eISSN 2783-6851

NEW TRENDS in COMPUTER SCIENCES

2026
Volume 4
Issue 1
Pages 131–145
https://doi.org/10.3846/ntcs.2026.26740

IMPACT OF UNIT AND ITEM NONRESPONSE ON POPULATION TOTAL ESTIMATION IN ASYMMETRIC BUSINESS SURVEY DATA

Vilma NEKRAŠAITĖ-LIEGĖ ✉, Elena STRELČIŪNAITĖ

Department of Mathematical Statistics, Vilnius Gediminas Technical University, Vilnius, Lithuania

Article History:
- received 19 April 2026
- accepted 15 June 2026

Abstract. Nonresponse is one of the main sources of nonsampling error affecting the accuracy of survey estimates, particularly in business surveys characterized by highly asymmetric data distributions. This study investigates the impact of unit and item nonresponse on the estimation of population totals using simulation experiments based on enterprise-level research and development (R&D) statistics. Several weighting adjustment methods for unit nonresponse and alternative imputation techniques for item nonresponse were evaluated under different levels and mechanisms of missingness. Estimator performance was assessed using relative bias, coefficient of variation, and relative root mean squared error. The results show that increasing nonresponse levels substantially reduce estimator accuracy, primarily due to increasing relative bias rather than estimator variability. The effectiveness of weighting adjustment methods depends strongly on the response mechanism: class-based response probability estimation performed best under missing completely at random conditions, while random forest–based estimation produced more accurate results when response behavior depended on enterprise size. The performance of imputation methods also varied across missingness mechanisms, with mean-of-donors hot-deck, nearest neighbour, and regression-based imputation performing best under different scenarios. The findings provide practical guidance for selecting appropriate nonresponse adjustment strategies in business surveys with strongly skewed variables.

Keywords: response rate, missing data adjustment, survey simulation, business statistics, hot-deck imputation, random forest, official statistics.

✉Corresponding author. E-mail: *vilma.nekrasaite-liege@vilniustech.lt*

1. Introduction

Survey data are widely used for estimating population characteristics in official statistics, economic analysis, and policy evaluation. However, nonresponse remains one of the main sources of nonsampling error affecting the quality and reliability of survey estimates. When sampled units fail to participate in a survey (unit nonresponse) or omit responses to particular questions (item nonresponse), estimators based solely on observed data may become biased, especially when the missingness mechanism is related to the variables of interest (Little & Rubin, 2019; Lohr, 2021). Various statistical models have been proposed to describe survey nonresponse mechanisms and to reduce the resulting estimation bias (Ayhan, 2023).

The treatment of nonresponse is particularly important in surveys involving asymmetric data distributions. Such datasets typically contain a large number of zero observations together with a small number of extremely large values, which is characteristic of business and innovation surveys, including research and development (R&D) statistics. In these situations, standard estimation and adjustment procedures may perform differently compared with approximately symmetric data, and careful selection of appropriate correction methods becomes essential (Durrant, 2009; Thompson, 2012).

Various approaches have been proposed to reduce the impact of nonresponse on survey estimates. Unit nonresponse is commonly addressed using weighting adjustments based on estimated response probabilities, assuming either equal response behavior across sampled units or homogeneous response patterns within predefined adjustment classes (Bethlehem, 2002; Little & Rubin, 2019; Särndal & Lundström, 2005). More recently, machine learning techniques such as random forests have been applied for response propensity estimation, offering flexible modeling of complex response mechanisms (Biemer & Lyberg, 2003; Larbi et al., 2025; Tivesten et al., 2012). In contrast, item nonresponse is usually handled through imputation methods, including donor-based procedures, nearest neighbor imputation, and regression-based techniques (van Buuren, 2018). The performance of these approaches depends strongly on both the missingness mechanism and the distributional properties of the data.

This study investigates the impact of both unit and item nonresponse on the estimation of population totals in the presence of asymmetric survey data. Using data from the State Data Agency's survey on research and development activities, simulation experiments were conducted to evaluate how different levels and mechanisms of missingness influence estimator accuracy. Several weighting adjustment methods were compared for handling unit nonresponse, including uniform response probability estimation, class-based response probability estimation, and random forest-based estimation. In addition, five imputation techniques were examined for handling item nonresponse, including three variants of hot-deck imputation, nearest neighbor imputation, and regression-based imputation.

The main objective of the study is to assess how estimate quality depends on both the nonresponse rate and the nonresponse mechanism and to determine which adjustment methods perform best under asymmetric data conditions. To achieve this objective, three research questions are addressed in the simulation study:

(U1) How does the rate of nonresponse affect the accuracy of population total estimators?

(U2) How do different unit nonresponse generation mechanisms influence the performance of weighting adjustment methods?

(U3) How do alternative imputation methods perform under item nonresponse in asymmetric survey data?

The results provide practical recommendations for selecting appropriate weighting and imputation strategies in the analysis of business survey data characterized by skewed distributions.

Although the literature includes approaches based on multiple imputation, calibration weighting, double sampling, and variance estimation under nonresponse, the present study focuses specifically on comparing point estimate performance under alternative weighting adjustment and single-imputation procedures.

The remainder of the paper is organized as follows. Section 2 describes the methodology, including the data, sampling design, nonresponse adjustment methods, and simulation framework. Section 3 presents the simulation results, which are discussed in Section 4. Finally, the main conclusions and directions for future research are summarized in Section 5.

2. Methodology

2.1. Data and sampling design

The empirical analysis in this study is based on data from the State Data Agency's 2024 survey on research and development activities. The dataset is treated as a finite population consisting of 2,894 enterprises and includes eleven variables describing enterprise characteristics and research activity indicators.

Three variables – economic activity classification, municipality code, and number of employees – were used for stratification and response modeling. Enterprise size, measured by the number of employees, was positively associated with R&D expenditure, although the relationship was relatively weak (Pearson correlation coefficient approximately 0.20). Economic activity classification and municipality code were treated as categorical auxiliary variables capturing structural differences among enterprises. Study variables describing research and development activity exhibit strong asymmetry with a large proportion of zero values and several extreme observations, which is typical for business survey data (Thompson, 2012).

Although Figure 1 presents the distributions of two study variables, R&D expenditure and R&D personnel, the simulation results reported in the following sections are presented only for R&D expenditure. The observed asymmetry was also reflected in a substantial proportion of zero values, which is characteristic of business survey data. Approximately 55% of enterprises reported zero R&D expenditure and 58% reported zero R&D personnel. The results obtained for the R&D personnel variable showed the same general patterns across all nonresponse scenarios and adjustment methods and therefore are not reported separately in order to avoid repetition.

To improve estimation efficiency, the population was stratified using aggregated auxiliary variables. Economic activity was grouped into seven sectors, municipalities into two regions, and enterprise size into three categories: small, medium-sized, and large enterprises. Combining these variables resulted in 42 strata.

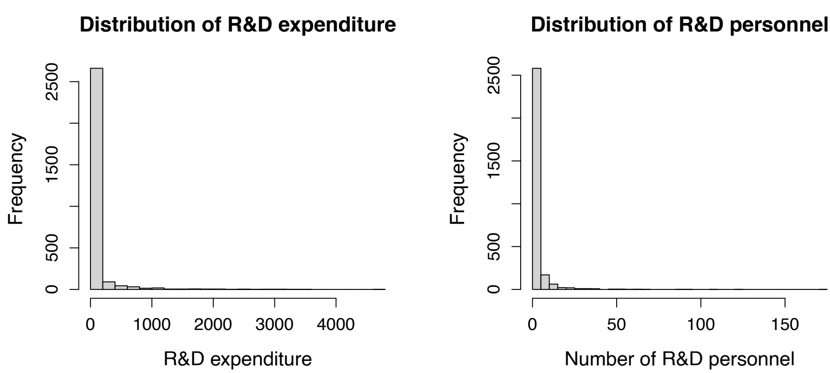

Figure 1. Distributions of selected study variables illustrating the asymmetric structure of R&D survey data

A stratified simple random sampling design without replacement was applied. The total sample size was n = 733, corresponding to approximately 25% of the population. Sample sizes within strata were allocated using Neyman optimal allocation, which distributes observations proportionally to the stratum sizes and within-stratum variability of the study variables in order to minimize the variance of estimators for a fixed total sample size (Krapavickaitė & Plikusas, 2005; Lohr, 2021; Thompson, 2012).

The resulting samples were subsequently subjected to different unit and item nonresponse mechanisms, requiring the application of nonresponse adjustment procedures described below.

2.2. Estimation under nonresponse

Nonresponse in survey sampling may occur either at the unit level, when an entire sampled unit does not respond, or at the item level, when only some study variables are missing for otherwise responding units. Both types of missing data may affect the accuracy of estimators and therefore require appropriate adjustment procedures. In this study, population totals are estimated using weighting adjustments for unit nonresponse and imputation methods for item nonresponse.

Let $U = \{1, \ldots, N\}$ denote a finite population and let $s \subset U$ be a stratified simple random sample with sampling weights $w_i = 1/\pi_i$, where π_i denotes the inclusion probability of the i-th unit. The parameter of interest is the population total

$$t = \sum_{i=1}^{N} y_i.$$

In the absence of nonresponse, the total is estimated by Horvitz-Thompson estimator

$$\hat{t} = \sum_{i \in s} w_i y_i.$$

When nonresponse occurs, modified estimators are required depending on the type of missingness. Estimation of population totals in the presence of nonresponse remains an important topic in survey statistics, especially when follow-up strategies or auxiliary information are available (Stefan & Hidiroglou, 2024).

2.3. Adjustment for unit nonresponse

Unit nonresponse reduces the effective sample size and may introduce bias if response behavior depends on study variables. A common approach to handling unit nonresponse is weighting adjustment based on estimated response probabilities (Brick, 2013; Haziza & Beaumont, 2017; Lohr, 2021; Särndal & Lundström, 2005).

Let $s_a \subset s$ denote the set of responding units. The adjusted estimator of the population total is defined as

$$\hat{t} = \sum_{i \in s_a} \frac{w_i y_i}{\hat{\kappa}_i},$$

where $\hat{t}$ denotes the estimated population total, y_i is the value of the study variable for unit i, w_i is the sampling weight, and $\hat{\kappa}_i$ denotes the estimated response probability of unit i.

Three alternative response probability estimators were considered in this study and are denoted by $\hat{\kappa}_1$, $\hat{\kappa}_2$ and $\hat{\kappa}_3$ in the simulation results and figures.

Uniform response probability estimation ($\hat{\kappa}_1$). Under the assumption that response probabilities are equal for all sampled units, the estimator is

$$\hat{\kappa}_i = \hat{\kappa} = \frac{\sum_{i \in s_a} w_i}{\sum_{i \in s} w_i}.$$

This adjustment corresponds to the missing completely at random assumption and provides unbiased estimates when response behavior does not depend on unit characteristics (Kalton & Kasprzyk, 1986; Little & Rubin, 2019).

Class-based response probability estimation ($\hat{\kappa}_2$). When response behavior varies across subpopulations, response probabilities may be estimated separately within homogeneous adjustment classes constructed from auxiliary variables. Let $h = 1, \ldots, H$ denote adjustment classes. The response probability then is equal to

$$\hat{\kappa}_i = \hat{\kappa}_h = \frac{\sum_{i \in s_{a,h}} w_i}{\sum_{i \in s_h} w_i}.$$

This approach reduces bias when auxiliary variables explain response heterogeneity (Bethlehem, 2002; Haziza & Beaumont, 2017; Särndal & Lundström, 2005).

Random forest-based response probability estimation ($\hat{\kappa}_3$). In addition to classical adjustment methods, response probabilities were estimated using a random forest model. In this study, the predictors included economic activity, municipality group, and enterprise size. Random forest methods allow flexible modeling of nonlinear relationships and interactions without requiring explicit specification of the response mechanism (Biemer & Lyberg, 2003; Larbi et al., 2025; McConville et al., 2017).

2.4. Adjustment for item nonresponse

Item nonresponse occurs when sampled units provide incomplete responses for selected study variables. In such cases, missing values may be replaced using imputation procedures based on observed data and auxiliary information (Kalton & Kasprzyk, 1986; Little & Rubin, 2019; van Buuren, 2018).

After imputation, the estimator of the population total takes the form

$$\hat{t} = \sum_{i \in s_a} w_i y_i + \sum_{i \in s_n} w_i \hat{y}_i,$$

where s_n denotes the set of units with missing values and $\hat{y}_i$ denotes the imputed value. Five imputation methods were investigated and are denoted by iv_1 – iv_5 in the simulation results and figures.

Random donor hot-deck imputation (iv_1) replaces each missing value with a randomly selected observed value from the same adjustment group (Andridge & Little, 2010).

Mean-of-donors hot-deck imputation (iv_2) replaces missing values with the average of several randomly selected donor observations in order to reduce variability (de Waal et al., 2011).

Ordered hot-deck imputation (iv_3) replaces missing values using the nearest available observed value within groups ordered according to selected auxiliary variables. This approach preserves local structure in the data and is particularly suitable for asymmetric distributions (Andridge & Little, 2010; Durrant, 2009).

Nearest neighbor imputation (iv_4) replaces missing values using values from units that are most similar with respect to auxiliary variables (Chen & Shao, 2000). Although widely used in practice, its performance may deteriorate in highly skewed datasets.

Regression-based imputation (iv_5) estimates missing values using a linear regression model fitted to complete observations (Kalton & Kasprzyk, 1986; Kim & Shao, 2021). Missing values are replaced by predicted values obtained from the fitted model. In this study, deterministic regression imputation was applied; that is, no additional random error term was added to the predicted values.

2.5. Simulation design

A simulation study was conducted to evaluate the performance of alternative nonresponse adjustment methods under different missingness scenarios. Both unit and item nonresponse were introduced artificially in order to analyze their effect on the accuracy of population total estimators.

Four levels of nonresponse were considered: 10%, 20%, 30%, 40%. Each scenario was replicated 1000 times to ensure stable estimates of bias, variance, and relative mean squared error.

Several mechanisms governing the generation of unit nonresponse were examined. First, missing completely at random (MCAR) nonresponse was simulated by assigning equal response probabilities to all sampled units. Second, response probability was assumed to depend on enterprise size (size-dependent), reflecting realistic response behavior observed in business surveys. Third, response probability was allowed to depend on selected auxiliary variables, namely economic activity, municipality group, and enterprise size (auxiliary-var.-dependent), corresponding to a missing at random mechanism (Kim & Shao, 2021; Little & Rubin, 2019).

Item nonresponse was simulated independently within responding units using mechanisms analogous to those applied in the unit nonresponse setting, namely MCAR, size-dependent and auxiliary-var.-dependent missingness mechanisms. The resulting incomplete datasets were analyzed using the imputation methods described in Section 2.4.

Estimator performance was evaluated using relative bias (RBias), coefficient of variation (cv), and relative root mean squared error (rRMSE), computed across simulation replications. Let $\hat{t}^{(r)}$ denote the estimator obtained in the r-th simulation run and t the true population total. The relative bias of the estimator was calculated as

$$RBias\left(\hat{t}\right) = \frac{\frac{1}{R}\sum_{r=1}^{R}\hat{t}^{(r)} - t}{t},$$

the coefficient of variation of the estimator was estimated as

$$cv\left(\hat{t}\right) = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}\left(\hat{t}^{(r)} - \bar{\hat{t}}\right)^2}}{t},$$

where $\bar{\hat{t}}$ denotes the mean of simulated estimates. The relative root mean squared error was computed as

$$rRMSE\left(\hat{t}\right) = \sqrt{RBias\left(\hat{t}\right)^2 + cv\left(\hat{t}\right)^2}.$$

These measures provide complementary information about estimate quality and are commonly used in simulation studies of survey sampling methods (Lohr, 2021).

Figure 2 summarizes the overall simulation framework used in this study. First, samples were selected from the stratified finite population using Neyman allocation. Unit and item nonresponse were then generated under three alternative missingness mechanisms: MCAR, size-dependent, and auxiliary-variable-dependent. The resulting incomplete datasets were adjusted using the weighting and imputation methods described in Sections 2.3 and 2.4. Population totals were subsequently estimated and evaluated using relative bias (RBias), coefficient of variation (cv), and relative root mean squared error (rRMSE). Finally, the performance of the alternative adjustment methods was compared with respect to the three research objectives of the study.

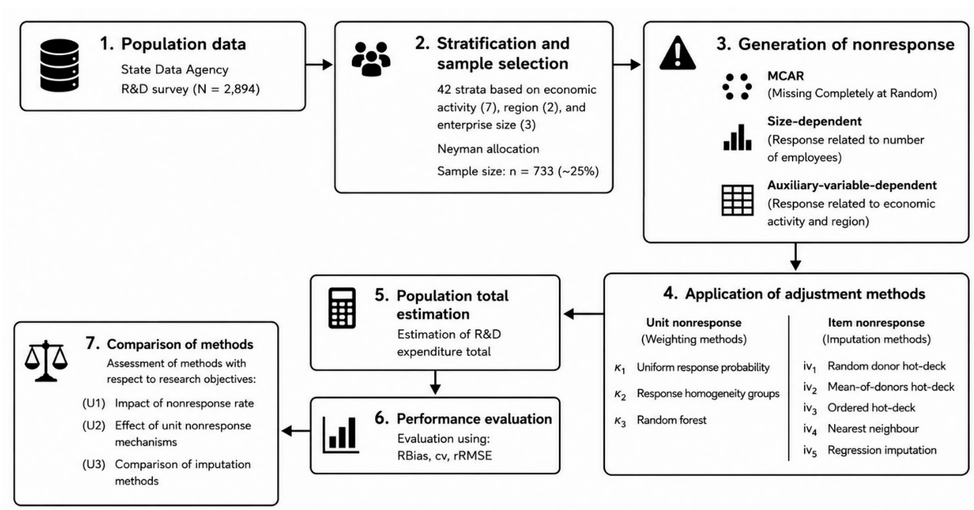

Figure 2. Overview of the simulation framework used to evaluate nonresponse adjustment methods

3. Results

The results are presented according to the three research objectives formulated in Introduction: the effect of nonresponse rate (U1), the effect of unit nonresponse mechanisms (U2), and the comparison of item nonresponse imputation methods (U3).

3.1. Effect of the nonresponse level on estimator accuracy

The first part of the simulation study examined how increasing levels of nonresponse affect the accuracy of population total estimators. Four nonresponse scenarios were considered, ranging from 10% to 40% missingness. In this part of the simulation study, nonresponse was generated under the MCAR mechanism in order to isolate the effect of increasing nonresponse rate on estimator accuracy independently of response heterogeneity. The distributions of the estimated population totals are illustrated in Figure 3, while the corresponding performance measures (RBias, cv, and rRMSE) are summarized in Table 1.

Table 1 confirms that estimator accuracy deteriorates as the nonresponse rate increases. In particular, the estimator without nonresponse adjustment shows a rapid increase in relative root mean squared error as the nonresponse rate grows, indicating substantial bias when missing observations are ignored.

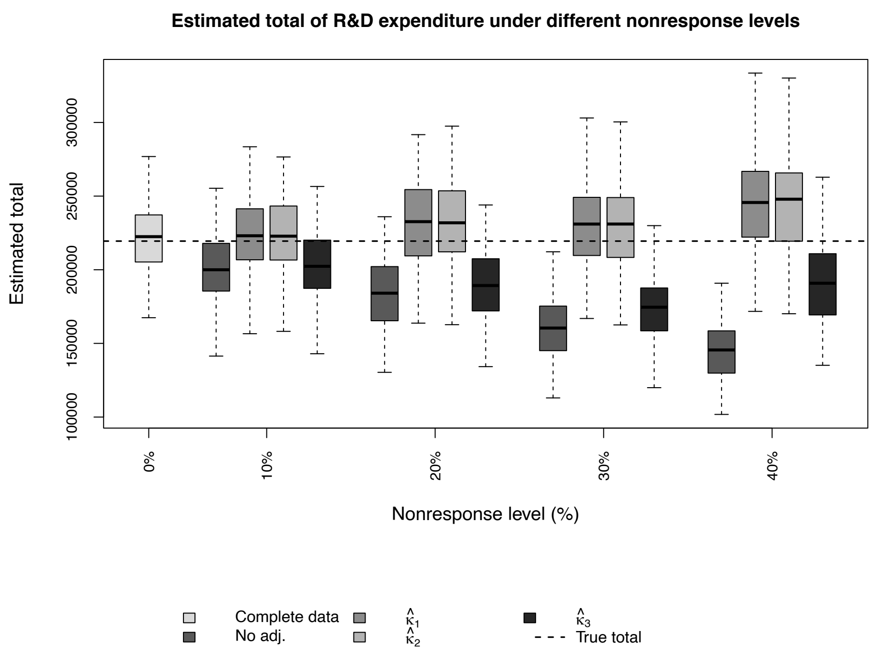

Figure 3. Distribution of estimated population totals of R&D expenditure under different nonresponse levels and adjustment methods ($\hat{\kappa}_1 - \hat{\kappa}_3$)

Table 1. Performance of population total estimator for R&D expenditure under different MCAR nonresponse levels

Nonresponse level	Method	$RBias\left(\hat{t}\right)$	$cv\left(\hat{t}\right)$	$rRMSE\left(\hat{t}\right)$
Complete data	No adjustment	0.008	0.110	0.110
10%	No adjustment	–0.082	0.103	0.131
10%	$\hat{\kappa}_1$	0.021	0.114	0.116
10%	$\hat{\kappa}_2$	0.025	0.111	0.114
10%	$\hat{\kappa}_3$	–0.071	0.103	0.125
20%	No adjustment	–0.158	0.103	0.189
20%	$\hat{\kappa}_1$	0.060	0.129	0.143
20%	$\hat{\kappa}_2$	0.066	0.124	0.141
20%	$\hat{\kappa}_3$	–0.132	0.105	0.168
30%	No adjustment	–0.266	0.100	0.284
30%	$\hat{\kappa}_1$	0.058	0.143	0.154
30%	$\hat{\kappa}_2$	0.052	0.134	0.143
30%	$\hat{\kappa}_3$	–0.204	0.106	0.230
40%	No adjustment	–0.338	0.094	0.351
40%	$\hat{\kappa}_1$	0.119	0.157	0.197
40%	$\hat{\kappa}_2$	0.121	0.149	0.191
40%	$\hat{\kappa}_3$	–0.126	0.140	0.188

Among the evaluated weighting adjustment methods, $\hat{\kappa}_2$ produced the smallest relative root mean squared error values across most nonresponse levels, while $\hat{\kappa}_1$ showed very similar performance. In contrast, $\hat{\kappa}_3$ resulted in noticeably larger estimation errors, especially when the nonresponse rate exceeded 30%.

The results obtained under the MCAR mechanism show that relative bias increases substantially as the nonresponse rate grows, whereas the increase in estimator variability measured by the coefficient of variation remains comparatively moderate. As a consequence, the relative root mean squared error of the estimators is driven primarily by the increase in relative bias rather than variability.

These findings indicate that ignoring nonresponse or applying inadequate adjustment procedures may lead to considerable deterioration in estimator accuracy even when the sampling design itself remains efficient. The results highlight the importance of selecting appropriate nonresponse correction methods, particularly when the proportion of missing data exceeds moderate levels.

3.2. Effect of unit nonresponse mechanisms on weighting adjustments

The second part of the simulation study evaluated the performance of different response probability estimation methods under alternative unit nonresponse mechanisms. The results are illustrated in Figure 4 and further summarized numerically in Table 2 using the performance measures RBias, cv, and rRMSE.

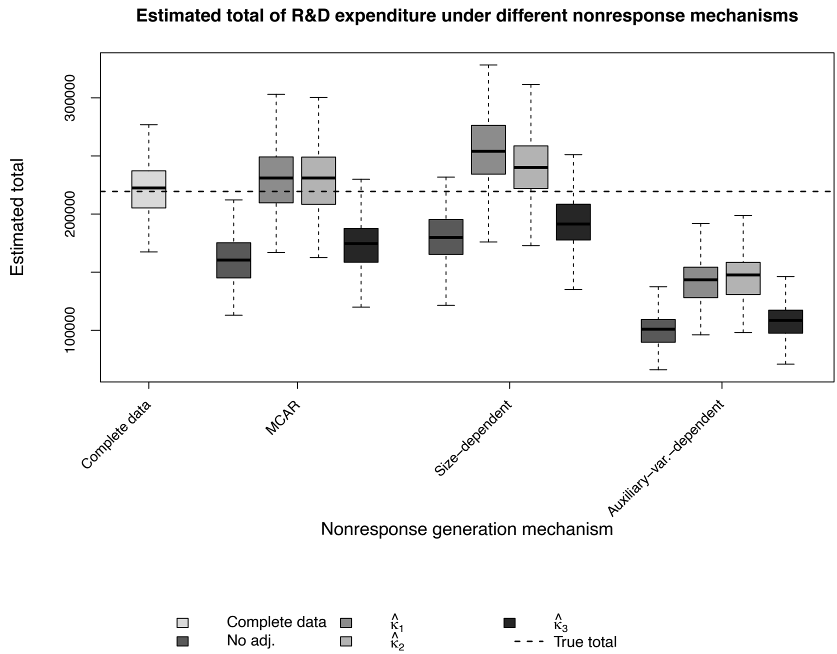

Figure 4. Distribution of estimated population totals of R&D expenditure across 1000 simulation replications under different nonresponse generation mechanisms and adjustment methods ($\hat{\kappa}_1 - \hat{\kappa}_3$)

Table 2. Performance of population total estimator for R&D expenditure under different unit nonresponse mechanisms (30% nonresponse)

Nonresponse mechanism	Method	$RBias(\hat{t})$	$cv(\hat{t})$	$rRMSE(\hat{t})$
Complete data	No adjustment	0.008	0.110	0.110
MCAR	No adjustment	–0.266	0.100	0.284
MCAR	$\hat{\kappa}_1$	0.058	0.143	0.154
MCAR	$\hat{\kappa}_2$	0.051	0.134	0.143
MCAR	$\hat{\kappa}_3$	–0.204	0.106	0.230
Size-dependent	No adjustment	–0.178	0.101	0.204
Size-dependent	$\hat{\kappa}_1$	0.170	0.143	0.222
Size-dependent	$\hat{\kappa}_2$	0.106	0.136	0.172
Size-dependent	$\hat{\kappa}_3$	–0.120	0.107	0.160
Auxiliary-variable-dependent	No adjustment	–0.544	0.071	0.549
Auxiliary-variable-dependent	$\hat{\kappa}_1$	–0.350	0.101	0.364
Auxiliary-variable-dependent	$\hat{\kappa}_2$	–0.336	0.101	0.351
Auxiliary-variable-dependent	$\hat{\kappa}_3$	–0.511	0.074	0.516

Table 2 shows that estimator performance depends not only on the level of nonresponse but also on the mechanism generating missing data. In particular, ignoring nonresponse leads to substantial deterioration in estimator accuracy across all mechanisms, with especially large relative root mean squared error values observed under the auxiliary-variable-dependent scenario.

The deterioration in estimator accuracy is particularly pronounced when the response mechanism depends on auxiliary variables related to research activity, indicating increased sensitivity of weighting adjustments to complex response structures.

When nonresponse was generated completely at random (MCAR), the response homogeneity groups probability adjustment method ($\hat{\kappa}_2$) produced the smallest relative root mean squared error, while $\hat{\kappa}_1$ showed slightly larger but comparable values. In contrast, $\hat{\kappa}_3$ resulted in noticeably larger estimation errors under this mechanism.

When response probability depended on enterprise size, the random forest-based response probability estimator ($\hat{\kappa}_3$) provided the most accurate results among the evaluated methods, yielding the smallest rRMSE values. This suggests that more flexible modeling of response behavior may improve estimator performance when response mechanisms depend on structural enterprise characteristics.

Under the auxiliary-variable-dependent mechanism, all evaluated adjustment methods showed substantially larger estimation errors compared with the MCAR and size-dependent scenarios. Although $\hat{\kappa}_2$ slightly outperformed the alternative weighting methods, the results indicate that complex response mechanisms remain difficult to correct using standard response probability adjustments.

Overall, the results demonstrate that correct identification of the underlying nonresponse mechanism plays a key role in selecting an appropriate weighting adjustment strategy.

3.3. Comparison of imputation methods for item nonresponse

The third part of the simulation study focused on the performance of alternative imputation methods for handling item nonresponse in asymmetric data. The results are illustrated in Figure 5 and further summarized numerically in Table 3 using the performance measures RBias, cv, and rRMSE.

Table 3 shows that estimator accuracy depends strongly on both the imputation method and the nonresponse mechanism. Under the MCAR mechanism, the mean-of-donors hot-deck imputation method (iv_2) produced the smallest relative root mean squared error, while nearest neighbour imputation (iv_4) resulted in substantially larger estimation errors.

When item nonresponse depended on enterprise size, the nearest neighbor imputation method (iv_4) yielded the most accurate estimates, followed by ordered hot-deck imputation (iv_3). In contrast, the random donor and mean-of-donors hot-deck methods (iv_1 and iv_2) produced noticeably larger estimation errors under this mechanism.

Under the auxiliary-variable-dependent mechanism, all imputation methods showed increased estimation errors compared with the MCAR and size-dependent scenarios. Among the evaluated approaches, regression imputation (iv_5) produced the smallest rRMSE values, whereas nearest neighbor imputation (iv_4) again showed the weakest performance.

Overall, the results indicate that the performance of imputation methods depends strongly on the structure of the missingness mechanism. Donor-based procedures that incorporate ordering with respect to auxiliary variables perform well when response behavior depends on enterprise characteristics, whereas regression-based imputation provides more stable results when missingness is related to multiple auxiliary variables.

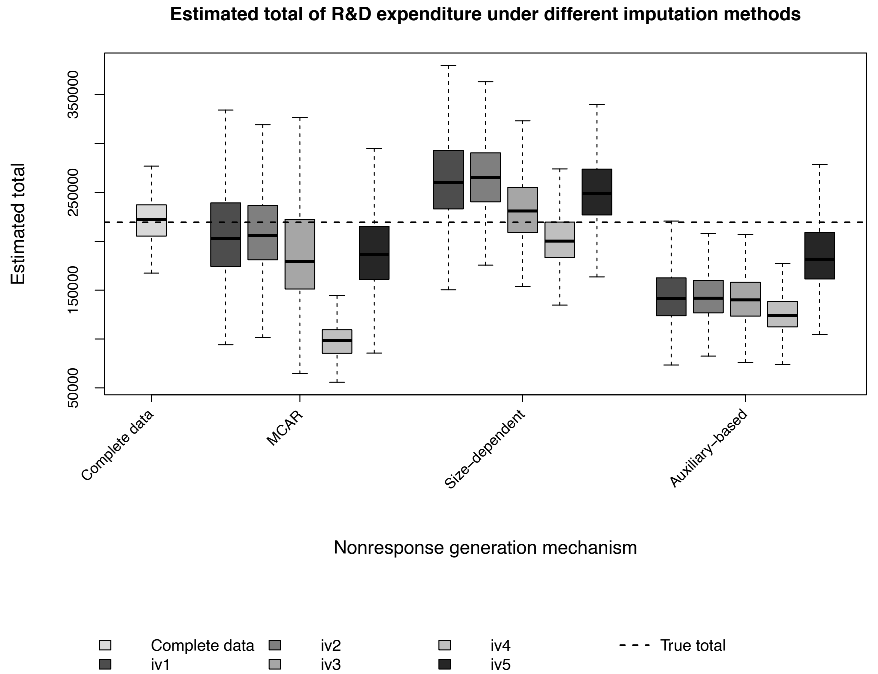

Figure 5. Distribution of estimated population totals of R&D expenditure across 1000 simulation replications under different imputation methods (iv_1 – iv_5) and nonresponse mechanisms

Table 3. Performance of population total estimator for R&D expenditure under different imputation methods and nonresponse mechanisms

Nonresponse mechanism	Method	$RBias(\hat{t})$	$cv(\hat{t})$	$rRMSE(\hat{t})$
Complete data	No adjustment	0.008	0.110	0.110
MCAR	iv_1	–0.049	0.222	0.227
MCAR	iv_2	–0.048	0.195	0.200
MCAR	iv_3	–0.129	0.266	0.295
MCAR	iv_4	–0.549	0.087	0.556
MCAR	iv_5	–0.134	0.209	0.248
Size-dependent	iv_1	0.205	0.196	0.284
Size-dependent	iv_2	0.214	0.171	0.274
Size-dependent	iv_3	0.063	0.153	0.166
Size-dependent	iv_4	–0.082	0.121	0.146
Size-dependent	iv_5	0.140	0.155	0.210
Auxiliary-variable-dependent	iv_1	–0.341	0.129	0.365
Auxiliary-variable-dependent	iv_2	–0.344	0.110	0.361
Auxiliary-variable-dependent	iv_3	–0.351	0.119	0.371
Auxiliary-variable-dependent	iv_4	–0.426	0.088	0.435
Auxiliary-variable-dependent	iv_5	–0.148	0.166	0.222

4. Discussion

The simulation results demonstrate that nonresponse has a substantial impact on the accuracy of population total estimators when analyzing asymmetric business survey data. The effect becomes particularly pronounced as the proportion of missing observations increases. One of the main findings of the study is that relative bias grows more rapidly than estimator variability measured by the coefficient of variation as nonresponse levels increase. This indicates that systematic deviations caused by missing data mechanisms represent a more important source of error than sampling variability under the considered conditions.

The comparison of weighting adjustment methods for unit nonresponse showed that estimator performance strongly depends on the assumed response mechanism. When nonresponse was generated completely at random, the response homogeneity group estimator ($\hat{\kappa}_2$) produced the most accurate results, while the uniform adjustment method ($\hat{\kappa}_1$) showed very similar performance. In contrast, when response behavior depended on enterprise size, the random forest-based response probability estimator ($\hat{\kappa}_3$) yielded the smallest estimation errors, indicating that more flexible modeling approaches may be advantageous when response mechanisms depend on structural enterprise characteristics. However, under the auxiliary-variable-dependent mechanism all weighting adjustment methods showed noticeably larger estimation errors, suggesting that complex response structures remain difficult to correct using standard response propensity models.

Although random forest-based response probability estimation provides a flexible framework capable of capturing nonlinear relationships between response indicators and auxiliary variables, its advantages were not consistent across all scenarios considered in this study. One possible explanation is that the variables used for stratification already captured a substantial part of response heterogeneity, thereby limiting the potential improvement achievable through more complex modeling approaches. Similar observations have been reported in previous studies of response propensity modeling in survey statistics (Bethlehem, 2002; Brick, 2013).

The results obtained for item nonresponse indicate that the performance of imputation methods depends strongly on the structure of the missingness mechanism. Under the MCAR mechanism, the mean-of-donors hot-deck imputation method (iv_2) produced the smallest estimation errors. When item nonresponse depended on enterprise size, the nearest neighbour imputation method (iv_4) achieved the best results, followed by ordered hot-deck imputation (iv_3). In contrast, when missingness depended on multiple auxiliary variables, regression-based imputation (iv_5) provided the most accurate estimates among the evaluated approaches. These findings suggest that no single imputation method performs best across all missingness scenarios in asymmetric datasets. Differences in estimator performance across methods appear to arise from their ability to preserve the distributional characteristics of asymmetric variables and to exploit auxiliary information under different missingness mechanisms.

Overall, the results highlight the importance of selecting nonresponse adjustment methods that are consistent with both the missingness mechanism and the distributional properties of the data. In particular, class-based weighting adjustment methods and flexible imputation procedures incorporating auxiliary information appear to provide robust solutions when their assumptions are consistent with the underlying missingness mechanism.

From a practical perspective, weighting adjustment methods based on response classes remain computationally simpler and easier to implement in large-scale official surveys than machine learning-based approaches. Although random forest-based adjustment showed improved performance under selected response mechanisms, its implementation requires additional computational effort, model tuning, and predictor selection. Therefore, the choice of adjustment method should consider not only estimation accuracy but also operational feasibility in large-scale survey production.

5. Conclusions

This study investigated the impact of unit and item nonresponse on the estimation of population totals in asymmetric survey data using simulation experiments based on enterprise-level research and development statistics. The performance of several weighting adjustment and imputation methods was evaluated under different levels and mechanisms of missingness. The present study focused on point estimate accuracy rather than variance estimation under nonresponse. Future research could extend the analysis by investigating variance estimation procedures under nonresponse, including resampling approaches such as bootstrap or jackknife.

The results confirm that increasing nonresponse levels substantially reduce estimator accuracy, primarily through increased relative bias rather than estimator variability. This highlights the importance of applying appropriate nonresponse adjustment procedures even when efficient sampling designs are used.

The simulation study demonstrated that the effectiveness of weighting adjustments depends strongly on the underlying response mechanism. Class-based response probability estimation performed best under MCAR conditions, while random forest–based response probability estimation provided more accurate results when response behavior depended on enterprise size. However, none of the evaluated weighting approaches fully compensated for complex auxiliary-variable-dependent response mechanisms.

Similarly, the performance of imputation methods depended on the structure of item nonresponse. Mean-of-donors hot-deck imputation performed best under MCAR conditions, nearest neighbor imputation under size-dependent missingness, and regression-based imputation under auxiliary-variable-dependent mechanisms. These findings indicate that no single imputation method is universally optimal for asymmetric business survey data.

Overall, the results provide practical guidance for selecting nonresponse adjustment strategies that are consistent with both the missingness mechanism and the distributional properties of the data. Future research could extend the analysis by investigating additional response mechanisms, alternative machine learning approaches, multiple imputation techniques, and imputation models adapted for zero-inflated variables such as Tobit or Heckman approaches.

References

Andridge, R. R., & Little, R. J. A. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, *78*(1), 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x

Ayhan, H. Ö. (2023). Models for survey nonresponse and bias adjustment techniques. *Statistics in Transition New Series, 24*(3), 139–149. https://doi.org/10.59170/stattrans-2023-040

Bethlehem, J. (2002). Weighting nonresponse adjustments based on auxiliary information. In R. M. Groves,D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 275–288). Wiley.

Biemer, P. P., & Lyberg, L. E. (2003). *Introduction to survey quality*. Wiley. https://doi.org/10.1002/0471458740

Brick, J. M. (2013). Unit nonresponse and weighting adjustments: A critical review. *Journal of Official Statistics*, *29*(3), 329–353. https://doi.org/10.2478/jos-2013-0026

Chen, J., & Shao, J. (2000). Nearest neighbor imputation for survey data. *Journal of Official Statistics*, *16*(2), 113–131.

de Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of statistical data editing and imputation*. Wiley. https://doi.org/10.1002/9780470904848

Durrant, G. B. (2009). Imputation methods for handling item-nonresponse in practice: Methodological issues and recent debates. *International Journal of Social Research Methodology*, *12*(4), 293–304. https://doi.org/10.1080/13645570802394003

Haziza, D., & Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, *32*(2), 206–226. https://doi.org/10.1214/16-STS608

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, *12*(1), 1–16.

Kim, J. K., & Shao, J. (2021). *Statistical methods for handling incomplete data*. CRC Press. https://doi.org/10.1201/9780429321740

Krapavickaitė, D., & Plikusas, A. (2005). *Imčių teorijos pagrindai*. Technika.

Larbi, K., Tsang, J., Haziza, D., & Dagdoug, M. (2025). *On the use of machine learning methods for the treatment of unit nonresponse in surveys*. Survey Methodology. Statistics Canada.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Lohr, S. L. (2021). *Sampling*: *Design and analysis* (3rd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9780429298899

McConville, K. S., Breidt, F. J., Lee, T. C. M., & Moisen, G. G. (2017). Model-assisted survey regression estimation with the LASSO. *Journal of Survey Statistics and Methodology*, *5*(2), 131–158. https://doi.org/10.1093/jssam/smw041

Särndal, C.-E., & Lundström, S. (2005). *Estimation in surveys with nonresponse*. Wiley. https://doi.org/10.1002/0470011351

Stefan, M., & Hidiroglou, M. A. (2024). Estimation of a population total under nonresponse using follow-up. *Journal of Survey Statistics and Methodology, 12*(5), 1365–1388. https://doi.org/10.1093/jssam/smae002

Thompson, S. K. (2012). *Sampling* (3rd ed.). Wiley. https://doi.org/10.1002/9781118162934

Tivesten, E., Jonsson, S., Jakobsson, L., & Norin, H. (2012). Nonresponse analysis and adjustment in a mail survey on car accidents. *Accident Analysis* & *Prevention*, *48*, 401–415. https://doi.org/10.1016/j.aap.2012.02.017

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9780429492259