Sampling with Probability Proportional to Aggregate Size in Heterogeneous Populations: A Study of Design and Efficiency

Daniel David M. Pamplona
University of the Philippines Visayas

Sampling with probability proportional to aggregate size (PPAS) is compared with traditional design-unbiased sampling methods under different simulated population scenarios in the estimation of the population total. The study considered both accuracy and precision of the estimates in the comparison. Heterogeneous populations were simulated by exploring varying behaviors of an auxiliary variable and its relationship with the target variable. Results show that the optimality of estimates using PPAS sampling improve as the association between the target variable and auxiliary variable strengthens. Furthermore, PPAS sampling estimates are more stable under large variability in the population.

Keywords: Probability Proportional to Aggregate Size Sampling, Nonparametric Bootstrap, Simple Random Sampling, Probability Proportional to Size Systematic Sampling

1. Introduction

Developments in sampling theory were brought by the emphasis on the use of auxiliary information in improving precision of estimates (Homa, Maurya, and Singh, 2013). Auxiliary information can be quantitative or categorical, examples are variables available in the register such as age, sex, and marital status. One approach that uses auxiliary variable is the probability proportional to size (PPS) sampling, a design-unbiased sampling method, where the inclusion probabilities $\pi_1, \ldots, \pi_N$ of the design is proportional to known, positive values $x_1, \ldots, x_N$ of an auxiliary variable. The PPS estimator is known to have small variance if x is more or less proportional to y, the target variable. However, PPS sampling was sometimes found to be difficult to carry out (Särndal, Swetson, and Wretman, 1992). It is assumed that the auxiliary variables covary with the target variable,

and thus carry information about the target variable. Such relationship is used advantageously in model-based and nonparametric estimation.

Another approach is the probability proportional to aggregate size (PPAS) sampling proposed by Midzuno (1952) using the unbiased ratio estimates of Lahiri (1951). In PPAS, the first unit is selected using PPS and the remaining $n - 1$ units are selected using simple random sampling without replacement (SRSWOR). Similar to PPS sampling, the variance is expected to be small when the auxiliary variable is proportional to the target variable. However, variance estimation for PPAS sampling method have been difficult in some cases. For large samples, Chauvet (2018) showed that the variance estimators for SRSWOR are also consistent for PPAS sampling.

Gauran and Poblador (2012) used sampling with PPAS to estimate total production area of top cereals and root crops across Philippine regions. The problem of negative estimated variance was encountered, so the use of the nonparametric bootstrap to estimate the standard error of the estimate of the total was adopted in the study. The performance of estimates under PPAS sampling was compared to PPS sampling and SRSWOR in terms of bias and precision of estimates. However, the study only used observed data, and little knowledge was gained in understanding the flexibility of the sampling method under varying population characteristics and model assumptions. To do this, a simulation study of heterogeneous populations needs to be carried out.

This study aims to identify the population characteristics where optimality of estimates is achieved using PPAS sampling as compared to SRSWOR and probability proportional to size systemic (PPSS) sampling. Data sets were simulated to explore the different behavior of the population of interest in relation to the auxiliary variable. Comparison of estimates were made in terms of bias and precision. Variance estimation of estimates in PPAS sampling is done with nonparametric bootstrap to address the similar issue of negative estimated variance.

2. Sampling Designs

Consider a finite population U of size N, with a variable of interest Y where y_i is the value for the unit $i \in U$. Suppose it is of interest to estimate the total $t_y = \sum_{i \in U} y_i$, and for all $i \in U$ a known measure x_i is available in the population. If Y relates to X in some functional manner, X is considered as auxiliary variable or subsidiary variable, with $t_x = \sum_{i \in U} x_i$.

Let $p_i > 0$ be some probability for unit i, with $\sum_{i \in U} p_i$. If the probabilities are chosen proportional to x_i, then $p_i = x_i / t_x$. A sample S of size n is selected according to some sampling design with probability of inclusion π_i for unit i. The Horvitz-Thompson (HT) estimator of the total is $\hat{t}_y = \Sigma_{i \in S} (y_1 / \pi_i)$.

2.1. Probability proportional to aggregate size sampling

Lahiri (1951) and Midzuno (1952) used information from aggregated size measure in developing unbiased ratio estimates. In the Lahiri-Midzuno scheme a sample of size n is chosen by selecting the first unit i with probability proportional to size measure x. The other $n-1$ units are selected using simple random sampling without replacement. It is known that the first order inclusion probability π_i for unit i and the probability of selecting a unique sample s of size n are given by

$$\pi_i = \frac{n-1}{N-1} + \frac{N-n}{N-1}\frac{x_i}{\sum_{j=1}^{N} x_j'} \tag{1}$$

$$P(s) = \frac{1}{N' t_x}\sum_{i=1}^{n} x_i, \tag{2}$$

where $N' = \binom{N-1}{n-1}$.

The unbiased estimator of the population total is

$$\hat{t}_{y,PPAS} = t_x \frac{\overline{y}}{\overline{x}} = \sum_{i=1}^{N} x_j \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}, \tag{3}$$

and the variance of the estimator can then be derived

$$\begin{aligned} \upsilon_{PPAS}(\hat{t}_y) &= E(\hat{t}_y^2) - t_y^2 \\ &= \sum_{s\in S} t_x^2 \frac{\overline{y}^2}{\overline{x}^2} P(s) - t_y^2 \\ &= t_x^2 \sum_{s\in S} \frac{(\Sigma y_i)^2}{(\Sigma x_i)^2} \frac{\Sigma x_i}{N' t_x^2} - t_y^2 \\ &= \frac{t_x}{N'} \sum_{s\in S} \frac{(\Sigma y_i)^2}{\Sigma x_i} - t_y^2. \end{aligned} \tag{4}$$

If $y = kx$, for some constant k, then

$$\begin{aligned} \upsilon_{PPAS}(\hat{t}_y) &= \frac{t_x}{N'} \sum_{s\in S} k^2 \frac{(\Sigma x_i)^2}{\Sigma x_i} - t_y^2 \\ &= \frac{t_x k^2}{N'} \sum_{j=1}^{N} x_j N' - t_y^2 \\ &= k^2 t_x^2 - t_y^2 \\ &= 0. \end{aligned} \tag{5}$$

The above result shows that efficiency of estimates relies on the proportional dependence of Y with the auxiliary variable X. It may be examined that optimal estimates can still be attained even if the relationship is not deterministic, i.e. $y = kx + \varepsilon$, where ε is a random error. It is also possible that the relationship is nonlinear in nature, thus the relationship is generalized by $y = f(x) + \varepsilon$, where $f(\cdot)$ takes possibly a non-linear form.

The estimator of the variance of the estimate of the total can be generalized to

$$\upsilon_{PPAS}(\hat{t}_y) = \left[\sum_{i=1}^{n} \sum_{i<j}^{n} \frac{(y_i + y_j)^2}{x_i + x_j} - \left(\sum_{i=1}^{n} y_i \right)^2 \right] \frac{t_x}{N-1} \tag{6}$$

this, however, might result to negative values, which poses a problem in variance estimation.

Gauran and Poblador (2012) explored the use of the nonparametric bootstrap to approximate the standard error of the estimate of total production area of cereals and root crops in the Philippines using PPAS sampling. The proposed procedure fixed the problem of negative variance. It was also observed that sampling method PPAS produced more precise estimates for small sample sizes when compared to probability proportional to size (PPS) sampling and simple random sampling without replacement (SRSWOR). The PPAS sampling method was also found to produce less biased estimates than SRSWOR and PPS sampling.

2.2. *Simple Random Sampling Without Replacement (SRSWOR)*

With SRSWOR, each unit in the population has equal probability of being selected as sample. This is indicated by the inclusion probability of any unit i as $\pi_i = n/N$. The design also assigns equal probability to each possible sample of size n. That is, selection probability of a sample j is $P(S_j) = 1/({}_N C_n)$. The unbiased estimator of the population total and the variance of its estimate, as shown by Lohr (2010), is given by:

$$\hat{t}_{SRS} = \sum\nolimits_{i=1}^{n} \frac{N}{n} y_i, \tag{7}$$

$$\upsilon\left(\hat{t}_{SRS}\right) = N^2 \left(\frac{N-n}{Nn} \right) S^2, \tag{8}$$

$$\text{where } s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y})^2; \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

2.3. Probability Proportional to Size: Systematic Sampling (PPSS)

In systematic sampling with unequal probabilities, an auxiliary variable denoted by X is used to reformat the selection probabilities. A unit whose size measure is $\geq X / n$ is discarded from the population and included as certainty unit. The probability that unit i is selected as sample is given by $\pi_i = n(X_i / X)$, and the joint inclusion probability of unit i and j is $\pi_{ij} = n(m_{ij} / X)$, where m_{ij} is the number of random numbers that select unit i and j simultaneously. The estimate of the total and the estimate of the variance of this estimate is generated by the Horvitz-Thompson estimator given by

$$\hat{t}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}, \tag{9}$$

$$\upsilon(\hat{t}_{HT}) = \sum_{i=1}^{n} \sum_{j>i} \left(\frac{\pi_i \pi_j - \pi_{ij}}{\pi_{ij}} \right) \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 . \tag{10}$$

3. Methodology

3.1. Simulation of different populations

To evaluate the performance of PPAS estimates, a simulation study was conducted. Each scenario postulates a model

$$y = f(x) + k\varepsilon, \ \ \varepsilon \sim N(0,1) \tag{11}$$

For this equation, the following characteristics were made to vary: functional form of $f(x)$, variance of X, multiplier (k) on the error term, and sampling rate. These variations in the model aim to capture the different patterns of association between the target and auxiliary variable.

Let $f(x)$ be either linear or nonlinear function of x. The different forms of $f(x)$ used in the simulation study are given by:

$$\text{Linear: } y = 3.2x + k\varepsilon \tag{12}$$

$$\textit{Quadratic: } y = 8.5x - 0.1x^2 + k\varepsilon \tag{13}$$

$$\textit{Exponential: } y = 12 \exp(x/20) + k\varepsilon \tag{14}$$

These different functional forms result to different population distributions. For instance, $f(x)$ that follows exponential form tend to create skewed populations as compared to linear and quadratic relationships.

The coefficients of the simulated model were chosen to generate roughly similar population total of the response Y under across all scenarios. The auxiliary

variable X was randomly generated from a normal distribution with mean 50 and variances 5, 40, and 200. Error terms were simulated from the standard normal distribution. The strength of relationship was controlled by the error multiplier (k) set to 1, 10, 30, in the functional model. For $k > 1$, the errors are magnified, and the model fit suffers from large errors. A similar approach was used by Kwong (2011) in simulating the different model fit for linear and nonlinear relationships between the target and auxiliary variable in nonparametric model-based estimation of population total.

The scenarios adopted in the simulation study aim to create heterogenous population characteristics where estimates of PPAS sampling are compared to SRSWOR and PPSS sampling estimates of the population total. The relative variability of the population is measured by the coefficient of variation (CV) of the target variable. The higher the CV value the greater the heterogeneity of the population. Populations with high variability often requires a larger sample which makes the procedure more complex and costly. Conversely, for populations with small CV, a smaller sample is sufficient.

3.2. Estimation of the variance of the estimated total when sampling with PPAS using the bootstrap

The bootstrap was developed by Efron (1992) inspired by earlier work on the Jackknife. Bootstrapping is an estimation procedure where inference about a population from sample data can be modelled by *resampling* the sample data and performing inference using the resample data. The new population is the original sample and the sampling distribution of the estimate is determined by the empirical distribution of the resample estimates. In surveys involving complex designs, the bootstrap is a common approach to variance estimation. The basic approach is as follows:

1. Observe a sample S with size n.
2. Compute $\hat{\theta}_s$ which estimates some model parameter θ.
3. For $k = 1, 2, \ldots, K$, generate a bootstrap sample S_k by sampling with replacement from the original observed sample.
4. Compute $\hat{\theta}_k$ for each bootstrap sample S_k, in the same way the original estimate $\hat{\theta}_s$ was calculated.
5. The parameter estimate and its corresponding variance is given by

$$\bar{\hat{\theta}} = \sum_{k=1}^{K} \hat{\theta} \frac{1}{k}, \tag{15}$$

$$V(\hat{\theta}_k) = \frac{1}{K-1}\sum_{k=1}^{k}\left(\hat{\theta}_k - \bar{\hat{\theta}}\right)^2. \quad (16)$$

The standard error of the estimate of the total from PPAS sampling was estimated by the bootstrap method with 200 replicates/resamples.

3.3. Evaluation of sampling for the population total with PPAS

Two evaluation measures were considered in this study: the mean absolute percentage difference for the accuracy of the estimator under a sampling method and the standard errors of the estimates for the precision of the estimates under a sampling method.

For each combination of model restriction and sampling rate (1%, 5%, 10%), the estimate of the total and its standard error estimate were generated using SRSWOR and PPAS sampling. The PPSS sampling estimates were only generated for the 1% sampling rate. This is because size measure of units in PPSS sampling are more distinct when considering small samples only. For the measurement of the bias/accuracy of an estimate of the total, $\hat{T}_{est}$, from a sampling method given the true total, T, the absolute percentage difference was used, APD_{est} (%),defined as:

$$APD_{est} = \left|\frac{\hat{T}_{est} - T}{T}\right| \cdot 100 \quad (17)$$

The more accurate the estimator/better sampling method is that with the lower mean APD_{est}.

The standard error for PPAS sampling was estimated by the bootstrap method. For SRSWOR sampling and PPS sampling, the standard errors of the estimate of the total were calculated using their respective theoretical estimators. A sampling method with lower standard error indicates a more precise estimate of the population total.

4. Results and Discussion

4.1 Variation in the target variable (Y) and its association with the auxiliary variable (X) under the simulated populations

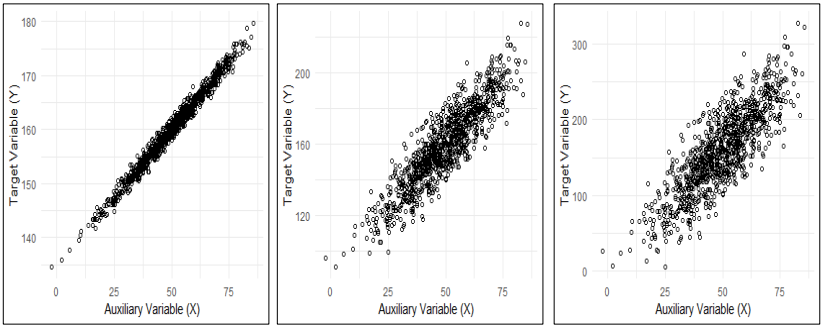
Heterogeneous populations in this study were simulated by varying population characteristics such as: the variability of the target variable (*Y*) and its strength and form of association with the auxiliary variable (*X*). For each scenario, the association between *Y* and *X* are examined. Pearson r correlation coefficient was used for linear association and Spearman rho coefficient for nonlinear monotonic association.

Tables 1 shows the CV of the population target variable resulting from each simulation scenario. As expected, the CV of increase as variation in X increase and as (k) gets larger. Also, CVs for simulated response under a quadratic form $f(x)$ are relatively lower than linear and exponential form, possibly due to the parabolic form of the function that restricts Y values to be scattered closely around the vertex (See Figure 2). Furthermore, at fixed variance of X, the association between X and Y weakens as the error multiplier (k) becomes larger.

Table 1. Variability of Target Population (Y) and Correlation Across Varying Var(X) and k

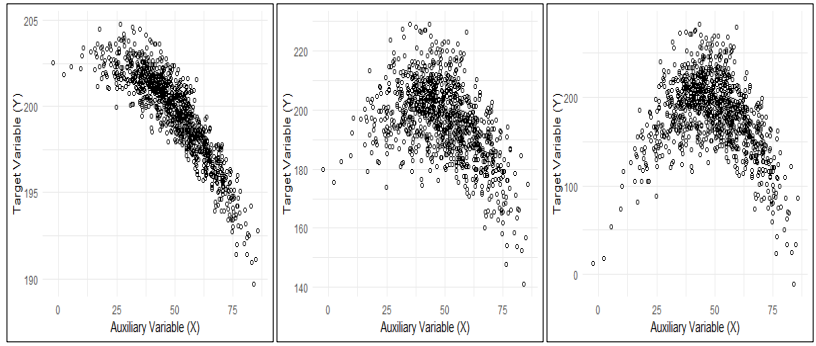
Functional Form of Target Variable, Y=f(x)	Variance of Auxiliary Variable, V(X)	5			40			200		
Linear: y = 3.2x + kε	Model Fit / Error Multiplier, k	1	10	30	1	10	30	1	10	30
	CV(%) of Y	0.99	0.57	0.20	0.99	0.89	0.55	0.99	0.98	0.83
Quadratic: y = 8.5x – 0.1x2 + kε	Pearson r (between X and Y)	0.05	0.08	0.19	0.13	0.14	0.22	0.29	0.29	0.34
	CV(%) of Y	0.02	0.06	0.17	0.06	0.09	0.18	0.21	0.22	0.29
Exponential y = 12 exp (x/20) + kε	CV (%) of Y	0.99	0.85	0.46	0.98	0.96	0.82	0.91	0.91	0.89
	Spearman rho, (between X and Y)	0.11	0.13	0.23	0.32	0.33	0.37	0.74	0.74	0.75

Figures 1, 2, and 3 illustrate the scatter plots of the simulated relationship between and under linear, quadratic, and exponential form, respectively. The plots are organized from left to right according to degree of variation of the auxiliary variable and the error multiplier . As a result, scatter and scope of points are wider in plots (c) than (a). These forms were considered to capture different population characteristics that can evaluate the performance of the designs considered in the study.



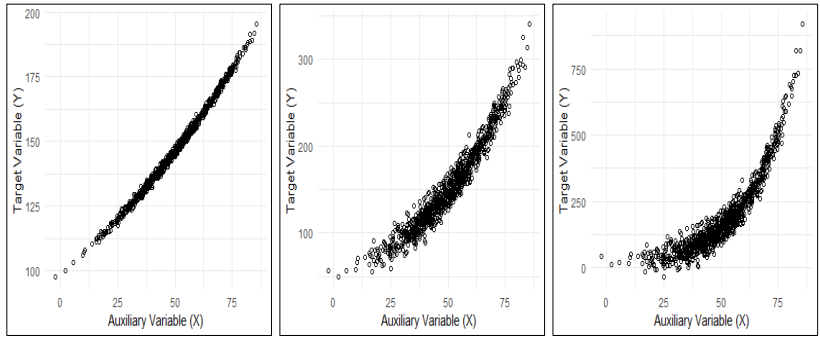
(a) (b) (c)

Figure 1. Scatter Plots of the Target Variable and the Auxiliary Variable for (a) Var(X)=5, k = 1; (b) Var(X)=40, k=10; (c) Var(X) =200, k=30



(a) (b) (c)

Figure 2. Scatter Plots of the Target Variable and the Auxiliary Variable for (a) Var(X)=5, k = 1; (b) Var(X)=40, k=10; (c) Var(X) =200, k=30



(a) (b) (c)

Figure 3. Scatter Plots of Target Variable and the Auxiliary Variable for (a) Var(X)=5, k = 1; (b) Var(X)=40, k=10; (c) Var(X) =200, k=30

4.2 Estimates of the total using the three sampling procedures

The estimates of population total using SRSWOR, PPAS sampling, and PPSS sampling at three sampling rates are presented in Table 2. At 1% sampling rate, PPAS produced estimates closer to the actual population total as compared to SRSWOR and PPSS sampling. But at 5% and 10% sampling rates, SRSWOR and PPAS sampling methods tend to produce similar estimates, and smaller bias. This is consistent with results of Chauvet (2018) that the two procedures are asymptotically equivalent in large samples. There is larger bias in PPSS estimates as compared to the other two procedures. This is expected, since PPSS estimates are known to perform better in small populations sizes with a heterogeneous target variable.

Table 2. Comparison of PPAS Sampling Estimates of the Population Total with those of Two Sampling Methods by Functional Form of the Target Variable (*Y*), by Variance of the Auxiliary Variable (*X*), and by Model Fit (*k*) Using Three Sampling Rates

f(x)	Vari-ance of X	Model fit (k)	Population Total	Sampling Rate						
				1%			5%		10%	
				SRS	PPAS	PPSS	SRS	PPAS	SRS	PPAS
Linear	5	1	147837	161271	147485	170657	161346	168249	161412	158816
		10	147982	164068	169151	174972	161309	162486	159074	160394
		30	148304	153343	160557	168162	158253	150365	158927	168373
	40	1	148882	153526	159607	158304	162146	158052	163256	158427
		10	149028	154631	156024	161727	164834	162171	160041	157863
		30	149350	136835	156209	148391	155808	168695	163433	152567
	200	1	155714	142184	158162	161583	157437	158398	162327	158133
		10	155859	177099	155292	159771	154314	158111	152807	160251
		30	156181	166743	160357	161727	167242	152383	162723	154797
Quadratic	5	1	174624	175970	157368	180345	174829	186288	174394	173227
		10	174769	177790	188184	175275	174042	177685	174950	175865
		30	175091	178207	176178	203278	173679	162476	172585	186246
	40	1	171221	171139	169052	169023	169094	169530	169937	169878
		10	171366	165163	167448	191130	170827	177833	172137	168607
		30	171689	170342	166410	157148	167599	189831	171092	161803
	200	1	155053	154876	170168	186876	151768	164432	157651	140461
		10	155198	163095	175128	166911	153240	162775	164020	156763
		30	155521	167908	159338	179160	152563	135691	155227	143010
Exponential	5	1	146532	148739	139062	155782	141872	152230	149967	146070
		10	146677	135740	152355	137258	150095	148927	147195	146624
		30	147000	144152	146586	131893	148739	139538	144002	152438
	40	1	152032	135171	156837	153242	140852	150428	155424	151311
		10	152177	176755	149715	160925	171407	152088	149220	152405
		30	152499	142599	148782	163247	164213	152889	164784	148310
	200	1	182514	111199	164746	209162	204261	170949	165332	203395
		10	182659	158605	142547	179828	193319	169637	169389	181517
		30	182981	140249	175103	233006	201185	198611	176594	187584

4.3 On the accuracy of the estimates of the total when sampling with PPAS compared to two sampling methods

The absolute percent difference provides a standard measure to compare observed bias of estimates. Consequently, the average of these values can be taken for each simulation setting given by the mean APD_{est}.

Table 3 shows these values across model restrictions by sampling design and sampling rate. In examining the values across the different forms of $f(x)$, SRSWOR appears to be affected by the different functions used. It has highest bias under exponential form, and it has lowest bias under quadratic form. However, PPAS estimates do not seem to be affected by the different forms of $f(x)$. In contrast to SRSWOR estimates, sampling with PPAS gave lower bias under exponential form, especially at higher sampling rates. In the same manner as in Table 2, it can be noted that as sampling rate goes higher, the estimates of PPAS sampling and SRSWOR also draw closer, and bias between the two procedures is almost similar. PPSS sampling estimates have higher bias as compared to the other two procedures, and it also appears to be unaffected by the different forms of $f(x)$.

The mean APD_{est} across different variance of auxiliary variable are also shown in Table 3. At variance equal to 5, SRSWOR have lowest bias compared to the other two designs at each sampling rate. When the variance of X is set to 40, PPAS sampling estimates show better results than the other designs across the different sampling rates. SRSWOR and PPAS sampling estimates show roughly the same bias at 10% sampling rate, once again showing the asymptotic property of sampling with PPAS. At 1% sampling rate, SRSWOR bias is larger than other designs under high variance of the auxiliary variable. PPSS sampling method also have better estimates under moderate variance of auxiliary variable.

Table 3 also provide the average absolute percent difference at different error multiplier (k) used in simulating the models. At 5% and 10% sampling rates, PPAS sampling estimates have higher bias under poor model fit (k = 30). Similarly, sampling with PPSS also have estimates with higher bias at high model errors. SRSWOR estimates, on the other hand, did not exhibit any changes in bias across the various model fit at each sampling rate. At lower sampling rates, SRSWOR still have the highest bias, but as sampling rates increase, estimates of SRSWOR and PPAS sampling become comparable.

Table 3. Comparison of the Mean Absolute Percentage Differences of PPAS Sampling Estimates of the Population Total with those of Two Sampling Methods Under Different Combinations of Functional Form of the Target Variable (Y), Variance of the Auxiliary Variable (X), and Model Fit (k) Using Three Sampling Rates

Characteristic	Values	Sampling Rate						
		1%			5%		10%	
		SRS	PPAS	PPSS	SRS	PPAS	SRS	PPAS
Functional Form $f(x)$	Linear	7.52	4.88	8.04	6.43	6.50	6.75	5.46
	Quadratic	2.43	5.54	9.36	1.17	6.06	1.20	3.82
	Exponential	13.36	5.83	8.90	6.90	3.76	4.13	2.44

Variance of X	5	4.28	5.58	9.97	3.65	5.67	3.36	4.58
	40	5.94	3.37	5.60	6.16	4.96	4.47	2.89
	200	13.09	7.29	10.73	4.68	5.70	4.23	4.26
Model fit k	1	8.17	5.32	8.04	5.01	5.19	4.40	4.29
	10	8.39	7.74	6.93	4.82	4.35	3.63	2.35
	30	6.76	3.19	11.33	4.67	6.80	4.04	5.08

4.4 On the precision of the estimates of the total when sampling with PPAS compared with sampling methods

Table 4 shows the standard errors of estimate of population total. A lower standard error means better precision of estimates which is valuable in any estimation procedure. Under the quadratic form $f(x)$, SRSWOR consistently have lower standard error across different sampling rates as compared to the other designs. Since the quadratic form resulted to a population with the least variability, SRSWOR is expected to perform well under this scenario. In contrast, in the population generated using the exponential form, SRSWOR estimates have the highest standard errors as compared to the other designs, especially at 1% and 5% sampling rate. The exponential form mimics a skewed population, and in theory, estimation using SRSWOR could be disadvantageous under this scenario.

Under the linear form of f(x), PPAS sampling estimates have remarkably low standard errors compared to the other designs when there is high variability in the population (Var (X) = 200), and the model fit is good (k = 1). And when model fit is poor (k = 30), standard errors of estimates when sampling with PPAS are noted to be higher but still better than the precision of SRSWOR estimates. Lohr (2010) pointed out that "if the model does not fit the data well, ratio or regression estimation might not increase precision for estimated means and totals". This behavior, however, is no longer present in the quadratic and exponential models. Standard errors under the PPSS sampling generally follow the same behavior as PPAS sampling estimates.

At 10% sampling rate, SRSWOR and PPAS sampling method produced comparable precision of estimates, with SRSWOR being slightly better under quadratic form of $f(x)$ while PPAS sampling estimates performing a little better at high variance of X, especially under linear and exponential form of $f(x)$.

Table 4. Comparison of the Estimated Standard Errors of PPAS Sampling Estimates of the Population Total with those of Two Sampling Methods Under Different Combinations of Functional Form of Target Variable (Y), Variance of Auxiliary Variable (X), and Model Fit (k) Using Three Sampling Rates

$f(x)$	Variance of X	Model fit (k)	Sampling Rate						
			1%			5%		10%	
			SRS	PPAS	PPSS	SRS	PPAS	SRS	PPAS
Linear	5	1	1107.2	13756.4	10176.3	1022.0	6308.9	647.2	4072.0
		10	5386.2	16740.0	15146.7	1720.6	5828.3	1205.5	4031.7
		30	9815.4	12377.0	19741.8	4569.5	6885.5	2841.8	5284.5
	40	1	7302.1	9839.3	6452.8	2730.1	3202.6	1751.6	2213.4
		10	3351.7	7545.8	5652.6	3004.5	4804.1	2125.6	2862.8
		30	12778.9	15173.7	14655.3	4772.2	6369.5	3108.2	4391.0
	200	1	18823.0	304.4	262.6	5713.9	164.3	4321.4	90.9
		10	15155.6	3706.5	2984.6	6222.8	1438.4	3889.9	1089.1
		30	19459.2	8662.4	6666.4	6303.6	3872.3	5043.3	2681.4
Quadratic	5	1	655.6	18999.2	18583.3	477.5	8583.2	369.8	5596.4
		10	4172.1	22867.7	19746.4	1302.7	7901.1	984.4	5496.7
		30	10034.1	15411.7	17416.1	4092.4	8515.9	2501.5	6529.0
	40	1	3339.1	22630.9	13208.0	1490.6	7376.7	1111.7	5156.8
		10	5592.6	17816.2	32558.8	2437.7	10093.9	1300.8	6127.5
		30	6808.9	21810.6	15779.9	4847.6	10691.7	3139.7	7368.0
	200	1	9968.9	21405.2	22907.1	6009.4	9554.5	2605.9	7050.9
		10	12486.1	11641.4	21923.1	5136.4	8868.5	2093.2	6173.8
		30	10831.7	23997.1	29625.9	4813.7	11364.1	3903.2	6443.7
Exponential	5	1	4508.5	9160.4	5708.4	2314.4	4407.3	1549.5	2753.0
		10	4316.2	11700.6	4177.5	2745.3	4051.9	1632.1	2781.4
		30	11981.9	10547.3	7967.7	4009.8	5564.5	3271.9	4332.6
	40	1	15385.6	2889.3	2104.5	6709.8	1277.1	5001.8	1057.8
		10	32079.1	4527.4	4667.7	6534.5	2175.9	3929.4	1740.6
		30	10702.0	10394.6	9773.3	9261.5	3998.1	5707.5	3543.2
	200	1	14852.7	28408.8	39129.3	16439.3	11725.0	9970.2	10421.3
		10	30794.5	11477.6	22029.2	33923.2	10075.5	9975.8	8564.3
		30	43574.1	17101.6	41768.3	18745.2	16855.4	11819.0	9218.5

5. Conclusions and Recommendations

The findings of the study are summarized as follows:

1. Sampling with PPAS produce more accurate estimates of the population total than SRSWOR at low sampling rates.

2. At high sampling rates, PPAS sampling and SRSWOR tend to produce similar estimates in terms of bias and precision. SRSWOR being slightly better at low population variability and PPAS being a little better at skewed populations.
3. Estimation under PPAS sampling works better under the linear form of relationship between the target variable Y and the auxiliary variable. If model fit is good, PPAS sampling estimates perform consistently well even under heterogeneity of target variable. This means that PPAS estimates are affected more by poor model fit between the target and auxiliary variable, rather than the variability in the population.
4. Under the exponential form of relationship between the auxiliary and target variable, Sampling with PPAS works best, in terms of bias and precision of estimates, when population variability is moderate, and the model fit is good.

The findings of the study may only reflect the simulated data and may not necessarily be true for other random generators. It is advisable to verify these findings by recreating the data using different random seeds used in the study. Since sampling with PPSS was found to be unsuitable for this population, other sampling designs can be considered to facilitate better comparison.

References

CHAUVET, G., 2018, Large sample properties of the Midzuno sampling scheme, Hal-01882304v1. Available at: https://hal.archives-ouvertes.fr/hal-01882304/document

EFRON B., 1992, Bootstrap methods: Another look at the jackknife, In: Kotz S., Johnson N.L. (eds) *Breakthroughs in Statistics*, Springer Series in Statistics (Perspectives in Statistics). Springer, New York, NY.

GAURAN, I. and POBLADOR, M., 2012, Sampling with probability proportional to aggregate size using nonparametric bootstrap in estimating total production area of top cereals and root crops across Philippine regions, *The Philippine Statistician,* 61 (1), pp. 87-108.

HOMA, F., MAURYA, S., and SINGH G.N., 2016, On the use of several auxiliary variables in estimation of current population mean in successive sampling, *Communication in Statistics* – Theory and Methods, 45 (11).

KWONG, A.A., 2011, Nonparametric model-based predictive estimation in survey sampling. *The Philippine Statistician,* 60, pp. 1-14.

LAHIRI, D.B., 1951, A method of sample selection providing unbiased ratio estimates, *Bulletin of the International Statistical Institute*, 53, pp. 133-140.

LOHR, S., 2010, *Sampling: Design and Analysis*, 2nd Ed. Boston: Brooks/Cole.

MIDZUNO, H., 1952, On the sampling system with probability proportional to sum of sizes, *Annals of the Institute of Statistical Mathematics*, 3, pp. 99-107.

SÄRNDAL, C., SWENSSON, B., WRETMAN J., 1992, *Model Assisted Survey Sampling*. Springer Series in Statistics, p 201.