Mathematics

AIMS Mathematics, 11(5): 13008–13022.
DOI: 10.3934/math.2026535
Received: 19 January 2026
Revised: 27 March 2026
Accepted: 27 March 2026
Published: 11 May 2026

https://www.aimspress.com/journal/Math

Research article

Robust population estimation under PPS sampling: An application to engineering data

L. S. Diab[1], Norah D. Alshahrani[2], Majdah Mohammed Badr[3] and Sohaib Ahmad[4,*]

[1] Department of Mathematics and Statistics, College of Science, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia
[2] Department of Mathematics, College of Science, University of Bisha, P.O. Box 551, Bisha 61922, Saudi Arabia
[3] Department of Mathematics and Statistics, College of Science, University of Jeddah, Jeddah, Saudi Arabia
[4] Department of Statistics, Abdul Wali Khan University, Mardan, Pakistan

* **Correspondence:** Email: sohaib_ahmad@awkum.edu.pk; Tel: +923208269320.

Abstract: The probability proportional to size (PPS) sampling method is frequently applied in engineering and industrial surveys when the auxiliary size measures can be obtained, and the population units are substantially heterogeneous. Nevertheless, classical PPS designs estimators are very susceptible to outliers and influential observations, which are common in engineering data because of measurement errors, extreme operating conditions, or structural variability. This research paper presents a powerful estimation method of the finite population mean using PPS sampling on auxiliary information. The proposed estimator has robust influence functions in the design-based PPS framework, enabling one to balance the impacts of extreme sample units by means of controlled weighting without breaking the sampling design on which it relies. Theoretical properties, such as design unbiasedness, consistency, and asymptotic variance of the estimator, are explored and compared with the traditional PPS estimators. Numerical findings show that the robust estimator significantly decreases the mean squared error compared to existing counterparts. The practical usefulness of the suggested methodology is demonstrated on an engineering dataset on the topic of operational performance measurements. Empirical findings prove the effectiveness of the robust estimator in delivering stable and reliable population estimates as opposed to the conventional PPS-based approaches. The results show that the concept of robustness remains critical in survey estimation when applied to engineering problems and can offer a convenient analysis device to

analysts who need to work with heterogeneous and contaminated data. The proposed method provides an effective and flexible alternative to strong population estimation where the PPS sampling designs are applicable.

Keywords: PPS sampling; robust estimator; multi-auxiliary variables; optimum values; efficiency
Mathematics Subject Classification: 62D05

1. Introduction

Proper estimation of population characteristics is a core goal of survey sampling, especially in all types of engineering and industrial research where robust inference is used to aid the decision-making process, quality control, and optimization of the system. In most practical scenarios, the population is extremely heterogeneous, and auxiliary information comes in the form of size measures including production capacity, energy usage, machine throughput, or structural dimensions. Probability proportional to size (PPS) sampling has thus gained significant popularity in engineering surveys, where it is possible to have a higher selection probability to units with larger sizes; therefore, better estimation efficiency is attained when the study variable is positively correlated with the size measure. Although it has benefits in efficiency, PPS sampling is not exempt from practical problems. With engineering systems increasingly becoming more complex and in scale, such anomalies have become more frequent, which highlights the necessity to have strong estimation methods in PPS sampling frameworks.

Strong statistics provide a conceptually sound method of reducing the impact of outliers, being efficient with standard data. Virtuous robustness has been investigated in great depth over the last several decades in both independent and identically distributed observations, as well as in regression and time series models. Nevertheless, the evolution of the strong approaches to complicated survey construction, especially PPS sampling, is comparatively meager. It is difficult to introduce the concept of robustness into design-based inference, since survey estimators should observe the probabilistic structure that the sampling design imposes on them, such as unequal selection probabilities and design weights. Current strategies for robust survey estimation have been more or less concentrated on simple random sampling designs. Other authors have suggested weight trimming, winsorization, or calibration-based methods to mitigate the effects of extreme survey weights. Although these techniques can be used to enhance estimator stability, they can commonly be ad hoc and could violate design-unbiasedness or interpretability. The influential observation problem is worsened in the context of PPS sampling, in which the probability of selection is unequal and highly fluctuating. Large size of units not only increases their chances of inclusion but also their large design weights, so robustness is a crucial issue. The authors in [1] discussed a robust estimator for estimating the mean under PPS sampling. [2] used a predicative approach based on PPS sampling using auxiliary information. [3] developed an improved estimator based on auxiliary information. [4] discussed double-sampling PPS sampling using auxiliary information. [5] suggested an improved family of estimators under PPS sampling. The authors in [6] discussed a penalized spline-based estimator under PPS sampling. [7] discussed PPS sampling with and without measurement error. [8] a robust regression-type estimator based on PPS sampling. [9] suggested a new approach to enhance efficiency of an estimator under PPS sampling. [10] recommended a new estimators based on PPS

sampling using measurement errors. [11] Presents a generalized class of estimators using two auxiliary variables. [12] suggested a new class of estimators for estimating the cumulative distribution function (CDF) under PPS sampling.

Engineering applications also pose special obstacles, thus encouraging the design of efficient PPS estimators. Engineering populations are often heavily tailed and highly nonlinear in correlation between the study and the auxiliary variables. To give a concrete example, large-scale components can be determinant to the behavior of the infrastructure being monitored: in manufacturing, process extremes might represent rare but significant process failures; and in the energy system, peak loads can be higher than the normal working loads. Traditional PPS estimators can be sensitive to such extremes and, therefore, give unstable estimates that do not depict the whole population structure.

This paper attempts to respond to these issues by proposing a well-developed estimation procedure that deals with a finite population under PPS sampling. The suggested methodology alters classical PPS estimators, introducing strong influence functions that curtail the influence of extreme sample units, but still retains the major design-based characteristics of the estimator. In contrast to all-model-based solutions, the given approach will be applied in the design-based framework so that the inference will be valid no matter what population distribution is used. It is especially critical in the engineering case, when assumptions about the model can be hard to check or can differ among systems and applications. Theoretical characteristics of the suggested robust estimator are discussed thoroughly. Design unbiasedness and consistency are proved under weak regularity conditions, and expressions of the asymptotic variance are given for statistical inference. An effective variance estimation process is also presented to explain the altered influence structure. Such theoretical findings give a solid basis to the practical application of the estimator and explain why the proposed approach is different than heuristic robustness methods prevalent in the application of survey analysis.

Several population cases are factored in, such as different rates of correlation between the study variable and the measure of size, different rates of contamination, and other possible outlier-generating processes. The robust estimator is compared with classical PPS estimators regarding its performance in terms of bias, variance, and mean squared error. The experimental applicability of the suggested methodology is also exemplified with real engineering data. The dataset contains operational measurements with large variability and extreme values, which are realistic engineering conditions. Experimental evidence demonstrates that the powerful PPS estimator generates more stable and interpretable population estimates than traditional estimates, which makes it useful to engineering practitioners and analysts.

1.1. Research gap

The probability proportional to size (PPS) sampling method is widely applied in engineering and industrial surveys because it is effective when one has a heterogeneous population where auxiliary size measures are available. However, well-established classical PPS estimators perform poorly in the presence of influential observations. Such irregularities are especially likely to occur in engineering data due to severe conditions of operation, errors in measurements, and structural inconsistency. This does not overrule the fact that current approaches toward PPS estimation rely, to a large extent, on assumptions of clean data, being very vulnerable to contamination.

Strong statistical tools have been extensively built to support independent and identically

distributed data and regression-based models; however, these have not been extensively extended to support complicated survey designs and, in particular, PPS sampling. The majority of the powerful survey estimation methods are based on basic random or stratified sampling, leaving PPS sampling quite unexplored. The unbalanced selection probabilities and the extremely varying weights of design in PPS sampling provide an extra set of difficulties not sufficiently met by the available robust methods.

Existing efforts to manage PPS sampling are typically based on ad hoc methods, including weight trimming, winsorization, or post-sampling corrections. Although such techniques can enhance numerical stability, they can often be poorly supported by theory, be contrary to design-based principles, and lead to bias or inconsistent estimators. Additionally, there has been little focus on robust variance estimation under PPS designs, in which case the statistical inference is restricted to a case of limited reliability in the circumstances where robustness corrections are implemented. The other significant gap is that not that many robust PPS estimators are used on real engineering data. Most methodological literature is based on artificial data or model examples of surveys that do not show the practical working in real-world engineering settings with heavy-tailed data distribution and complicated populations. Therefore, the generalization of robust PPS approaches to engineering decision-making has not been adequately proven. Moreover, little comparative performance analysis that systematically determines robustness efficiency trade-offs at different levels of contamination and size-variable relationships is available. Extensive literature has not been done to determine whether robust PPS estimators can be efficient in uncontaminated conditions and offer protection against extreme values.

Overall, this leaves a definite gap in a theoretically sound, design-consistent, robust estimation framework specific to PPS sampling, with valid variance estimation and validated to work with engineering data. Sealing these loopholes would greatly increase the accuracy of the population estimation in surveys conducted in engineering operations and would help in the development of a sound survey sampling methodology.

The rest of this paper is structured in the following way: Section 2 presents the methodology, notations, and symbols under PPS sampling designs. Section 3 presents some well-known existing estimators. Section 4 presents the proposed robust estimator and elaborates on the theoretical properties of the estimator. Section 5 provides the numerical study and compares estimators. Section 6 presents the discussion of the work.

2. Materials and methods

Let a population $K= \{K_1, K_2,\ldots, K_N\}$ consist of N recognizable components. The study variable is represented by Y, where X, Z, and Q are the auxiliary variables. The auxiliary variables are positively correlated, and the correlation between the study and the rank of the auxiliary variables is minimal. A sample of size n is selected with the help of PPS sampling without replacement. Let $P_i=\frac{Q_i}{\sum_{i=1}^{n} Q_i}$ be the PPS sampling for the $i^{t\square}$ units.

$$u_i=\frac{y_i}{NP_i}, \quad v_i=\frac{x_i}{NP_i}, \quad w_i=\frac{z_i}{NP_i},$$

$$\bar{u}=\frac{\sum_{i=1}^{n} u_i}{n}=\bar{y}_{pps}, \ \bar{v}=\frac{\sum_{i=1}^{n} v_i}{n}=\bar{x}_{pps}, \ \bar{w}=\frac{\sum_{i=1}^{n} z_i}{n}=\bar{z}_{pps},$$

$E(\zeta_0^2)= \frac{C_u^2}{n}$, $E(\zeta_1^2)= \frac{C_v^2}{n}$, $E(\zeta_2^2)= \frac{C_z^2}{n}$,

$E(\zeta_0\zeta_1)= \Lambda\rho_{uv}C_uC_v$, $E(\zeta_0\zeta_2)= \Lambda\rho_{uz}C_uC_z$, $E(\zeta_1\zeta_2)= \Lambda\rho_{vz}C_vC_z$,

$\rho_{uv}=\frac{\sum_{i=1}^{N}P_i(u_i-\bar{Y})(v_i-\bar{X})}{S_uS_v}$, $\rho_{uw}=\frac{\sum_{i=1}^{N}P_i(u_i-\bar{Y})(w_i-\bar{Z})}{S_uS_w}$, $\rho_{vw}=\frac{\sum_{i=1}^{N}P_i(v_i-\bar{X})(w_i-\bar{Z})}{S_uS_w}$

$S_u^2=\sum_{i=1}^{N}P_i(u_i-\bar{Y})^2$, $S_v^2=\sum_{i=1}^{N}P_i(v_i-\bar{X})^2$, $S_w^2=\sum_{i=1}^{N}P_i(w_i-\bar{Z})^2$,

$S_u=\sqrt{\sum_{i=1}^{N}P_i(u_i-\bar{Y})^2}$, $S_v=\sqrt{\sum_{i=1}^{N}P_i(v_i-\bar{X})^2}$, $S_w=\sqrt{\sum_{i=1}^{N}P_i(w_i-\bar{Z})^2}$,

$S_{uv}=\sum_{i=1}^{N}P_i(u_i-\bar{Y})(v_i-\bar{X})$, $S_{uw}=\sum_{i=1}^{N}P_i(u_i-\bar{Y})(w_i-\bar{Z})$, $S_{vw}=\sum_{i=1}^{N}P_i(v_i-\bar{X})(w_i-\bar{Z})$,

where $\Lambda=\left(\frac{1}{n}\right)$.

3. Existing estimators

For the purpose of comparison, here we discuss some well-known and related estimators along with their properties. We have:

(1) The usual estimator under PPS sampling, given by:

$$\hat{\bar{Y}}_{UPPS} = \bar{u}, \tag{1}$$

$$\mathrm{Var}\left(\hat{\bar{Y}}_{UPPS}\right) = \lambda\bar{Y}^2C_u^2. \tag{2}$$

(2) The ratio estimator under PPS sampling, given by:

$$\hat{\bar{Y}}_{RPPS}= \bar{u}\left(\frac{\bar{X}}{\bar{v}}\right), \tag{3}$$

$$\mathrm{Bias}(\hat{\bar{Y}}_{RPPS}) \cong \bar{Y}\Lambda(C_u^2-\rho_{uv}C_uC_v), \tag{4}$$

$$\mathrm{MSE}(\hat{\bar{Y}}_{RPPS}) \cong \bar{Y}^2\Lambda\ (C_u^2+C_v^2+2\rho_{uv}C_uC_v). \tag{5}$$

(3) The product estimator under PPS sampling, given by:

$$\hat{\bar{Y}}_{PPPS}= \bar{u}\left(\frac{\bar{v}}{\bar{X}}\right), \tag{6}$$

$$\mathrm{Bias}(\hat{\bar{Y}}_{PPPS}) \cong \bar{Y}\Lambda\ \rho_{uv}C_uC_v, \tag{7}$$

$$\text{MSE}(\hat{\bar{Y}}_{PPPS}) \cong \bar{Y}^2 \Lambda \ (C_u^2 + C_v^2 - 2\rho_{uv}C_uC_v). \tag{8}$$

(4) The multivariate ratio estimator under PPS sampling, given by:

$$\widehat{\bar{Y}}_{MRPPS}=\psi_1 \ \bar{u} \ \left(\frac{\bar{X}}{\bar{v}}\right)+\psi_2 \ \bar{u} \ \left(\frac{\bar{W}}{\bar{z}}\right), \tag{9}$$

$$\text{Bias}\left(\hat{\bar{Y}}_{MRPPS}\right) \cong \Lambda \ \{\psi_1(C_v^2 - \rho_{uv}C_uC_v) + \psi_2(C_w^2 - \rho_{uw}C_uC_w)\}, \tag{10}$$

$$\text{MSE}\left(\hat{\bar{Y}}_{MRPPS}\right)=\frac{\Omega_{11} \ \Omega_{22}-\Omega_{12}^2}{\Omega_{11} \ +\Omega_{22}-2\Omega_{12}}, \tag{11}$$

where

$$\Omega_{11}=\Lambda \ \bar{Y}^2(C_u^2 + C_v^2 - 2\rho_{uv}C_uC_v),$$
$$\Omega_{22}=\Lambda \ \bar{Y}^2(C_u^2 + C_w^2 - 2\rho_{uw}C_uC_w),$$
$$\Omega_{12}= \ \Lambda \ \bar{Y}^2(C_u^2 - \rho_{uv}C_uC_v - \rho_{uw}C_uC_w + \rho_{vw}C_vC_w).$$

(5) The regression estimator with variance, given by:

$$\hat{\bar{Y}}_{RegRPPS}= \quad \bar{u}+b(\bar{X} - \bar{v}), \tag{12}$$

$$\text{Var}(\hat{\bar{Y}}_{RegRPPS})=\text{MSE}(\hat{\bar{Y}}_{RegRPPS}) \cong \Lambda\bar{Y}^2C_u^2(1 - \rho_{uv}^2). \tag{13}$$

(6) The ratio and product exponential type estimators, given by:

$$\hat{\bar{Y}}_{BRPPS} = \bar{u} \ \exp\left(\frac{\bar{X}-\bar{v}}{\bar{X}+\bar{v}}\right), \tag{14}$$

and

$$\hat{\bar{Y}}_{BRPPS}= \ \bar{u} \ \exp\left(\frac{\bar{v}-\bar{X}}{\bar{v}+\bar{X}}\right), \tag{15}$$

$$\text{Bias}\left(\hat{\bar{Y}}_{BRPPS}\right) \cong \ \lambda \ \bar{Y}\left(\frac{3}{8} \ C_v^2 - \frac{1}{2}\rho_{uv}C_uC_v\right), \tag{16}$$

$$\text{MSE}\left(\hat{\bar{Y}}_{BRPPS}\right) \cong \ \lambda \ \bar{Y}^2\left(C_u^2 + \frac{1}{4}C_v^2 - \rho_{uv}C_uC_v\right), \tag{17}$$

and

$$\text{Bias}\left(\hat{\bar{Y}}_{BPPPS}\right) \cong \ \lambda \ \bar{Y}\left(\rho_{uv}C_uC_v - \frac{1}{4}C_v^2\right), \tag{18}$$

$$\text{MSE}\left(\hat{\bar{Y}}_{BPPPS}\right) \cong \lambda\ \bar{Y}^2\left[C_u^2 + \frac{1}{4}C_v^2 + \rho_{uv}C_uC_v\right]. \tag{19}$$

4. Recommended work

Population estimation is critically important when making informed decisions in engineering and industrial systems, in which proper evaluation of resource consumption, the performance measure, and the system's behavior directly affect the planning process, the optimization effort, and risk management. Engineering populations are usually very heterogeneous and composed of units that differ in scale, capacity, and in the nature of operations. When this is the case, probability proportional to size (PPS) sampling is often used since it effectively makes use of auxiliary size data to enhance the accuracy of estimation. Nevertheless, the usefulness of the PPS sampling is highly dependent on the strength of the related estimators.

In actual engineering data, there is no such thing as anomaly extreme observations; rather, they are part of the complex system's behavior. Common causes of outliers and heavy-tailed distributions are equipment failures, peak operational loads, environmental stressors, and measurement errors. Theoretically unbiased classical PPS estimators, however, put undue weight on such extreme units because of dissimilar selection probabilities and excessive design weights. This makes population estimates unstable, misleading, and inappropriate in making engineering decisions, especially when sample sizes are medium and when there are a few large units that make up the population. Although these issues are common, no serious efforts to estimate under PPS sampling have been made in survey sampling literature, particularly in the engineering context. Current methods can be based on ad hoc solutions like weight trimming or data winsorization and can affect design-based validity or lead to bias. Furthermore, most powerful statistical methods are formulated on assumptions that cannot be applied to complex survey designs and thus have restricted applicability to practice.

This study was motivated by the necessity to fill in the gap between sound statistical theory and design-based PPS sampling technique in engineering surveys. Estimation procedures resistant to extreme observations have a high demand, and maintaining the probabilistic structure and inferential guarantees of the sampling design is critical. These techniques would give engineers and analysts a more credible estimate of the population, resulting in improved operational planning, system reliability, and resource allocation efficiency.

In addition to this, there is no empirical evidence that can prove the usefulness of robust PPS estimators on actual engineering data. Engineering practitioners need approaches that are theoretically and practically viable and interpretable. The proposed study will provide a methodological contribution to engineering practice by creating and implementing a powerful estimator in the context of the PPS framework, which can be statistically sound and that is immediately applicable to practice. To conclude, the rationale behind the study is the increasing complexity of engineering systems and the drawbacks of existing PPS estimators to address these problems. These problems are significant to tackle in order to promote high-quality inference and population estimation in engineering surveys. The suggested estimator is given by:

$$\hat{\bar{Y}}_{PropPPS} = \bar{u}\left(\frac{\bar{v}}{\bar{X}}\right)^{W_1}\left(\frac{\bar{w}}{\bar{Z}}\right)^{W_2}, \tag{20}$$

The error terms are defined as:

$$e_0=\frac{\bar{u}-\bar{Y}}{\bar{Y}},\ e_1=\frac{\bar{v}-\bar{X}}{\bar{X}},\ e_2=\frac{\bar{w}-\bar{Z}}{\bar{Z}}.$$

Solving (20) up to first-order approximation, we have:

$$\hat{\bar{Y}}_{PropPPS}=\bar{Y}\ (1+e_0)(1+e_1)^{\mathrm{W}_1}(1+e_2)^{\mathrm{W}_2}, \tag{21}$$

$$\hat{\bar{Y}}_{PropPPS}=\bar{Y}\left\{1+e_0+\mathrm{W}_1e_1+\mathrm{W}_2e_2+\mathrm{W}_1e_0e_1+\mathrm{W}_2e_0e_2+\mathrm{W}_1\mathrm{W}_2e_1e_2+\frac{\mathrm{W}_1\ (\mathrm{W}_1-1)}{2}\ e_1^2+\frac{\mathrm{W}_2\ (\mathrm{W}_2-1)}{2}\ e_2^2+\cdots\right\}. \tag{22}$$

To obtain bias of $\hat{\bar{Y}}_{PropPPS}$, we take expectation on both sides of (22):

$$\mathrm{E}\left(\hat{\bar{Y}}_{PropPPS}\right)=\ \bar{Y}\left\{1+E(e_0)+\mathrm{W}_1E(e_1)+\mathrm{W}_2E(e_2)+\mathrm{W}_1E(e_0e_1)+\mathrm{W}_2E(e_0e_2)+\mathrm{W}_1\mathrm{W}_2E(e_1e_2)+\frac{\mathrm{W}_1\ (\mathrm{W}_1-1)}{2}\ E(e_1^2)+\frac{\mathrm{W}_2\ (\mathrm{W}_2-1)}{2}\ E(e_2^2)+\cdots\right\}, \tag{23}$$

$$\mathrm{Bias}\left(\hat{\bar{Y}}_{PropPPS}\right)=\ \bar{Y}\left\{\mathrm{W}_1\rho_{uv}C_uC_v+\mathrm{W}_2\rho_{uw}C_uC_w+\mathrm{W}_1\mathrm{W}_2\rho_{vw}C_vC_w+\frac{\mathrm{W}_1\ (\mathrm{W}_1-1)}{2}\ C_v^2+\frac{\mathrm{W}_2\ (\mathrm{W}_2-1)}{2}\ C_w^2+\cdots\right\}. \tag{24}$$

The MSE of $\hat{\bar{Y}}_{PropPPS}$ is given by:

$$\mathrm{MSE}\left(\hat{\bar{Y}}_{PropPPS}\right)=\bar{Y}^2\mathrm{E}[e_0^2+\mathrm{W}_1^2e_1^2+\mathrm{W}_2^2e_2^2+2\mathrm{W}_1e_0e_1+\ \ 2\mathrm{W}_2e_0e_2+2\mathrm{W}_1\mathrm{W}_2e_1e_2+\cdots], \tag{25}$$

$$\mathrm{MSE}\left(\hat{\bar{Y}}_{PropPPS}\right)=\lambda\bar{Y}^2[C_u^2+\mathrm{W}_1^2C_v^2+\mathrm{W}_2^2C_w^2+2\mathrm{W}_1\rho_{uv}C_uC_v+\ \ 2\mathrm{W}_2\rho_{uw}C_uC_w+2\mathrm{W}_1\mathrm{W}_2\rho_{vw}C_vC_w+\cdots]. \tag{26}$$

Differentiating w.r.t W_1 and W_2 is given by:

$$\mathrm{W}_1=\frac{-C_u\mathrm{K}_{y1.2}}{C_v},\ \mathrm{W}_2=\frac{-C_u\mathrm{K}_{y2.1}}{C_w},$$

where

$$\mathrm{K}_{y1.2}=\frac{\rho_{uv}-\rho_{uw}\cdot\rho_{vw}}{1-\rho_{vw}^2},\quad \mathrm{K}_{y2.1}=\frac{\rho_{vw}-\rho_{uv}\cdot\rho_{vw}}{1-\rho_{vw}^2}$$

$$\mathrm{MSE}\left(\hat{\bar{Y}}_{PropPPS}\right)=\ \lambda\bar{Y}^2C_u^2\ \left\{1-\frac{\rho_{uv}^2+\rho_{uw}^2-2\rho_{uv}\rho_{uw}\rho_{vw}}{1-\rho_{vw}^2}\right\}. \tag{27}$$

5. Numerical study

Here, we discuss some real datasets, which are taken from engineering data. The summary statistics are detailed as follows:

Population-I: [Source: [13]]

Y = Teaching,

X = Placements,

Z = Internship,

Q = Infrastructure.

N=26, n=5, λ=0.2, $\bar{Y}$=2.807692, $\bar{X}$=2.884615, $\bar{Z}$=2.769231, $\bar{u}$=16.93436, $\bar{v}$=16.30256, $\bar{w}$ =16.06564, S_u^2 =1.873669, S_v^2 =2.251578, S_w^2 =1.582643, S_u =1.36882, S_v =1.500526, S_w=1.500526, C_u=0.487525, C_v=0.5201823, C_w=0.5201823, ρ_{uv}=0.7817236, ρ_{uw}=0.6975543, ρ_{vw}=0.6457886

Population-II: [Source: [14]]

Y = Water,

X = coarseagg,

Z = fineagg,

Q = strength.

N=1030, n=50, λ=0.02, $\bar{Y}$= 181.5673, $\bar{X}$= 972.9189, $\bar{Z}$= 773.5805, $\bar{u}$= 5182.585, $\bar{v}$= 27486.63, $\bar{w}$= 22044.73, S_u^2= 13972.79, S_v^2= 367324, S_w^2= 249880.7, S_u= 118.2066, S_v= 606.0726, S_w= 606.0726, C_u= 0.6510345, C_v= 0.6229426, C_w= 0.6229426, ρ_{uv}= 0.9648626, ρ_{uw}= 0.7872318, ρ_{vw}= 0.798734

Population-III: [Source: [15]]

Y =Age,

X = BMI,

Z = Antral follicle count,

Q = Testosterone_Level(ng/dL).

N=1000, n=50, λ=0.02, $\bar{Y}$= 31.771, $\bar{X}$= 17.469, $\bar{Z}$= 26.387, $\bar{u}$= 772.4925, $\bar{v}$= 425.206, $\bar{w}$= 638.1502, S_u^2= 310.5192, S_v^2= 127.4147, S_w^2= 173.6103, S_u= 17.62155, S_v= 11.28781, S_w= 11.28781, C_u= 0.5546427, C_v= 0.6461625, C_w= 0.6461625, ρ_{uv} 0.6284066, ρ_{uw}= 0.8934947, ρ_{vw}= 0.7901385.

6. Discussion

Table 1 provides the mean squared error of all considered estimators, and Table 2 provides the PREs of all considered estimators. Figures 1–3 show the numerical results with the help of graphs. The findings of this research indicate that the issue of robustness through PPS-based population estimation benefits the reliability of inference significantly when statistical observations are known, which is prevalent in engineering data. The results illustrate that such behavior may result in unstable estimates and inflated variance, which negatively affect the usefulness of classical PPS estimators in a practical engineering application.

The suggested robust estimator is useful to confine the effect of the extreme sample units without ignoring the PPS sampling framework. In the actual engineering data application, the robust estimator has a lower cost (mean squared error) compared to the classical PPS estimator at

contamination. Notably, the efficiency loss at uncontaminated or almost normal conditions was small, which implies that the robustness efficiency is balanced. This is a property that is especially desirable in engineering practice, where the actual amount of contamination is usually unknown, and data conditions might be different in working conditions.

Table 1. Mean square error (MSE) using engineering data.

Estimators	Population-I	Population-II	Population-III
$\widehat{\bar{Y}}_{Upps}$	0.3747337	279.4558	6.210383
$\widehat{\bar{Y}}_{Rpps}$	0.1762308	19.31162	5.546151
$\widehat{\bar{Y}}_{Ppps}$	1.426475	1051.319	23.73259
$\widehat{\bar{Y}}_{MRPPS}$	0.1267084	32.67255	3.794257
$\widehat{\bar{Y}}_{RegPPS}$	0.145737	19.29368	3.757935
$\widehat{\bar{Y}}_{BRPPS}$	0.1688275	85.41891	3.771021
$\widehat{\bar{Y}}_{BPPPS}$	0.7939494	601.4224	12.86424
$\widehat{\bar{Y}}_{PropPPS}$	**0.1218607**	**19.08191**	**1.152941**

However, in spite of these strong points, there are some limitations of the study that should be mentioned. The performance of the estimators can be affected by the tuning parameters used; despite the standard guidelines used, the optimum parameter selection in the PPS settings is still under research. The study also concentrated on PPS sampling, without replacement; hence, a more significant development of the study is to extend the sampling methodology to a multi-stage, cluster, or adaptive sampling method typically applied in large-scale engineering surveys. Further research should look into the application of the suggested robust estimator alongside calibration and model-assisted models and tools and its behavior in the case of alternative contamination and nonresponse. Research into computational efficiency when dealing with large engineering populations and programming implementations would also increase practical adoption. In general, the results of this study can be used to emphasize the significance of robustness in PPS-based population estimation and show that effective design-based inference is possible even when engineering information is both complex and contaminated. The suggested powerful estimator is a useful methodological contribution to both survey statisticians and engineering practitioners.

Table 2. Mean square error (MSE) using engineering data.

Estimators	Population-I	Population-II	Population-III
$\widehat{\bar{Y}}_{Upps}$	100	100	100
$\widehat{\bar{Y}}_{Rpps}$	212.638	1447.087	111.9765
$\widehat{\bar{Y}}_{Ppps}$	26.26992	26.58146	26.16817
$\widehat{\bar{Y}}_{MRPPS}$	295.745	855.323	163.6785
$\widehat{\bar{Y}}_{RegPPS}$	257.1301	1448.432	165.2605
$\widehat{\bar{Y}}_{BRPPS}$	221.9625	327.1592	164.6871
$\widehat{\bar{Y}}_{BPPPS}$	47.19869	46.46582	48.27634
$\widehat{\bar{Y}}_{PropPPS}$	**307.5099**	**1464.507**	**538.6556**

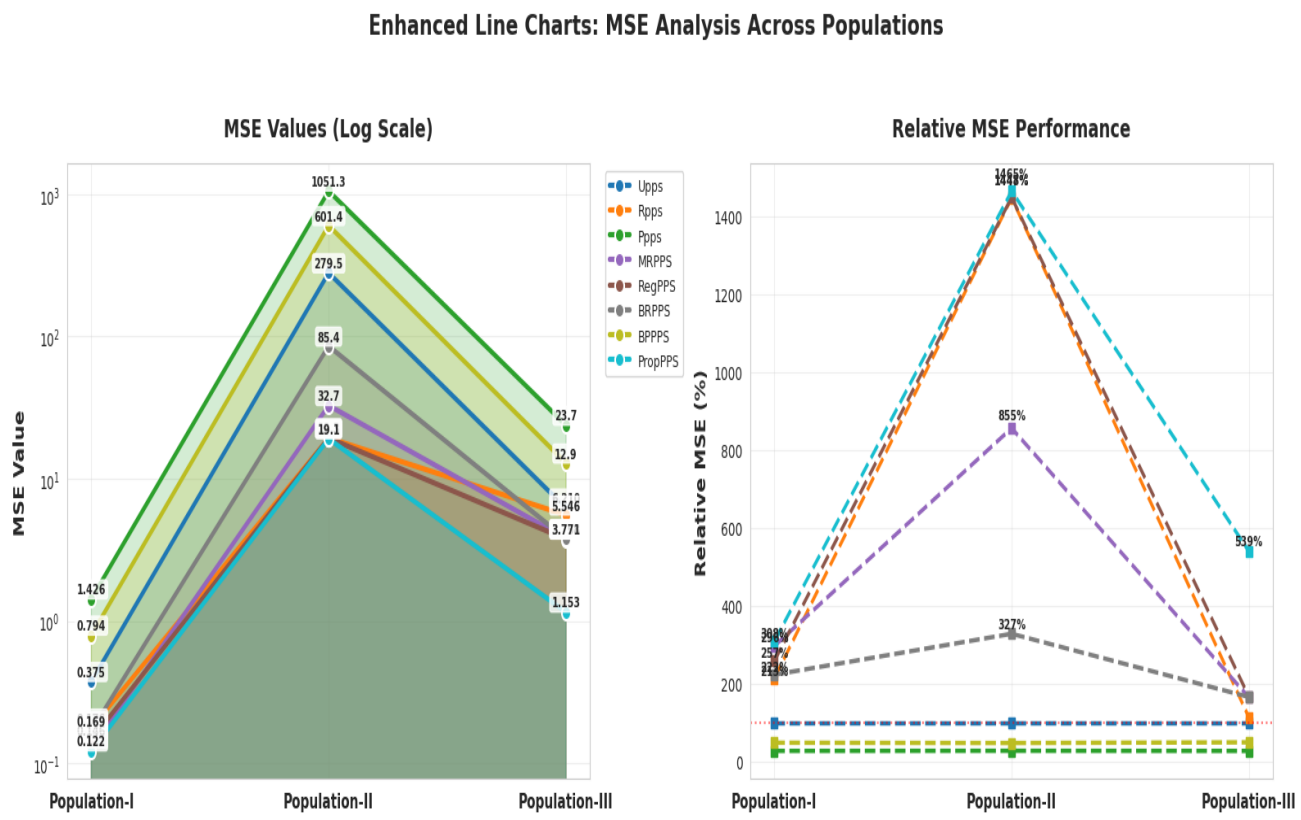

Figure 1. MSEs of all considered estimators using enhanced line charts.

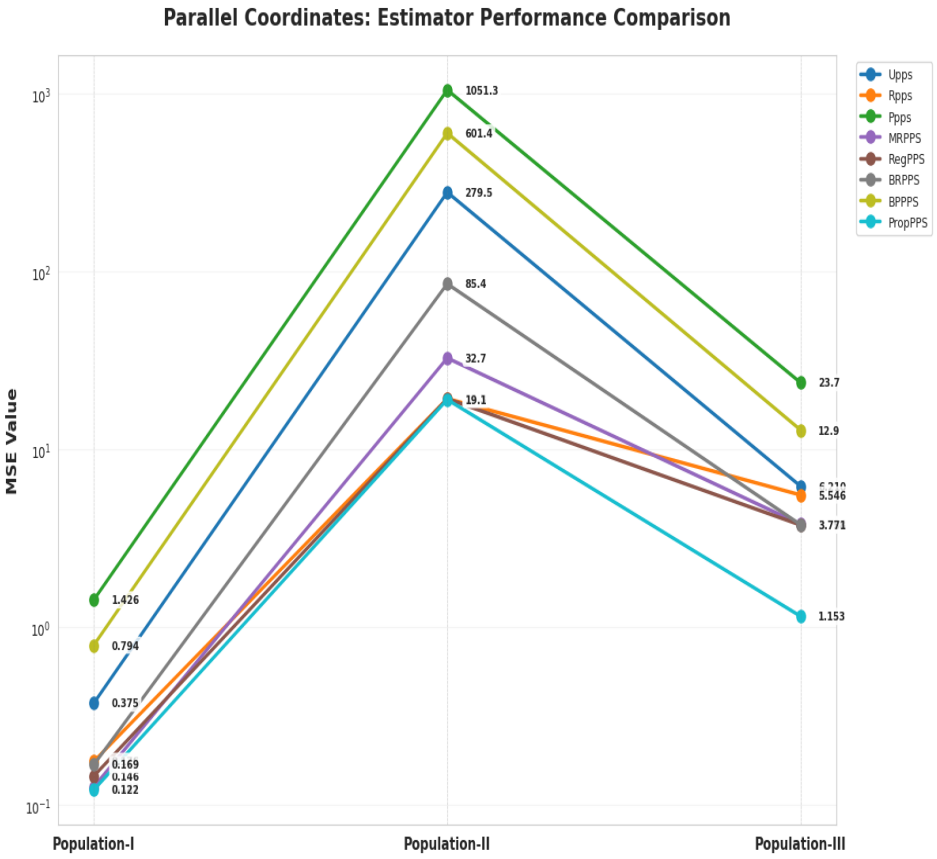

Figure 2. MSEs of all considered estimators using parallel coordinates.

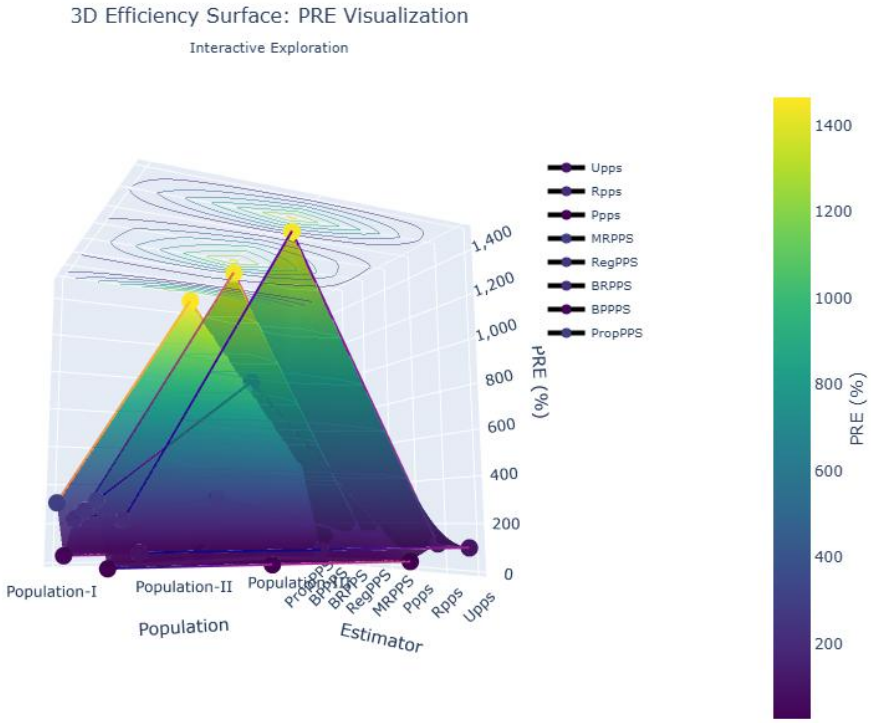

Figure 3. PREs of all estimators using 3D surface.

7. Conclusions

In this article, we proposed an improved estimator for estimating the population mean under PPS sampling using auxiliary information. To assess the efficiency of the proposed estimator, we used engineering datasets. The proposed robust estimator can successfully regulate the effect of extreme sample units without modifying the basic properties of PPS sampling. Theoretical studies had identified important design properties, such as consistency and valid variance estimation, where inference is statistically sound. The estimator was proven to be superior to classical PPS estimators in terms of bias reduction and mean squared error in a wide range of contamination cases, although the estimator is equally efficient when the data is clean. The proposed method has practical significance, supported by the engineering data application. The robust estimator also generated more stable estimations and confidence intervals than the classical PPS estimator, which resulted in a more correct and practical conclusion for engineering decision-makers. These have a direct implication on operation planning, energy management, and system optimization, where inaccurate population estimates may lead to inefficient or expensive decisions.

The presented methodology was also found to be practical, confirmed by the application to real engineering data. The powerful estimator generated less biased and more sensible population estimates, less variance inflation, and narrower confidence intervals. These improvements are especially significant to the engineering decision-making process, where poor population estimates may cause resource allocation to be conducted inefficiently, designing the system in a sub-optimal fashion, and creating more operational risk. In general, the results indicate that robustness should be introduced into the PPS-based estimation process to overcome the challenges of contemporary engineering data.

8. Future directions

- A number of future research opportunities can be identified through this research. First, the suggested robust estimation framework could be applicable to large-scale engineering surveys by extending the framework to more complicated sampling designs, e.g., multi-stage, cluster, and stratified PPS sampling. Second, more effective estimator performance in a variety of engineering settings can be achieved by the development of data-driven or adaptive procedures for choosing robust tuning parameters.
- Future research could also study the combination of the concept of robustness with calibration as well as model-assisted estimation, and how the usage of auxiliary information could be improved by protection against outliers. Another valuable research direction is how to solve the problem of nonresponse and measurement error in the robust PPS framework, especially in the case of engineering systems that are incomplete or noisy.
- Practically speaking, it would be simpler to establish computationally effective algorithms and more user-friendly software implementations in order to allow more engineering practitioners to adopt strong PPS estimators. Lastly, the suggested methodology should be extended to a wider variety of engineering fields and include energy systems, infrastructure monitoring, reliability engineering, and environmental engineering, offering additional empirical evidence and proving its flexibility.

- Finally, the present research provides a solid and substantially theoretically justified method of population estimation in PPS sampling that acts as a valuable instrument to enhance the quality of survey inference in engineering, with promising future methodological developments.

Author contributions

L. S. Diab: Writing – original draft; Norah D. Alshahrani: Writing – original draft; Majdah Mohammed Badr: Writing – review, Conceptualization, Formal analysis; Sohaib Ahmad: Writing – review, Conceptualization, Formal analysis. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Funding statement

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2601).

Data availability

All data and source are available within the manuscript.

Conflict of interest

The authors declare no conflict of interest

References

1. A. R. El-Saeed, S Ahmad, B. Aloraini, Robust estimator for estimation of population mean under PPS sampling: Application to radiation data, *J. Radiat. Res. Appl. Sci.*, **18** (2025), 101384. https://doi.org/10.1016/j.jrras.2025.101384
2. S. Khan, M. Farooq, Ahmad, S. Khan, Improved Estimator for the Estimation of Population Mean Using a Predictive Approach Under PPS Sampling, *VFAST Trans. Math.*, **12** (2024), 01–16. https://doi.org/10.21015/vtm.v12i2.1942
3. M. Azeem, S. Iftikhar, M. Ijaz, N. Salahuddin, M. Ilyas, An improved estimator of population mean under PPS sampling with application to radiation data sets, *J. Radiat. Res. Appl. Sci.*, **18** (2025), 101543. https://doi.org/10.1016/j.jrras.2025.101543

4. J. Wang, S. Ahmad, M. Arslan, S. A. Lone, A. H. Abd Ellah, M. A. Aldahlan, et al., Estimation of finite population mean using double sampling under probability proportional to size sampling in the presence of extreme values, *Heliyon*, **9** (2023), e21418. https://doi.org/10.1016/j.heliyon.2023.e21418
5. S. Ahmad, J. Shabbir, E. Zahid, M. Aamir, Improved family of estimators for the population mean using supplementary variables under PPS sampling, *Sci. Prog.*, **106** (2023), 00368504231180085. https://doi.org/10.1177/00368504231180085
6. H. Zheng, R. J. Little, Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples, *J. Off. Stat.*, **19** (2003), 99.
7. R. R. Sinha, B. Khanna, Estimation of ratio and product of two population means under pps sampling with and without measurement error, In: *Statistical Modeling and Applications on Real-Time Problems*, CRC Press, 2024, 34–59. https://doi.org/10.1201/9781003481263-3
8. M. Hussein Mohamud, F. A. Mohamud, Estimation of the mean using robust regression and probability proportional to size sampling, *Statistical Theory Relat. Fields*, **9** (2025), 213–222. https://doi.org/10.1080/24754269.2025.2516339
9. R. Latpate, J. Kshirsagar, V. Kumar Gupta, G. Chandra, Probability proportional to size sampling, In: *Advanced sampling methods*, Singapore: Springer, 2021, 85–98. https://doi.org/10.1007/978-981-16-0622-9_7
10. R. R. Sinha, B. Khanna, Estimation of population mean under probability proportional to size sampling with and without measurement errors, *Concurrency Comput.: Pract. Exper.*, **34** (2022), e7023. https://doi.org/10.1002/cpe.7023
11. S. Ahmad, J. Shabbir, E. Zahid, M. Aamir, M. Alqawba, New generalized class of estimators for estimation of finite population mean based on probability proportional to size sampling using two auxiliary variables: A simulation study, *Sci. Prog.*, **106** (2023), 00368504231208537. https://doi.org/10.1177/00368504231208537
12. S. Shah, E. Mahmoudi, H. Iftikhar, P. C. Rodrigues, R. I. Gonzales Medina, J. L. López-Gonzales, A Novel Family of CDF Estimators Under PPS Sampling: Computational, Theoretical, and Applied Perspectives, *Axioms*, **14** (2025), 796. https://doi.org/10.3390/axioms14110796
13. A. Verma, Clustering-Engineering College Data, 2020. Available from: https://www.kaggle.com/datasets/ankitverma2010/clustering-engineering-college-data?select=Engg_College_Data.csv.
14. V. Shanawad, Civil Engineering: Cement Manufacturing Dataset, 2021. Available from: https://www.kaggle.com/datasets/vinayakshanawad/cement-manufacturing-concrete-dataset.
15. S. Dalvi, A Dataset for Exploratory Data Analysis, Feature Engineering, 2025. Available from: https://www.kaggle.com/datasets/samikshadalvi/pcos-diagnosis-dataset.