISSN 1087-6545. doi: 10.1080/08839514.2018.1448143. URL http://dx.doi.org/10.1080/08839514.2018.1448143.

Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

Donald B Rubin. *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons, 2004.

Esther-Lydia Silva-Ramírez, Rafael Pino-Mejías, and Manuel López-Coello. Single imputation with multilayer perceptron and multiple imputation combining multilayer perceptron and k-nearest neighbours for monotone patterns. *Applied Soft Computing*, 29:65–74, 2015.

Aivin V. Solatorio and Olivier Dupriez. Realtabformer: Generating realistic relational and tabular data using transformers, 2023. URL https://arxiv.org/abs/2302.02041.

Anmol Somanchi. Missing the poor, big time: A critical assessment of the consumer pyramids household survey. *Retrieved December*, 12:2021, 2021.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Heajung Suh and Jongwoo Song. A comparison of imputation methods using machine learning models. *CSAM (Communications for Statistical Applications and Methods)*, 30(3):331–341, 2023.

Yige Sun, Jing Li, Yifan Xu, Tingting Zhang, and Xiaofeng Wang. Deep learning versus conventional methods for missing data imputation: A review and comparative study. *Expert Systems with Applications*, 227: 120201, 2023.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Zhenhua Wang, Olanrewaju Akande, Jason Poulos, and Fan Li. Are deep learning models superior for missing data imputation in large surveys? evidence from an empirical comparison. *arXiv preprint arXiv:2103.09316*, 2021.

World Bank. Synthetic data for an imaginary country, full population, 2023, 2023. URL https://microdata.worldbank.org/index.php/catalog/study/WLD_2023_SYNTH-CEN-EN_v01_M.

Xinyu Yang, Yu Sun, Xinyang Chen, et al. Frequency-aware generative models for multivariate time series imputation. *Advances in Neural Information Processing Systems*, 37:52595–52623, 2024.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pp. 5689–5698. PMLR, 2018.

Hengrui Zhang, Liancheng Fang, Qitian Wu, and Philip S Yu. Diffputer: Empowering diffusion models for missing data imputation. In *The Thirteenth International Conference on Learning Representations*, 2025.

He Zhao, Ke Sun, Amir Dezfouli, and Edwin V Bonilla. Transformed distribution matching for missing value imputation. In *International Conference on Machine Learning*, pp. 42159–42186. PMLR, 2023.

Shuhan Zheng and Nontawat Charoenphakdee. Diffusion models for missing value imputation in tabular data. In *NeurIPS 2022 First Table Representation Workshop*, 2022. URL https://openreview.net/forum?id=4q9kFrXC2Ae.

Table 3: Related imputation benchmark datasets.

Dataset	Full Name	Link
Housing	California Housing (Zhang et al., 2025; Du et al., 2024; Jarrett et al., 2022)	https://www.kaggle.com/datasets/camnugent/california-housing-prices
Letter	Letter Recognition (Zhang et al., 2025; Du et al., 2024; Jarrett et al., 2022; Yoon et al., 2018)	https://archive.ics.uci.edu/dataset/59/letter+recognition
Credit	Default of Credit Card Clients (Zhang et al., 2025; Du et al., 2024; Yoon et al., 2018)	https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients
News	Online News Popularity (Zhang et al., 2025; Yoon et al., 2018)	https://archive.ics.uci.edu/dataset/332/online+news+popularity
Concrete	Concrete Compressive Strength (Du et al., 2024; Jarrett et al., 2022; Zheng & Charoenphakdee, 2022)	https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength
Wine	Wine Quality (Du et al., 2024; Jarrett et al., 2022; Zheng & Charoenphakdee, 2022)	https://archive.ics.uci.edu/dataset/186/wine+quality
Diabetes	Diabetes (Du et al., 2024; Jarrett et al., 2022; Zheng & Charoenphakdee, 2022)	https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset/data
Spam	Spam Base (Du et al., 2024; Jarrett et al., 2022; Yoon et al., 2018)	https://archive.ics.uci.edu/dataset/94/spambase

A Appendix

A.1 Ethical Considerations and Limitations

SynthCPHS was generated with the goal of maintaining data utility while ensuring strong privacy protection. We leveraged the `synthpop` framework for fully synthetic data generation, an approach known for its statistical disclosure control properties (Nowok et al., 2017). By design, the synthesized records contain no actual individuals, and thus the risk of re-identification or sensitive attribute disclosure is extremely low. Prior evaluations support this: Nowok et al. (2017) note that fully synthetic data pose minimal disclosure risk, and an independent risk analysis by Elliot (2015) similarly found the disclosure risk in a synthpop-generated dataset to be "very small". In practice, additional safeguards (such as excluding any accidentally replicated unique cases) can be applied to further reduce even the perceived risk of identification.

Nevertheless, this synthetic dataset is not intended as a substitute for the original CMIE CPHS data. For real-world policy and socioeconomic research, direct access to the authentic CPHS data remains indispensable, as only the original data can provide fully reliable and legally accountable insights for decision-making.

A.2 Code

Code is available here: https://anonymous.4open.science/r/1H249F74N5F-3J4G8-JJ80/

A.3 Configurations

We conduct all 14 methods' experiments (SynthCPHS and SubSDIC) with:

- **Operating System:** Rocky Linux 8 (a rebuild of Red Hat Enterprise Linux 8)
- **CPU:** 2x AMD EPYC 7763 64-Core Processor 1.8GHz (128 cores in total)
- **RAM:** 1000 GiB
- **GPU:** 4x NVIDIA A100-SXM-80GB GPUs (each with 6912 FP32 CUDA cores)
- **Interconnect:** Dual-rail Mellanox HDR200 InfiniBand
- **Cluster:** 90 Dell PowerEdge XE8545 servers
- **Software:** CUDA 11.4, Python 3.9.20, PyTorch Paszke (2019) 2.6.0

A.4 Datasets

We include only those datasets that have been used in at least three imputation studies. The full names and links for these datasets are provided in Table 3.

A.5 SubSDIC: Feature Descriptions and Construction Details

Table 4 lists the 19 variables, including the target variables, in the dataset along with the corresponding SDIC "fake" survey questions created to collect the data. Features prefixed with "cat_" indicate categorical variables, while those prefixed with "con_" indicate continuous variables.

Table 4: Detailed Feature Description of SubSDIC

Feature Name	Question Construct (H=households, I=individuals)
cat_hid	Household identifier / Machine Generated (H)
cat_geo1	Geographic area - Admin 1 (H)
cat_geo2	Geographic area - Admin 2 (H)
cat_urbrur	Urban or rural indicator of household location (H)
con_hhsize	Household size, i.e., number of individuals in the household (H)
cat_statocc	Does the household own, rent, or occupies this dwelling for free? (H)
con_exp_09	How much does the household spend per year on? (H)
con_exp_10	How much does the household spend per year on? (H)
con_tot_exp	Total monthly household expenditure across all categories (H)
cat_relation	What is the relationship of [name] to the head of household? (I)
cat_sex	Is [name] male or female? (I)
con_age	How old is [name]? (I)
cat_marstat	What is [name's] marital status? (I)
cat_religion	What is the religion of [name]? (I)
cat_school_attend	Is [name] attending school or preschool? (I)
con_yrs_school	How many years has [name] attended school? (I)
cat_act_status	What is [name's] status of activity? (I)
cat_occupation	What is/was [name's] main occupation? (I)
cat_educ_attain	What is the highest level of school that [name] has completed? (I)

A.6 Generate Missingness

In this paper, missing values are synthetically introduced using procedures adapted from the R-miss-tastic platform (Mayer et al., 2019), a widely used repository for standardized missing data workflows and reproducible experiments. Specifically, we rely on their R implementation to generate missingness under three mechanisms: MCAR, MAR, and MNAR. The missingness is generated feature-wise using logistic models, without relying on fixed missingness patterns. **This section briefly describe how missing values are generated under MCAR, MAR, and MNAR mechanisms in their framework, using the notation in the main text problem definition section.**

A.6.1 Missing Completely at Random (MCAR)

In the MCAR setting, missingness is independent of both observed and unobserved data. For each selected variable j, missing entries are assigned uniformly at random:

$$m_{ij} \sim \text{Bernoulli}(1-p),$$

where $p \in (0,1)$ denotes the target missingness ratio. This mechanism ensures no structural dependence in the missingness pattern.

A.6.2 Missing at Random (MAR)

Under MAR, the missingness in variable $\mathbf{X}_j$ depends only on other observed variables. We define:

$$\mathbb{P}(m_{ij}=1 \mid \mathbf{x}_{i,-j}^{\text{obs}}) = \frac{1}{1+\exp\left(-\left(\mathbf{x}_{i,-j}^{\text{obs}}\right)^\top \boldsymbol{\beta}_j\right)},$$

where $\mathbf{x}_{i,-j}^{\text{obs}} = \{x_{ik} \mid m_{ik}=1,\ k \neq j\}$ and $\boldsymbol{\beta}_j$ is a learned coefficient vector for feature $\mathbf{X}_j$. A logistic regression model is fitted using these covariates to estimate observation probabilities.

A.6.3 Missing Not at Random (MNAR)

In the MNAR setting, the missingness in $\mathbf{X}_j$ depends on the value x_{ij} itself (even if it is missing), in addition to other features. The observation probability is defined as:

$$\mathbb{P}(m_{ij} = 1 \mid x_{ij}, \mathbf{x}_{i,-j}^{\text{obs}}) = \frac{1}{1 + \exp\left(-[x_{ij}, \mathbf{x}_{i,-j}^{\text{obs}}]^\top \boldsymbol{\beta}_j\right)}.$$

Here, x_{ij} is explicitly included as a predictor, distinguishing MNAR from MAR. If x_{ij} is missing during mask generation, it is temporarily imputed (e.g., with the mean) but removed prior to model training.

A.7 Implementations and Hyperparameters

Implementation of models: We implemented all 14 imputation methods based on open access GitHub repository following:

- Mean/Mode: Implemented using the NumPy package.
- MOT (Muzellec et al., 2020): `https://github.com/BorisMuzellec/MissingDataOT`.
- MissForest (Stekhoven & Bühlmann, 2012): Implemented using the missforest package.
- DSAN (Lee & Kim, 2023): `https://github.com/uos-dmlab/Structued-Data-Quality-Analysis/tree/master`.
- DSN (Lee & Kim, 2023): Developed from DSAN by removing the attention layer.
- TabCSDI (Zheng & Charoenphakdee, 2022): `https://github.com/pfnet-research/TabCSDI`.
- Remasker (Du et al., 2024): `https://github.com/tydusky/remasker`.
- DiffPuter (Zhang et al., 2025): Originally available at `https://github.com/hengruizhang98/DiffPuter`, but now removed.
- For HyperImputer (Jarrett et al., 2022), MICE (Van Buuren & Groothuis-Oudshoorn, 2011), MIRACLE (Kyono et al., 2021), SoftImpute (Hastie et al., 2015), MIWAE (Mattei & Frellsen, 2019), and GAIN (Yoon et al., 2018), we use implementations at: `https://github.com/vanderschaarlab/hyperimpute`.

The codes for all methods are available in the anonymous GitHub repository.

Hyperparameter settings of models: Most of the methods included in our benchmark recommend using a single set of hyperparameters across different datasets. For such methods, we adopt the default hyperparameters provided in their official GitHub repositories and ensure sufficient training epochs or steps to achieve convergence of the training loss. The anonymous GitHub repository provides the implementation with the default hyperparameters applied across all methods.

A.8 Experiment Results

All the experiments results, including KS test, JS divergence, and Kendall's W, can be found in the folder of the supplementary material named "Experiment_result".