symmetry

MDPI

Article

Enhanced Estimation Methods Using Auxiliary Information for Rare and Clustered Populations

Pannarat Guayjarernpanishk [1], Supawadee Wichitchan [2], Chawalit Boonpok [2], Monchaya Chiangpradit [2] and Nipaporn Chutiman [2,*]

[1] Department of Mathematics, Faculty of Science, Khon Kaen University, Khon Kaen 40002, Thailand; panngu@kku.ac.th
[2] Department of Mathematics, Faculty of Science, Mahasarakham University, Maha Sarakham 44150, Thailand; supawadee.wi@msu.ac.th (S.W.); chawalit.b@msu.ac.th (C.B.); monchaya.c@msu.ac.th (M.C.)
* Correspondence: nipaporn.c@msu.ac.th

Abstract

This research proposes two new estimators which use auxiliary information to derive estimates of population means in cases where populations are considered rare and clustered in line with the General Inverse Adaptive Cluster Sampling (GIACS) framework. A ratio-type estimator is the first to be proposed, while a correlation-type estimator is the second. In this study, theoretical analyses were conducted to derive explicit first-order approximations of the bias and mean squared error for each estimator. The results show that, although the estimators are biased in finite samples, the bias decreases as the sample size increases and converges to zero, implying that the estimators are asymptotically unbiased. Simulation data with symmetric distributions were generated to evaluate their performance. The findings showed that the GIACS estimators using auxiliary information are more efficient than those without. When comparing the performance of estimators that use auxiliary information, it was found that the two proposed estimators, the ratio-type and correlation-type estimators, offered superior efficiency when compared to regression-type estimators. These results confirm that applying auxiliary information can significantly improve the accuracy of estimations within the GIACS framework, making the proposed estimators highly suitable for practical applications involving rare and clustered population.

Keywords: auxiliary information; clustered populations; general inverse sampling; adaptive cluster sampling; mean square error; relative efficiency

1. Introduction

Sampling for rare and clustered populations is an important issue in survey statistics because traditional sampling methods, including SRSWOR (simple random sampling without replacement), often yield inefficient estimates when the population is clustered. The likelihood of discovering rare units within the sample is thus reduced, while the estimator exhibits a variance which is usually higher than with methods specifically designed for rare and clustered populations. Thompson [1] first proposed adaptive cluster sampling (ACS), which selects an initial sample unit using SRSWOR. Should any unit satisfy a specific condition, the sample is augmented by its neighboring units, thereby automatically "expanding the networks" around the unit of interest and reducing the variance of the estimator when the population is clearly clustered. Numerous investigations of rare and clustered populations have employed ACS, which was particularly useful when investigating the COVID-19 pandemic [2–4], as well as IoT-based applications and autonomous

Received: 2 January 2026
Revised: 5 February 2026
Accepted: 10 February 2026
Published: 11 February 2026

systems [5,6], surveys of informal businesses [7,8], hydroacoustic fisheries surveys [9], the Gulf of Alaska rockfish survey [10], herpetofauna surveys in tropical rainforests [11], and geophysical surveys in CO2 emissions [12].

Over the past decade, ACS has been further developed theoretically, particularly in the development of estimators based solely upon the investigated variable. For instance, ref. [13] presented bootstrap confidence intervals for adaptive cluster sampling, and [14] introduced a quasi-Bayesian approach in ACS. In addition, estimators using auxiliary information have also been developed. For instance, refs. [15,16] proposed a ratio-type estimator using the mean of the auxiliary variable. Refs. [17,18] developed estimators whereby other auxiliary variables, such as the population coefficient of variation, provided the additional information. Their approach also made use of the kurtosis of the auxiliary variable, the coefficients of skewness, and the correlation coefficient linking the variable under investigation and the auxiliary variable. Furthermore, the work of [19–22] proposed enhanced estimators of the exponential ratio-type and product-type, while ref. [23] developed estimators using some robust measures of the auxiliary variable for estimator development in ACS.

However, while ACS is highly efficient in many situations, it still has a significant limitation when the population contains very few units that meet the specified criterion. This results in incomplete sampling of the desired target units. To address this limitation, general inverse adaptive cluster sampling (GIACS) was the technique proposed by Salehi and Seber [24]. This approach integrates elements of general inverse sampling [25] and adaptive cluster sampling. A sequential selection process is used in GIACS together with predefined stopping rules to control the number of rare units randomly included in the sample. This technique requires the random selection of initial units on a sequential basis until a predetermined number of rare units is acquired. The randomization terminates when the random stopping condition is reached. If any of the randomly selected units meet the specified criteria, the network of those units is added to the sample. Therefore, GIACS is a more controlled and efficient sampling method for rare and clustered populations.

Nevertheless, since the development of estimators in GIACS remains limited and most studies have focused only on estimators using the variable of interest, ref. [26] initiated the development of estimators which are able to utilize auxiliary information within the GIACS framework. They proposed a regression-type estimator and demonstrated that using auxiliary information can significantly reduce bias and mean squared error, especially in cases involving a rare and clustered population network. Although the regression-type estimator provides an important first step in incorporating auxiliary information into the GIACS framework, its estimation error remains relatively large for rare and highly clustered populations. In this study, we propose a methodological and theoretical extension by developing two alternative estimators using auxiliary information: a correlation-type estimator and a ratio-type estimator. The correlation-type estimator belongs to the class of difference estimators and replaces the regression coefficient with a correlation coefficient, which is often known or can be reliably obtained from historical surveys or large auxiliary databases, making it practically attractive. In contrast, the ratio-type estimator is formulated as a multiplicative estimator based on a proportional structure and belongs to the class of ratio estimators. It is well established in sampling theory that ratio estimators tend to outperform regression estimators when the relationship between the study variable and the auxiliary variable passes through or lies close to the origin—a condition that commonly arises in rare and clustered populations. Section 2 presents the background of General Inverse Adaptive Cluster Sampling, while in Section 3, the proposed estimators are introduced through theoretical analysis of their properties, namely bias and MSE. In addition, Section 4 presents simulation studies to evaluate estimator efficiency. Conclusions are then presented in Section 5.

https://doi.org/10.3390/sym18020334

2. Background of General Inverse Adaptive Cluster Sampling

The population of size N units consist of a value set $U = (y_1, y_2, \ldots, y_N)$. Let Y be the variable of interest, and y_i be the y-value associated with the ith unit. The population of N units is divided into two groups based on their y-values. The group whose y-values satisfy a given condition ($Cond$) is $U_M = \{y : y_i \in Cond;\ i = 1, 2, \ldots N\}$, where M is the total number of unknown units in U_M, and the group whose y-values do not satisfy the specified condition ($Cond$) is $U_{N-M} = U_{M'} = \{y : y_i \notin Cond;\ i = 1, 2, \ldots N\}$. Sampling commences with the random selection of n_I units via the SRSWOR approach. Should there be, among these n_I units, no fewer than the pre-specified number of rare units, r, from U_M which can satisfy the condition stated ($Cond$), this marks the end of the initial sampling stage; if not, the sampling process will proceed further until r units from U_M are satisfied, whereupon our final initial sample size will be n_T or n_F units are selected in the total. This final initial sample size, however, is set as $n_F = N$, which can be classified into three cases as follows:

Condition 1 $(C1)$: number of units contained within the initial sample, n_I, meeting the designated condition is no fewer than r, so the final initial sample size reached equals $n_T = n_I$.

Condition 2 $(C2)$: number of units contained within the initial sample, n_I, is fewer than r, so the sampling process proceeds further until r units from U_M meet the condition; that is $n_I < n_T < n_F$ or $n_T = n_F$ & $|S_M| = r$ in which S_M indicates the index set of sample units, representing the numbers of U_M.

Condition 3 $(C3)$: number of units contained within the initial sample, n_I, is fewer than r, so the sample proceeds further to n_F. Should the number of units which meet the designated condition fall below r, this results in $n_T = n_F$ & $|S_M| < r$.

In the case of initial units which are selected and meet the designated condition $(Cond)$ the sample will be expanded to include the whole of the network of those units. The final sample therefore comprises those initial units which were selected via the general inverse sampling approach, in addition to all of the units which are part of the network linked to those selected initial units.

The following unbiased estimator for population mean was proposed by [24] under GIACS in each condition as shown below:

For Condition 1 $(C1)$: $\overline{y}_{GIACS} = \frac{1}{n_I}\sum_{i=1}^{n_I} \overline{y}_i^*$, in which $\overline{y}_i^*$ represents the mean value of the variable under investigation in the network which contains unit i from the initial sample, which has the definition of $\overline{y}_i^* = \frac{1}{m_i}\sum_{j\in\psi_i} y_j$, whereby ψ_i indicates the network containing unit i, and m_i shows how many units the network contains.

For Condition 2 $(C2)$: using Murthy's estimator $\overline{y}_{GIACS} = \dfrac{1}{N}\sum_{i=1}^{n_I}\dfrac{\Pr(s|i)}{\Pr(s)}\overline{y}_i^* = \dfrac{1}{N}\left[\dfrac{N}{n_I - 1}\sum_{i\in S_M}\overline{y}_i^* + \sum_{i\in S_{M'}}\overline{y}_i^*\right]$, $\overline{y}_{GIACS} = \frac{1}{N}\left(\hat{M}\overline{y}_M^* + (N - \hat{M})\overline{y}_{M'}^*\right)$, in which $\Pr(s)$ indicates the probability of acquiring sample s, $\Pr(s|i)$ indicates the conditional probability of acquiring s on the condition that ith unit was initially selected, p_i indicates the probability of selecting unit i when drawing first. In application, M is unknown, but M has an unbiased estimator proposed by [25] as: $\hat{M} = \frac{N(r-1)}{(n_T-1)}$.

For Condition 3 $(C3)$: $\overline{y}_{GIACS} = \frac{1}{n_F}\sum_{i=1}^{n_F}\overline{y}_i^*$.

Therefore, an unbiased estimator of the population mean under GIACS as follows:

$$\overline{y}_{GIACS} = \begin{cases} \dfrac{1}{n_I}\sum_{i=1}^{n_I}\overline{y}_i^* & if\ C1, \\ \dfrac{1}{N}\left(\hat{M}\overline{y}_M^* + (N - \hat{M})\overline{y}_{M'}^*\right) & if\ C2, \\ \dfrac{1}{n_F}\sum_{i=1}^{n_F}\overline{y}_i^* & if\ C3. \end{cases} \tag{1}$$

The variance of $\overline{y}_{GIACS}$ for each condition as follows:

For Condition 1 ($C1$): $V(\overline{y}_{GIACS}) = \frac{(N-n_I)}{N}\frac{S_y^{*2}}{n_I}$, where $S_y^{*2} = (N-1)^{-1}\sum_{i=1}^{N}(\overline{y}_i^* - \mu_y)^2$.

For Condition 2 ($C2$): Using the variance of Murthy's estimator [27], $V(\overline{y}_I^*) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j<i}^{N}\left(1 - \sum_{s\ni i,j}\frac{\Pr(s|i)\Pr(s|j)}{\Pr(s)}\right)\left(\frac{\overline{y}_i^*}{p_i} - \frac{\overline{y}_j^*}{p_j}\right)^2 p_i p_j$, in which $\Pr(s)$ indicates the probability of acquiring sample s, $\Pr(s|i)$ indicates the conditional probability of acquiring s on the condition that ith unit was initially selected, p_i indicates the probability of selecting unit i when drawing first.

Therefore, $\overline{y}_{GIACS}$ has the variance of:

$$V(\overline{y}_{GIACS}) = \begin{cases} \frac{(N-n_I)}{N}\frac{S_y^{*2}}{n_I} & if\ C1, \\ V(\overline{y}_I^*) & if\ C2, \\ \frac{(N-n_F)}{N}\frac{S_y^{*2}}{n_F} & if\ C3. \end{cases} \quad (2)$$

Ref. [26] presented the regression-type estimator for GIACS:

$$\overline{y}_{GIACS,Reg} = \overline{y}_{GIACS} + B^*(\mu_x - \overline{x}_{GIACS}) \quad (3)$$

where $B^* = \frac{\sum_{i=1}^{N}\overline{x}_i^*\overline{y}_i^* - N\mu_y\mu_x}{\sum_{i=1}^{N}\overline{x}_i^{*2} - N\mu_x^2}$ the estimator $\overline{x}_{GIACS}$ can be calculated similarly to $\overline{y}_{GIACS}$, using the information from the auxiliary variable. The estimators of μ_y and μ_x are $\overline{y}_{GIACS}$ and $\overline{x}_{GIACS}$, respectively, while $\sum_{i=1}^{N}\overline{x}_i^*\overline{y}_i^*$ and $\sum_{i=1}^{N}\overline{x}_i^{*2}$ can be estimated via the application of Murthy's estimator [24,27]. These estimators are shown below as: $\hat{T}_{xy}^* = \sum_{i=1}^{n_T}\frac{\Pr(s|i)}{\Pr(s)}\overline{x}_i^*\overline{y}_i^*$ and $\hat{T}_{xx}^* = \sum_{i=1}^{n_T}\frac{\Pr(s|i)}{\Pr(s)}\overline{x}_i^*\overline{x}_i^*$.

For Condition 1 (C1): $\hat{T}_{xy}^* = \frac{N}{n_I}\sum_{i=1}^{n_I}\overline{x}_i^*\overline{y}_i^*$ and $\hat{T}_{xx}^* = \frac{N}{n_I}\sum_{i=1}^{n_I}\overline{x}_i^*\overline{x}_i^*$.

Condition 2 $(C2)$: $\hat{T}_{xy}^* = \hat{M}\left(r^{-1}\sum_{i\in S_M}\overline{x}_i^*\overline{y}_i^*\right) + \left(N - \hat{M}\right)\left[(n_T - r)^{-1}\sum_{i\in S_{M'}}\overline{x}_i^*\overline{y}_i^*\right]$, and $\hat{T}_{xx}^* = \hat{M}\left(r^{-1}\sum_{i\in S_M}\overline{x}_i^*\overline{x}_i^*\right) + \left(N - \hat{M}\right)\left[(n_T - r)^{-1}\sum_{i\in S_{M'}}\overline{x}_i^*\overline{x}_i^*\right]$.

Condition 3 $(C3)$: $\hat{T}_{xy}^* = \frac{N}{n_F}\sum_{i=1}^{n_F}\overline{x}_i^*\overline{y}_i^*$ and $\hat{T}_{xy}^* = \frac{N}{n_F}\sum_{i=1}^{n_F}\overline{x}_i^*\overline{y}_i^*$.

3. Proposed Estimators Using Auxiliary Information in GIACS

This research presents two estimators using auxiliary information. The first estimator, motivated by [26], is a ratio estimator applied to construct an estimator in GIACS. The proposed estimator is as follows:

$$\overline{y}_{GIACS,R} = \frac{\overline{y}_{GIACS}}{\overline{x}_{GIACS}}\mu_x = \hat{R}^*\mu_x,\ \text{where}\ \hat{R}^* = \frac{\overline{y}_{GIACS}}{\overline{x}_{GIACS}}. \quad (4)$$

The first proposed estimator is called a ratio-type estimator in GIACS. In ratio estimation, auxiliary information from a variable that is positively and linearly related to the study variable is used to improve the estimation of the population mean.

$\overline{y}_{GIACS,R}$ is the biased estimator. The bias of $\overline{y}_{GIACS,R}$ is:

$$Bias\left(\overline{y}_{GIACS,R}\right) = \frac{\mu_y}{\mu_x^2}V(\overline{x}_{GIACS}) - \frac{1}{\mu_x}COV(\overline{y}_{GIACS}, \overline{x}_{GIACS})$$

With increasing sample size, there is a convergence to zero of the bias, showing that the proposed estimator has a mean squared error which is asymptotically equivalent to its variance.

$MSE\left(\overline{y}_{GIACS,R}\right)$ can be considered under three separate conditions as follows: Given $\alpha_i^* = \overline{y}_i^* - R^*\overline{x}_i^* \approx \overline{y}_i^* - \hat{R}^*\overline{x}_i^*$, it follows that $\overline{y}_{GIACS,R} = \overline{\alpha}^* + R^*\mu_x$.

For condition 1 $(C1)$ and condition 3 $(C3)$, consider Equation (2) by replacing $\overline{y}_i^*$ in the formula for S_y^{*2} with α_i^*.

For condition 2 $(C2)$ $V\left(\overline{y}_{GIACS,R}\right)_{C2} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j<i}^{N}\left[1 - \sum_{s\ni i,j}\frac{\Pr(s|i)\Pr(s|j)}{\Pr(s)}\right]\left(\frac{\alpha_i^*}{p_i} - \frac{\alpha_j^*}{p_j}\right)^2 p_i p_j$.

Accordingly, $\overline{y}_{GIACS,R}$ has a mean squared error which approximates to:

$$MSE\left(\overline{y}_{GIACS,R}\right) \approx V\left(\overline{y}_{GIACS,R}\right) = \begin{cases} \frac{(N-n_I)}{N}\frac{S_\alpha^{*2}}{n_I} & if\ C1, \\ V\left(\overline{y}_{GIACS,R}\right)_{C2} & if\ C2, \\ \frac{(N-n_F)}{N}\frac{S_\alpha^{*2}}{n_F} & if\ C3. \end{cases} \tag{5}$$

The estimator of the mean squared error of $\overline{y}_{GIACS,R}$ is:

$$M\hat{S}E\left(\overline{y}_{GIACS,R}\right) = \begin{cases} \frac{(N-n_I)}{N}\frac{s_{\alpha I}^{*2}}{n_I} & if\ C1, \\ \hat{V}\left(\overline{y}_{GIACS,R}\right)_{C_2} & if\ C2, \\ \frac{(N-n_F)}{N}\frac{s_{\alpha F}^{*2}}{n_F} & if\ C3. \end{cases} \tag{6}$$

where $\hat{V}\left(\overline{y}_{GIACS,R}\right)_{C_2} = \left(\frac{\hat{M}}{N}\right)^2\left(\frac{(N-n_T+1)(n_Tr-n_T-r)-N(n_T-2)}{N(n_T-2)(r-1)}\right)\frac{s_{\alpha M}^{*2}}{r} + \frac{1}{N^2}\hat{V}(\hat{M})\left(\overline{\alpha}_M^* - \overline{\alpha}_{M'}^*\right)^2 + \left(\frac{(N-n_T+1)(n_T-r-1)}{N(n_T-1)(n_T-2)}\right)s_{\alpha M'}^{*2}$, $\hat{V}(\hat{M}) = N\frac{(n_T-r)(r-1)(N-n_T+1)}{(n_T-1)^2(n_T-2)} = \left[1 - \frac{(n_T-1)}{N}\right]\frac{\hat{M}(N-\hat{M})}{(n_T-2)}$. $s_{\alpha M}^{*2} = (r-1)^{-1}\sum_{i\in S_M}\left(\alpha_i^* - \overline{\alpha}_M^*\right)^2$, $\overline{\alpha}_M^* = r^{-1}\sum_{i\in M}\alpha_i^*$, $s_{\alpha M'}^{*2} = (n_T - r - 1)^{-1}\sum_{i\in S_{M'}}\left(\alpha_i^* - \overline{\alpha}_{M'}^*\right)^2$, $\overline{\alpha}_{M'}^* = (n_T - r)^{-1}\sum_{i\in M'}\alpha_i^*$, $s_{\alpha I}^{*2} = (n_I - 1)^{-1}\sum_{i=1}^{n_I}\left(\alpha_i^* - \overline{\alpha}_I^*\right)^2$, $\overline{\alpha}_I^* = n_I^{-1}\sum_{i=1}^{n_I}\alpha_i^*$, and $s_{\alpha F}^{*2} = (n_F - 1)^{-1}\sum_{i=1}^{n_F}\left(\alpha_i^* - \overline{\alpha}_F^*\right)^2$, $\overline{\alpha}_F^* = n_F^{-1}\sum_{i=1}^{n_F}\alpha_i^*$.

The second estimator, motivated by [26,28], employs information derived from the correlation coefficient between $\overline{y}_i^*$ and $\overline{x}_i^*$. Equation (2) then has a regression coefficient which is substituted by the correlation coefficient between $\overline{y}_i^*$ and $\overline{x}_i^*$. This estimator is called a correlation-type estimator in GIACS and is defined as follows:

$$\overline{y}_{GIACS,\rho} = \overline{y}_{GIACS} + \hat{\rho}^*(\mu_x - \overline{x}_{GIACS}) \tag{7}$$

where $\hat{\rho}^* = \frac{\hat{T}_{xy}^* - N\overline{y}_{GIACS}\overline{x}_{GIACS}}{\sqrt{\left(\hat{T}_{xx}^* - N\overline{x}_{GIACS}^2\right)}\sqrt{\left(\hat{T}_{yy}^* - N\overline{y}_{GIACS}^2\right)}}$ is the estimator of the population correlation coefficient between $\overline{y}_i^*$ and $\overline{x}_i^*$, $\rho* = \frac{S_{x*y*}}{\sqrt{S_{x*}^2}\sqrt{S_{y*}^2}} = \frac{\sum_{i=1}^{N}\overline{x}_i^*\overline{y}_i^* - N\mu_y\mu_x}{\sqrt{\sum_{i=1}^{N}\overline{x}_i^{*2} - N\mu_x^2}\sqrt{\sum_{i=1}^{N}\overline{y}_i^{*2} - N\mu_y^2}}$.

For Condition 1 $(C1)$: $\hat{T}_{yy}^* = \frac{N}{n_I}\sum_{i=1}^{n_I}\overline{y}_i^*\overline{y}_i^*$.

Condition 2 $(C2)$: $\hat{T}_{yy}^* = \hat{M}\left(r^{-1}\sum_{i\in S_M}\overline{y}_i^*\overline{y}_i^*\right) + \left(N - \hat{M}\right)\left[(n_T - r)^{-1}\sum_{i\in S_{M'}}\overline{y}_i^*\overline{y}_i^*\right]$.

Condition 3 $(C3)$: $\hat{T}_{yy}^* = \frac{N}{n_F}\sum_{i=1}^{n_F}\overline{y}_i^*\overline{y}_i^*$.

$\overline{y}_{GIACS,\rho}$ is the biased estimator. The bias of $\overline{y}_{GIACS,\rho}$ is:

$$Bias(\overline{y}_{GIACS,\rho}) = \frac{\rho*}{2S_{y*}^2}COV\left(\overline{x}_{GIACS}, s_{y*}^2\right) + \frac{\rho*}{2S_{x*}^2}COV\left(\overline{x}_{GIACS}, s_{x*}^2\right) - \frac{\rho*}{S_{x*y*}}COV\left(\overline{x}_{GIACS}, S_{x*y*}\right)$$

Given $\gamma_i^* = \overline{y}_i^* - \rho^*\overline{x}_i^* \approx \overline{y}_i^* - \hat{\rho}^*\overline{x}_i^*$, it follows that $\overline{y}_{GIACS,\rho} = \overline{\gamma}^* + \rho^*\mu_x$.

$MSE\left(\overline{y}_{GIACS,\rho}\right)$ in the same way as $\overline{y}_{GIACS,R}$ by replacing α_i^* with γ_i^*, gives:

Therefore, the mean squared error of $\overline{y}_{GIACS,\rho}$ is approximated as:

$$MSE\left(\overline{y}_{GIACS,\rho}\right) \approx V\left(\overline{y}_{GIACS,\rho}\right) = \begin{cases} \frac{(N-n_I)}{N}\frac{S_{\gamma}^{*2}}{n_I} & if\ C1, \\ V\left(\overline{y}_{GIACS,\rho}\right)_{C2} & if\ C2, \\ \frac{(N-n_F)}{N}\frac{S_{\gamma}^{*2}}{n_F} & if\ C3. \end{cases} \tag{8}$$

The estimator of the mean squared error of $\overline{y}_{GIACS,\rho}$ is:

$$M\hat{S}E\left(\overline{y}_{GIACS,\rho}\right) = \begin{cases} \frac{(N-n_I)}{N}\frac{s_{\gamma I}^{*2}}{n_I} & if\ C1, \\ \hat{V}\left(\overline{y}_{GIACS,\rho}\right)_{C_2} & if\ C2, \\ \frac{(N-n_F)}{N}\frac{s_{\gamma F}^{*2}}{n_F} & if\ C3. \end{cases} \tag{9}$$

Appendix A shows the details of the derivations of the bias and mean squared error for the proposed estimators.

To summarize, our proposed estimators offer two complementary ways of incorporating auxiliary information into the GIACS framework. Both estimators are designed to improve efficiency and reduce estimation error in rare and clustered populations.

4. Simulation and Discussion

4.1. Simulation Studies

This study presented two data sets as follows:

Data set 1: A clustered population of 400 square units (arranged in 20 rows and 20 columns) was generated to represent both the auxiliary variable and the variable under investigation using a bivariate Poisson cluster process. The study region incorporated five randomly positioned parent points, with offspring generated around each parent in line with a symmetric Gaussian distribution which exhibited a dispersion parameter of 0.05. The expected number of offspring per parent was set to 1000 [29], as shown in Figures 1 and 2, in which μ_y = 59.7375 and the x and y data had a correlation coefficient of 0.9973.

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	34	108	35	8	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	11	293	1022	375	7	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	2	308	1182	427	7	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	44	103	40	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	16	33	8	0	0	0	0	5	0	0	0	0	0	0
0	0	0	0	0	21	434	645	123	0	0	0	19	67	48	16	17	16	0	0
0	0	0	0	42	746	1230	208	0	0	4	277	818	306	193	657	337	16	0	0
0	0	0	0	19	177	243	26	0	0	2	362	1306	599	345	1371	558	18	0	0
0	0	0	0	0	5	4	0	0	0	5	38	222	84	91	207	133	8	0	0
4	0	0	0	0	0	0	0	3	7	0	0	0	0	0	0	0	25	0	0
0	0	0	0	4	2	0	0	0	0	0	0	0	32	355	735	120	3	0	0
0	0	0	6	179	383	97	0	0	0	0	0	0	50	727	1241	245	13	0	0
0	0	0	11	576	1357	287	8	0	0	0	0	0	0	130	270	62	0	0	0
0	0	0	0	181	517	126	3	0	0	0	0	0	0	0	0	0	0	0	0

Figure 1. y data of Population I.

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	8	29	9	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	3	74	257	98	3	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	73	290	109	2	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	10	26	10	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	5	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	4	8	2	0	0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	5	108	160	31	0	0	0	4	17	13	4	4	4	0	0
0	0	0	0	11	189	310	52	0	0	1	55	205	76	48	160	84	5	0	0
0	0	0	0	5	45	60	7	0	0	1	87	323	147	87	337	138	5	0	0
0	0	0	0	0	1	1	0	0	0	1	9	57	21	22	55	37	2	0	0
1	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	1	0	0
0	0	0	0	2	1	0	0	0	0	0	0	0	7	91	177	31	1	0	0
0	0	0	2	44	96	25	0	0	0	0	0	0	13	185	311	60	3	0	0
0	0	0	3	148	337	72	2	0	0	0	0	0	0	30	66	14	0	0	0
0	0	0	0	47	129	31	1	0	0	0	0	0	0	0	0	0	0	0	0

Figure 2. x data of Population I.

Data set 2: A population of the auxiliary variable X uses real data, which is the data of ring-necked ducks from an area of 5000 square kilometers in central Florida [29]. The total area was partitioned into 50 sampling units, each comprising 100 km^2 (arranged in 5 rows and 10 columns). The variable of interest Y was generated using a linear model of the form $y = 4x + \varepsilon$ model, where $\varepsilon \sim N(0, y_i)$. This specification ensures a symmetric error distribution centered at zero [30,31]. This is shown in Figures 3 and 4, with μ_y = 1641.48, while the correlation coefficient linking the x and y data is equal to 0.9937.

0	91	632	302	0	0	0	0	3054	0
9089	50,249	1020	4551	20	0	0	89	0	236
0	0	0	0	442	0	14	0	7615	0
0	0	0	0	0	0	0	0	0	0
0	0	4	0	0	0	0	0	0	4666

Figure 3. y data of Population II.

0	20	200	75	0	0	0	0	675	0
4000	13,500	234	1335	4	0	0	35	0	55
0	0	0	0	97	0	4	0	1815	0
0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	2	0	0	1283

Figure 4. x data of Population II.

In both of the populations, the definition of the criterion under which units could be added to the sample was $Cond = \{y : y > 0\}$, while the networks of the x data had positions which were aligned with the y data.

In the simulation study, GIACS was implemented using a sequential inverse sampling algorithm with a fixed stopping rule r. At each iteration, simple random sampling without replacement (SRSWOR) was used to select an initial sample of size n_I. If the number of units satisfying the specified condition ($Cond$) was greater than or equal to r, no further random selection was performed. Otherwise, sampling continued sequentially until the number of initial units meeting the condition equaled r. The values of r and n_I varied across scenarios, while n_F was set to 100 and 25 for Data Set 1 and Data Set 2, respectively.

For all initial units satisfying *Cond*, the entire network of those units was then added to the final sample using an adaptive cluster sampling procedure.

Each estimator underwent a total of $T = 10,000$ iterations. The expected sequence of final initial sample sizes from general inverse sampling is defined as $E(n_T)$, and the expected final sample size for adaptive cluster sampling is defined as $E(\nu)$.

The proposed estimators using auxiliary variables are biased estimators, and the estimated absolute relative biases are defined as: $A\hat{R}B\left(\overline{y}_{GIACS,Reg}\right) = \frac{1}{T}\sum_{i=1}^{T} \frac{\left|\left(\overline{y}_{GIACS,Reg}\right)_i - \mu_y\right|}{\mu_y}$, $A\hat{R}B\left(\overline{y}_{GIACS,R}\right) = \frac{1}{T}\sum_{i=1}^{T} \frac{\left|\left(\overline{y}_{GIACS,R}\right)_i - \mu_y\right|}{\mu_y}$, and $A\hat{R}B\left(\overline{y}_{GIACS,\rho}\right) = \frac{1}{T}\sum_{i=1}^{T} \frac{\left|\left(\overline{y}_{GIACS,\rho}\right)_i - \mu_y\right|}{\mu_y}$.

The estimators have an estimated mean square error shown as: $M\hat{S}E(\overline{y}_{GIACS}) = \frac{1}{T}\sum_{i=1}^{T}\left[(\overline{y}_{GIACS})_i - \mu_y\right]^2$, $M\hat{S}E\left(\overline{y}_{GIACS,Reg}\right) = \frac{1}{T}\sum_{i=1}^{T}\left[\left(\overline{y}_{GIACS,Reg}\right)_i - \mu_y\right]^2$ $M\hat{S}E\left(\overline{y}_{GIACS,R}\right) = \frac{1}{T}\sum_{i=1}^{T}\left[\left(\overline{y}_{GIACS,R}\right)_i - \mu_y\right]^2$, and $M\hat{S}E\left(\overline{y}_{GIACS,\rho}\right) = \frac{1}{T}\sum_{i=1}^{T}\left[\left(\overline{y}_{GIACS,\rho}\right)_i - \mu_y\right]^2$.

The proposed estimators have relative efficiency (RE) when compared to $\overline{y}_{GIACS}$ which can be defined by:

$RE1\left(\overline{y}_{GIACS,R}\right) = \frac{M\hat{S}E(\overline{y}_{GIACS})}{M\hat{S}E(\overline{y}_{GIACS,R})}$ and $RE1\left(\overline{y}_{GIACS,\rho}\right) = \frac{M\hat{S}E(\overline{y}_{GIACS})}{M\hat{S}E\left(\overline{y}_{GIACS,\rho}\right)}$.

The proposed estimator has relative efficiency (RE) which can be compared to $\overline{y}_{GIACS,Reg}$ and is defined as

$RE2\left(\overline{y}_{GIACS,R}\right) = \frac{M\hat{S}E\left(\overline{y}_{GIACS,Reg}\right)}{M\hat{S}E(\overline{y}_{GIACS,R})}$ and $RE2\left(\overline{y}_{GIACS,\rho}\right) = \frac{M\hat{S}E\left(\overline{y}_{GIACS,Reg}\right)}{M\hat{S}E\left(\overline{y}_{GIACS,\rho}\right)}$.

The relative efficiency (RE) is used to compare the performance of the proposed estimator with that of a competing estimator. An RE value greater than one indicates that the proposed estimator is more efficient than the competing estimator.

4.2. Discussion

This study examined scenarios in which the study variable is positively correlated with the auxiliary variables. Under these settings, numerical analyses were conducted to evaluate the estimated absolute relative bias $\left(A\hat{R}B\right)$, estimated mean squared error $\left(M\hat{S}E\right)$, and relative efficiency (RE) of the proposed estimators. The numerical results for Population I are reported in Tables 1–3, while those for Population II are presented in Tables 4–6.

Table 1. Estimated absolute relative bias of the proposed estimators for Population I.

r	n_I	$E(n_T)$	$E(\nu)$	$ARB\left(\overline{y}_{GIACS,Reg}\right)$	$ARB\left(\overline{y}_{GIACS,R}\right)$	$ARB\left(\overline{y}_{GIACS,\rho}\right)$
3	5	12.5857	70.3417	0.1778	0.3163×10^{-3}	0.0097
3	10	13.0820	72.5616	0.2297	0.2226×10^{-3}	0.0059
3	15	16.2061	79.2187	0.1241	0.2561×10^{-3}	0.0145
3	20	20.4720	87.6895	0.1482	0.3013×10^{-3}	0.0145
4	10	16.7825	81.395	0.1797	0.1691×10^{-3}	0.0124
4	15	18.0858	84.5285	0.1617	0.2327×10^{-3}	0.0121
4	20	21.5862	91.1579	0.1295	0.3850×10^{-4}	0.0106
5	15	20.8697	91.4705	0.1384	0.7366×10^{-4}	0.0141
5	20	23.1604	95.4267	0.1649	0.3013×10^{-4}	0.0070
5	25	26.8037	100.7264	0.1122	0.2678×10^{-4}	0.0141
5	30	30.0993	107.0549	0.0593	0.5357×10^{-4}	0.0143
10	30	41.7951	122.1784	0.0841	0.1841×10^{-4}	0.0142
10	40	44.5705	124.8715	0.0984	0.1171×10^{-4}	0.0136
10	50	51.3650	131.6681	0.0636	0.3348×10^{-4}	0.0075

https://doi.org/10.3390/sym18020334

Table 2. Estimated mean squared error of the estimators for Population I.

r	n_I	$\hat{MSE}(\overline{y}_{GIACS})$	$\hat{MSE}\left(\overline{y}_{GIACS,Reg}\right)$	$\hat{MSE}\left(\overline{y}_{GIACS,R}\right)$	$\hat{MSE}\left(\overline{y}_{GIACS,\rho}\right)$
3	5	1548.6267	1155.3721	0.0980	873.6444
3	10	897.4571	612.5440	0.0825	506.3457
3	15	654.0586	410.9425	0.0731	369.1502
3	20	495.4843	368.0048	0.0607	279.5675
4	10	878.4212	631.9332	0.0733	495.3398
4	15	550.4195	320.5219	0.0605	310.7505
4	20	599.0873	393.9126	0.0561	337.7988
5	15	519.0368	333.8360	0.0572	292.7755
5	20	543.7874	345.7475	0.0511	306.3742
5	25	510.0624	312.0688	0.0444	287.3891
5	30	419.4019	307.2793	0.0361	236.3688
10	30	291.8287	165.5705	0.0261	164.4759
10	40	234.4412	122.6420	0.0178	120.2456
10	50	194.9591	115.2926	0.0145	109.9623

Table 3. Relative efficiency of the proposed estimators for Population I.

r	n_I	Compared with $\overline{y}_{GIACS}$		Compared with $\overline{y}_{GIACS,Reg}$	
		$RE1\left(\overline{y}_{GIACS,R}\right)$	$RE1\left(\overline{y}_{GIACS,\rho}\right)$	$RE2\left(\overline{y}_{GIACS,R}\right)$	$RE2\left(\overline{y}_{GIACS,\rho}\right)$
3	5	15,802.3133	1.7726	11,789.5112	1.3225
3	10	10,878.2679	1.7724	7424.7758	1.2097
3	15	8947.4501	1.7718	5621.6484	1.1132
3	20	8162.8386	1.7723	6062.6820	1.3163
4	10	11,983.9181	1.7734	8621.1896	1.2758
4	15	9097.8430	1.7713	5297.8826	1.0314
4	20	10,678.9180	1.7735	7021.6150	1.1661
5	15	9074.0699	1.7728	5836.2937	1.1402
5	20	10,641.6321	1.7749	6766.0959	1.1285
5	25	11,487.8919	1.7748	7028.5766	1.0859
5	30	11,617.7812	1.7744	8511.8920	1.3000
10	30	11,181.1762	1.7743	6343.6973	1.0067
10	40	13,170.8539	1.9497	6890.0000	1.0199
10	50	13,445.4552	1.7730	7951.2138	1.0485

Table 4. Estimated absolute relative bias of the proposed estimators for Population II.

r	n_I	$E(n_T)$	$E(\nu)$	$\hat{ARB}\left(\overline{y}_{GIACS,Reg}\right)$	$\hat{ARB}\left(\overline{y}_{GIACS,R}\right)$	$\hat{ARB}\left(\overline{y}_{GIACS,\rho}\right)$
3	5	9.2102	15.3681	0.1816	0.0367	0.0132
3	8	9.9232	16.1825	0.1783	0.0176	0.0150
4	8	11.9530	18.0843	0.1584	0.0114	0.0222
4	10	12.7163	18.8986	0.1622	0.0075	0.0222
5	10	15.7913	21.6347	0.0880	0.0097	0.0214
5	15	16.7308	22.4200	0.1093	0.0076	0.0063

Table 5. Estimated mean squared error of the estimators for Population II.

r	n_I	$\hat{MSE}(\overline{y}_{GIACS})$	$\hat{MSE}\left(\overline{y}_{GIACS,Reg}\right)$	$\hat{MSE}\left(\overline{y}_{GIACS,R}\right)$	$\hat{MSE}\left(\overline{y}_{GIACS,\rho}\right)$
3	5	1,115,744.0409	817,341.0988	88,224.6723	575,799.7640
3	8	774,357.8219	583,268.5982	26,258.1505	395,310.3420
4	8	703,027.8236	472,216.0826	8701.1413	357,373.0960
4	10	643,806.1262	473,681.2131	7338.2065	331,566.7900
5	10	467,187.7845	250,591.2071	3884.6609	240,861.1568
5	15	358,956.1348	246,956.8459	3129.4146	184,155.1328

Table 6. Relative efficiency of the proposed estimators for Population II.

r	n_I	Compared with $\overline{y}_{GIACS}$		Compared with $\overline{y}_{GIACS,Reg}$	
		$RE1\left(\overline{y}_{GIACS,R}\right)$	$RE1\left(\overline{y}_{GIACS,\rho}\right)$	$RE2\left(\overline{y}_{GIACS,R}\right)$	$RE2\left(\overline{y}_{GIACS,\rho}\right)$
3	5	12.6466	1.9377	9.2643	1.4195
3	8	29.4902	1.9589	22.2129	1.4755
4	8	80.7972	1.9672	54.2706	1.3214
4	10	87.7334	1.9417	64.5500	1.4286
5	10	120.2648	1.9397	64.5079	1.0404
5	15	114.7039	1.9492	78.9147	1.3410

Because estimators incorporating auxiliary information are generally biased, their performance is first assessed using the estimated absolute relative bias. Tables 1 and 4 report the ARB values for Populations I and II, respectively, and reveal consistent patterns across both populations. For the regression-type, ratio-type, and correlation-type estimators, increasing either the initial sample size or the predetermined number of rare units leads to a systematic reduction in $A\hat{R}B$. Notably, the proposed ratio-type and correlation-type estimators exhibit extremely small $A\hat{R}B$ values—often effectively zero—across all combinations of initial sample size and rare-unit thresholds, indicating negligible bias in practice.

The estimated mean squared errors for Populations I and II are summarized in Tables 2 and 5, respectively. Across all simulation settings, estimators utilizing auxiliary information uniformly achieve lower $M\hat{S}Es$ than the estimator that does not use auxiliary variables. Moreover, both proposed estimators consistently outperform the regression-type estimator in terms of $M\hat{S}E$. Among the three, the ratio-type estimator yields the smallest $M\hat{S}Es$ across nearly all designs considered, and this improvement is statistically significant, demonstrating a clear efficiency advantage.

A similar pattern is observed when comparing relative efficiencies. Tables 3 and 6 report the RE values for Populations I and II, respectively, showing that all estimators incorporating auxiliary information achieve RE values greater than one, confirming their superior efficiency relative to the benchmark estimator. When compared directly with the regression-type estimator, the correlation-type estimator exhibits a modest increase in efficiency, although the difference is generally small. However, the efficiency of the ratio-type estimator was substantially better than that of the regression-type estimator, since the relationship linking the variable of interest and the auxiliary variable passes through the origin.

Although the bias and MSE are derived using large-sample approximations, the simulation results show that the proposed estimators remain stable and highly efficient even for small initial sample sizes and strong spatial clustering. This confirms that the asymptotic theory provides a reliable description of their finite-sample performance under the GIACS framework, which is especially important for rare and clustered populations.

5. Conclusions

This study proposed two new estimators for population mean estimation in rare and clustered populations under the general inverse adaptive cluster sampling (GIACS) framework: a ratio-type estimator and a correlation-type estimator. The auxiliary variable provided crucial information which was used in the development of our proposed estimators. For both estimators, the bias and the mean squared error were derived, providing a theoretical foundation for their performance in populations considered rare and clustered.

Our findings reveal that estimators utilizing auxiliary information achieve substantially improved performance relative to estimators that do not incorporate such information. In particular, both proposed estimators exhibit estimated absolute relative bias that decreases systematically as the initial sample size and the predetermined number of rare units increase, approaching zero in practical sampling scenarios. Across a wide range of realistic designs, the proposed estimators consistently yield lower mean squared errors and higher relative efficiencies than the conventional regression-type estimator [26]. Among them, the ratio-type estimator shows the largest efficiency gains, reflecting the strong linear association—passing through the origin—between the study variable and the auxiliary variable in the considered populations. In this study, the proposed ratio-type and correlation-type estimators were constructed using a single auxiliary variable under the assumption that the study variable and the auxiliary variable exhibit a strong linear association. At this stage, issues related to robustness—such as performance under strongly nonlinear relationships or in the presence of outliers—have not yet been examined and are therefore left for future investigation. Future research should focus on developing nonlinear or transformation-based extensions of the proposed estimators to better accommodate nonlinear relationships between the study variable and the auxiliary variables. In addition, this approach may be further advanced through the development of estimators that incorporate additional types of auxiliary information, including robust measures such as the median or M-estimators, as well as the use of multiple auxiliary variables combined with dimension-reduction techniques. These extensions may enhance estimator efficiency under the GIACS structure.

Author Contributions: Conceptualization, N.C. and P.G.; methodology and software, S.W. and N.C.; validation, M.C. and C.B.; formal analysis, P.G. and N.C.; writing—original draft preparation, P.G. and M.C.; writing—review and editing, N.C., C.B. and S.W.; supervision, N.C. and P.G.; project administration, N.C.; funding acquisition, N.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research project was financially supported by Mahasarakham University (grant number 6904001).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The details of the derivations used to obtain the bias and mean squared error for the proposed estimators presented in Section 3 are shown as follows, offering a theoretical basis underpinning the performance of the estimators and their associated properties.

The bias and mean square error of the ratio-type estimator in GIACS are as follows:

To prove bias of $\overline{y}_{GIACS,R}$, let $e_x = \frac{\overline{x}_{GIACS}-\mu_x}{\mu_x}$ and $e_y = \frac{\overline{y}_{GIACS}-\mu_y}{\mu_y}$. Then the ratio-type estimator can be written as $\overline{y}_{GIACS,R} = \mu_y\left(1+e_y\right)\left(1+e_x\right)^{-1}$. Assuming $|e_x| < 1$ and using the binomial expansion of $\left(1+e_x\right)^{-1}$.

$\overline{y}_{GIACS,R}$ can be written as $\overline{y}_{GIACS,R} = \mu_y(1+e_y)(1-e_x+e_x^2+O(e_x))$, where $O(e_x)$ denote the higher order terms of e_x and $O(e_x)$ closed to zero.

Therefore, $\overline{y}_{GIACS,R} \approx \mu_y(1+e_y-e_x+e_x^2-e_ye_x)$, $E\left(\overline{y}_{GIACS,R}\right)-\mu_y = \mu_y[E(e_y)-E(e_x)+E(e_x^2)-E(e_ye_x)]$, where $E(e_y)=0$ and $E(e_x)=0$, $Bias\left(\overline{y}_{GIACS,R}\right) = E\left(\overline{y}_{GIACS,R}\right)-\mu_y = \mu_y\left[E(e_x^2)-E(e_ye_x)\right]$, $E(e_x^2) = \frac{1}{\mu_x}E(\overline{x}_{GIACS}-\mu_x)^2 = \frac{1}{\mu_x}V(\overline{x}_{GIACS})$ and $E(e_ye_x) = \frac{1}{\mu_y\mu_x}E\left[(\overline{y}_{GIACS}-\mu_y)(\overline{x}_{GIACS}-\mu_x)\right] = \frac{1}{\mu_y\mu_x}COV(\overline{x}_{GIACS},\overline{y}_{GIACS})$.

Then $Bias\left(\overline{y}_{GIACS,R}\right) = \frac{\mu_y}{\mu_x^2}V(\overline{x}_{GIACS}) - \frac{1}{\mu_x}COV(\overline{y}_{GIACS},\overline{x}_{GIACS})$.

As the sample size increase, $V(\overline{x}_{GIACS})$ and $COV(\overline{y}_{GIACS},\overline{x}_{GIACS})$ are converge to zero, implying that $Bias\left(\overline{y}_{GIACS,R}\right)$ is converge to zero.

To prove $MSE\left(\overline{y}_{GIACS,R}\right)$, $\overline{y}_{GIACS,R} = \frac{\overline{y}_{GIACS}}{\overline{x}_{GIACS}}\mu_x = \hat{R}^*\mu_x$ can be rewritten as $\overline{y}_{GIACS,R} = \overline{y}_{GIACS} + \hat{R}^*(\mu_x-\overline{x}_{GIACS})$. For a large sample size n_T, we assume $\hat{R}^* \approx R^*$, so $\overline{y}_{GIACS,R} = (\overline{y}_{GIACS} - R^*\overline{x}_{GIACS}) + R^*\mu_x$.

The bias will converge to zero with rising sample size, indicating that the proposed estimator has a mean squared error which is asymptotically equivalent to its variance.

To prove $MSE\left(\overline{y}_{GIACS,R}\right)$ for the Condition 2 ($C2$):

The variance of Murthy's estimator is used as given in [24,27], $\overline{y}_{GIACS,R}$ has a mean square error of:

$$MSE\left(\overline{y}_{GIACS,R}\right)_{C_2} \approx V\left(\overline{y}_{GIACS,R}\right)_{C_2} = V[E(\overline{\alpha}^*|s)] = V(\overline{\alpha}^*) - E[V(\overline{\alpha}^*|s)],$$

where $V(\overline{\alpha}^*) = \frac{1}{N^2}\sum_{i=1}^{N} V\left(\frac{\alpha_i^*}{p_i}I_i\right) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i\neq j} COV\left(\frac{\alpha_i^*}{p_i}I_i, \frac{\alpha_i^*}{p_i}I_i\right)$. I_i serves as an indicator function which is equal to 1 upon the selection of unit i as the first unit and is equal to 0 in other cases. For $I_iI_j = 0$, $i \neq j$; $E(I_iI_j) = 0$, and $COV(I_iI_j) = -p_ip_j$, so that $V(\overline{\alpha}^*) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j<i}\left(\frac{\alpha_i^*}{p_i} - \frac{\alpha_j^*}{p_j}\right)^2 p_ip_j$. $V(\overline{\alpha}^*|s) = \frac{1}{N^2}V\left(\sum_{i=1}^{N}\frac{\alpha_i^*}{p_i}I_i|s\right) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{\alpha_i^{*2}}{p_i^2}\left[\Pr(I_i=1|s) - \Pr(I_i=1|s)^2\right] - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}\frac{\alpha_i^*\alpha_j^*}{p_ip_j}\Pr(I_i=1|s)\Pr(I_j=1|s) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j<i}\frac{\Pr(s|i)P(s|j)}{\Pr(s)^2}\left(\frac{\alpha_i^*}{p_i} - \frac{\alpha_j^*}{p_j}\right)p_ip_j$, which gives $E[V(\overline{\alpha}^*|s)] = \frac{1}{N^2}\sum_s \Pr(s)\sum_{i=1}^{N}\sum_{j<i}\frac{\Pr(s|i)\Pr(s|j)}{\Pr(s)^2}\left(\frac{\alpha_i^*}{p_i} - \frac{\alpha_j^*}{p_j}\right)p_ip_j = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j<i}\frac{\Pr(s|i)\Pr(s|j)}{\Pr(s)}\left(\frac{\alpha_i^*}{p_i} - \frac{\alpha_j^*}{p_j}\right)p_ip_j$.

From [27], the general equation of $V\left(\overline{y}_{GIACS,R}\right)_{C2}$ is:

$$V\left(\overline{y}_{GIACS,R}\right)_{C2} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j<i}\left[1 - \sum_{s\ni i,j}\frac{\Pr(s|i)\Pr(s|j)}{\Pr(s)}\right]\left(\frac{\alpha_i^*}{p_i} - \frac{\alpha_j^*}{p_j}\right)^2 p_ip_j$$

The estimator of the mean square error of $\overline{y}_{GIACS,R}$ can be considered under three separate conditions as follows:

$$M\hat{S}E\left(\overline{y}_{GIACS,R}\right) = \begin{cases} \frac{(N-n_I)}{N}\frac{s_{\alpha I}^{*2}}{n_I} & if\ C1, \\ \hat{V}\left(\overline{y}_{GIACS,R}\right)_{C_2} & if\ C2, \\ \frac{(N-n_F)}{N}\frac{s_{\alpha F}^{*2}}{n_F} & if\ C3, \end{cases}$$

where C1: $s_{\alpha I}^{*2} = (n_I-1)^{-1}\sum_{i=1}^{n_I}(\alpha_i^* - \overline{\alpha}_I^*)^2$, $\overline{\alpha}_I^* = n_I^{-1}\sum_{i=1}^{n_I}\alpha_i^*$ and
C3: $s_{\alpha F}^{*2} = (n_F-1)^{-1}\sum_{i=1}^{n_F}(\alpha_i^* - \overline{\alpha}_F^*)^2$, $\overline{\alpha}_F^* = n_F^{-1}\sum_{i=1}^{n_F}\alpha_i^*$.

$$C2 \;:\; \hat{V}\left(\overline{y}_{GIACS,R}\right)_{C_2} = \frac{1}{N^2}\sum_{i=1}^{n_T}\sum_{j<i}^{n_T}\left(\frac{\Pr(s|i,j)}{\Pr(s)} - \frac{\Pr(s|i)\Pr(s|j)}{\Pr(s)^2}\right)\left(\frac{\alpha_i^*}{p_i} - \frac{\alpha_j^*}{p_j}\right)^2 p_i p_j.$$

$\Pr(s|i,j)$ is the probability of getting the sample s, given that the units i and j were selected in the first two draws. From Equation (6) and the Appendix in [27], together with Equation (3) and Appendix A in [24], it follows that

$$\hat{V}\left(\overline{y}_{GIACS,R}\right)_{C_2} = \frac{N}{(n_T-1)^2(n_T-2)}\left[\frac{r(r-2)(N-n_T+1)-N(n_T-2)}{r^2}\sum_{i\in S_M}\sum_{j<i:j\in S_M}\left(\alpha_i^*-\alpha_j^*\right)^2\right.$$
$$\left.+(N-n_T+1)\left[\left(\left[\frac{(r-1)}{r}\sum_{i\in S_M}\sum_{j\in S_{M'}}\left(\alpha_i^*-\alpha_j^*\right)^2+\sum_{i\in S_{M'}}\sum_{j>i:j\in S_{M'}}\left(\alpha_i^*-\alpha_j^*\right)^2\right)\right]\right]$$

where $\sum_{i\in S_M}\sum_{j<i:j\in S_M}\left(\alpha_i^*-\alpha_j^*\right)^2 = \frac{1}{2}\sum_{i\in S_M}\sum_{j\in S_M}\left(\alpha_i^*-\overline{\alpha}_M^*+\overline{\alpha}_M^*-\alpha_j^*\right)^2 = r(r-1)s_{\alpha M}^{*2}$, $\sum_{i\in S_M}\sum_{j\in S_{M'}}\left(\alpha_i^*-\alpha_j^*\right)^2 = \sum_{i\in S_M}\sum_{j\in S_{M'}}\left(\alpha_i^*-\overline{\alpha}_M^*+\overline{\alpha}_M^*-\overline{\alpha}_{M'}^*-\alpha_j^*+\overline{\alpha}_{M'}^*\right)^2 = (n_T-r)(r-1)s_{\alpha M}^{*2}+$ $(n_T-r)r\left(\overline{\alpha}_M^*-\overline{\alpha}_{M'}^*\right)^2 + r(n_T-r-1)s_{\alpha M'}^{*2}$, and $\sum_{i\in S_{M'}}\sum_{j>i:j\in S_{M'}}\left(\alpha_i^*-\alpha_j^*\right)^2 = \frac{1}{2}\sum_{i\in S_{M'}}\sum_{j\in S_{M'}}$ $\left(\alpha_i^*-\overline{\alpha}_{M'}^*+\overline{\alpha}_{M'}^*-\alpha_j^*\right)^2 = (n_T-r)(n_T-r-1)s_{\alpha M'}^{*2}$.

Therefore, $\hat{V}\left(\overline{y}_{GIACS,R}\right)_{C_2}$ is:

$$\hat{V}\left(\overline{y}_{GIACS,R}\right)_{C_2} = \left(\frac{\hat{M}}{N}\right)^2\left(\frac{(N-n_T+1)(n_Tr-n_T-r)-N(n_T-2)}{N(n_T-2)(r-1)}\right)\frac{s_{\alpha M}^{*2}}{r}$$
$$+\frac{1}{N^2}\hat{V}(\hat{M})\left(\overline{\alpha}_M^*-\overline{\alpha}_{M'}^*\right)^2+\left(\frac{(N-n_T+1)(n_T-r-1)}{N(n_T-1)(n_T-2)}\right)s_{\alpha M'}^{*2}$$

The correlation-type estimator in GIACS has the bias and mean square error which can be shown below as:

To prove bias of $\overline{y}_{GIACS,\rho} = \overline{y}_{GIACS} + \hat{\rho}*(\mu_x - \overline{x}_{GIACS})$, where $\hat{\rho}^*$ is the estimator of the population correlation coefficient between $\overline{y}_i^*$ and $\overline{x}_i^*$, $\rho* = \frac{S_{x*y*}}{\sqrt{S_{x*}^2}\sqrt{S_{y*}^2}}$. $\hat{\rho}* = \frac{s_{x*y*}}{\sqrt{s_{x*}^2}\sqrt{s_{y*}^2}} = \frac{\hat{T}_{xy}^*-N\overline{y}_{GIACS}\overline{x}_{GIACS}}{\sqrt{\left(\hat{T}_{xx}^*-N\overline{x}_{GIACS}^2\right)}\sqrt{\left(\hat{T}_{yy}^*-N\overline{y}_{GIACS}^2\right)}}$.

Let $e_1 = \frac{s_{y*}^2-1}{S_{y*}^2}$, $e_2 = \frac{s_{x*}^2-1}{S_{x*}^2}$ and $e_3 = \frac{s_{x*y*}^2-1}{S_{x*y*}^2}$. Then the correlation-type estimator can be written as

$$\overline{y}_{GIACS,\rho} = \mu_y(1+e_y) - \frac{S_{x*y*}(1+e_3)}{\sqrt{S_{x*}^2(1+e_2)}\sqrt{S_{y*}^2(1+e_1)}}\mu_x e_x$$

$$\overline{y}_{GIACS,\rho} - \mu_y = \mu_y e_y - \frac{S_{x*y*}}{S_{x*}S_{y*}}\mu_x e_x(1+e_3)(1+e_2)^{-\frac{1}{2}}(1+e_1)^{-\frac{1}{2}}$$

Assuming $|e_1| < 1$ and $|e_2| < 1$, using the binomial expansion, the first order approximation is considered and ignore the higher order approximation. Hence,

$$\overline{y}_{GIACS,\rho} - \mu_y = \mu_y e_y - \rho*\mu_x e_x(1+e_3)\left(1-\frac{e_2}{2}+\frac{3}{8}e_2^2-\ldots\right)\left(1-\frac{e_1}{2}+\frac{3}{8}e_1^2-\ldots\right)$$

$$\overline{y}_{GIACS,\rho} - \mu_y = \mu_y e_y - \rho*\mu_x\left(e_x - \frac{e_xe_1}{2} - \frac{e_xe_2}{2} + e_xe_3\right)$$

$$Bias(\overline{y}_{GIACS,\rho}) = E\left(\overline{y}_{GIACS,\rho}-\mu_y\right) = \mu_yE(e_y) - \rho*\mu_x\left[E(e_x) - \frac{1}{2}E(e_xe_1) - \frac{1}{2}E(e_xe_2) + E(e_xe_3)\right]$$

$$Bias(\overline{y}_{GIACS,\rho}) = \rho*\mu_x\left[\frac{1}{2}E(e_xe_1)\frac{1}{2}E(e_xe_2) - E(e_xe_3)\right]$$

where $E(e_y) = 0$, $E(e_x) = 0$, $E(e_xe_1) = \frac{1}{\mu_x S_{y*}^2}COV(\overline{x}_{GIACS}, s_{y*}^2)$, $E(e_xe_2) = \frac{1}{\mu_x S_{x*}^2}COV(\overline{x}_{GIACS}, s_{x*}^2)$ and $E(e_xe_3) = \frac{1}{\mu_x S_{x*y*}}COV(\overline{x}_{GIACS}, S_{x*y*})$.

To prove $MSE\left(\overline{y}_{GIACS,\rho}\right)$, similarly to $\overline{y}_{GIACS,R}$, $\overline{y}_{GIACS,\rho} = \overline{y}_{GIACS} + \hat{\rho}^*(\mu_x - \overline{x}_{GIACS})$ can be rewritten as $\overline{y}_{GIACS,\rho} = (\overline{y}_{GIACS} - \hat{\rho}^*\overline{x}_{GIACS}) + \hat{\rho}^*\mu_x$. For a large sample size, we assume $\hat{\rho}^* \approx \rho^*$, and therefore let $\gamma_i^* = \overline{y}_i^* - \rho^*\overline{x}_i^* \approx \overline{y}_i^* - \hat{\rho}^*\overline{x}_i^*$. It follows that $\overline{y}_{GIACS,\rho} = \overline{\gamma}^* + \rho^*\mu_x$. We then prove $MSE\left(\overline{y}_{GIACS,\rho}\right)$ in the same way as $\overline{y}_{GIACS,R}$ by replacing α_i^* with γ_i^*.

References

1. Thompson, S.K. Adaptive Cluster Sampling. *J. Am. Stat. Assoc.* **1990**, *85*, 1050–1059. [CrossRef]
2. Olayiwola, O.M.; Ajayi, A.O.; Onifade, O.C.; Wale-Orojo, O.; Ajibade, B. Adaptive Cluster Sampling with Model Based Approach for Estimating Total Number of Hidden COVID-19 Carriers in Nigeria. *Stat. J. IAOS* **2020**, *36*, 103–109. [CrossRef]
3. Chandra, G.; Tiwari, N.; Nautiyal, R. Adaptive Cluster Sampling-Based Design for Estimating COVID-19 Cases with Random Samples. *Curr. Sci.* **2021**, *120*, 1202–1210. [CrossRef]
4. Stehlík, M.; Kiseľák, J.; Dinamarca, A.; Alvarado, E.; Plaza, F.; Medina, F.A.; Stehlíková, S.; Marek, J.; Venegas, B.; Gajdoš, A.; et al. REDACS: Regional Emergency-Driven Adaptive Cluster Sampling for Effective COVID-19 Management. *Stoch. Anal. Appl.* **2023**, *41*, 474–508. [CrossRef] [PubMed]
5. Hwang, J.; Bose, N.; Fan, S. AUV Adaptive Sampling Methods: A Review. *Appl. Sci.* **2019**, *9*, 3145. [CrossRef]
6. Giouroukis, D.; Dadiani, A.; Traub, J.; Zeuch, S.; Markl, V. A Survey of Adaptive Sampling and Filtering Algorithms for the Internet of Things. In *Proceedings of the 14th ACM International Conference on Distributed and Event-based Systems (DEBS '20)*; ACM: New York, NY, USA, 2020; pp. 27–38. [CrossRef]
7. Aga, G.; Francis, D.C.; Jolevski, F.; Rodriguez Meza, J.; Wimpey, J.S. An Application of Adaptive Cluster Sampling to Surveying Informal Businesses. *J. Surv. Stat. Methodol.* **2023**, *11*, 1246–1266. [CrossRef]
8. Aga, G.; Francis, D.C.; Wild, M. Surveying Informal Enterprises: Applying Stratified Adaptive Cluster Sampling Using CAPI with Implementation and Monitoring Tools. In Proceedings of the Workshop on Statistical Data Collection, Ottawa, ON, Canada, 10–12 October 2017; pp. 1–13. Available online: https://unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2017/mtg3/DC2017_1-3_Aga_Francis_and_Wild_World_Bank_AD.pdf (accessed on 10 December 2025).
9. Conners, M.; Schwager, S. The Use of Adaptive Cluster Sampling for Hydroacoustic Surveys. *ICES J. Mar. Sci.* **2002**, *59*, 1314–1325. [CrossRef]
10. Hanselman, D.H.; Quinn, T.J.; Lunsford, C.; Heifetz, J.; Clausen, D. Applications in Adaptive Cluster Sampling of Gulf of Alaska Rockfish. *Fish. Bull.* **2003**, *101*, 501–513.
11. Noon, B.R.; Ishwar, N.M.; Vasudevan, K. Efficiency of Adaptive Cluster and Random Sampling in Detecting Terrestrial Herpetofauna in a Tropical Rainforest. *Wildl. Soc. Bull.* **2006**, *34*, 59–68. [CrossRef]
12. Boomer, K.; Werner, C.; Brantley, S. CO_2 Emissions Related to the Yellowstone Volcanic System: 1. Developing a Stratified Adaptive Cluster Sampling Plan. *J. Geophys. Res.* **2000**, *105*, 10817–10830. [CrossRef]
13. Christman, M.C.; Pontius, J.S. Bootstrap Confidence Intervals for Adaptive Cluster Sampling. *Biometrics* **2000**, *56*, 503–510. [CrossRef]
14. Meeden, G.; Qureshi, M.N. Adaptive Cluster Sampling, a Quasi Bayesian Approach. *Surv. Methodol.* **2024**, *50*, 209–234.
15. Chao, C.T. Ratio Estimation On Adaptive Cluster Sampling. *J. Chin. Stat. Assoc.* **2004**, *42*, 307–327. [CrossRef]
16. Dryver, A.L.; Chao, C. Ratio Estimators in Adaptive Cluster Sampling. *Environmetrics* **2007**, *18*, 607–620. [CrossRef]
17. Chutiman, N. Adaptive Cluster Sampling Using Auxiliary Variable. *J. Math. Stat.* **2013**, *9*, 249–255. [CrossRef]
18. Yadav, S.K.; Misra, S.; Mishra, S.S.; Chutiman, N. Improved Ratio Estimators of Population Mean In Adaptive Cluster Sampling. *J. Stat. Appl. Probab. Lett.* **2016**, *3*, 1–6.
19. Chaudhry, M.S.; Hanif, M. Product and Exponential Product Estimators in Adaptive Cluster Sampling under Different Population Situations. *Proc. Pak. Acad. Sci. A Phys. Comput. Sci.* **2016**, *53*, 447–457.
20. Singh, R.; Mishra, R. Improved Exponential Ratio Estimators in Adaptive Cluster Sampling. *J. Stat. Appl. Probab. Lett.* **2022**, *9*, 19–29. [CrossRef]
21. Singh, R.; Mishra, R. Ratio-Cum-Product Type Estimators for Rare and Hidden Clustered Population. *Sankhya B* **2023**, *85*, 33–53. [CrossRef]
22. Mishra, R.; Singh, R.; Raghav, Y.S. On Combining Ratio and Product Type Estimators for Estimation of Finite Population Mean in Adaptive Cluster Sampling Design. *Braz. J. Biom.* **2024**, *42*, 412–420. [CrossRef]
23. Qureshi, M.N.; Kadilar, C.; Noor Ul Amin, M.; Hanif, M. Rare and Clustered Population Estimation Using the Adaptive Cluster Sampling with Some Robust Measures. *J. Stat. Comput. Simul.* **2018**, *88*, 2761–2774. [CrossRef]

24. Salehi, M.; Seber, G.A.F. A General Inverse Sampling Scheme and Its Application to Adaptive Cluster Sampling. *Aust. N. Z. J. Stat.* **2004**, *46*, 483–494. [CrossRef]
25. Christman, M.C.; Lan, F. Inverse Adaptive Cluster Sampling. *Biometrics* **2001**, *57*, 1096–1105. [CrossRef]
26. Chutiman, N.; Wichitchan, S.; Boonpok, C.; Chiangpradit, M.; Guayjarernpanishk, P. A Refined Regression Estimator for General Inverse Adaptive Cluster Sampling. *Mathematics* **2025**, *13*, 3751. [CrossRef]
27. Salehi, M.M.; Seber, G.A.F. Theory & Methods: A New Proof of Murthy's Estimator Which Applies to Sequential Sampling. *Aust. N. Z. J. Stat.* **2001**, *43*, 281–286. [CrossRef]
28. Hasan, M.Z.; Sultana, M.; Fatema, K.; Hossain, M.A.; Hossain, M.M. A New Regression Type Estimator and Its Application in Survey Sampling. *Open J. Stat.* **2020**, *10*, 1010–1019. [CrossRef]
29. Smith, D.R.; Conroy, M.J.; Brakhage, D.H. Efficiency of Adaptive Cluster Sampling for Estimating Density of Wintering Waterfowl. *Biometrics* **1995**, *51*, 777–788. [CrossRef]
30. Shahzad, U.; Ahmad, I.; Al-Noor, N.H.; Benedict, T.J. Use of Calibration Constraints and Linear Moments for Variance Estimation Under Stratified Adaptive Cluster Sampling. *Soft Comput.* **2022**, *26*, 11185–11196. [CrossRef]
31. Subzar, M.; Alqurashi, T.; Chandawat, D.; Tamboli, S.; Raja, T.A.; Attri, A.K.; Wani, S.A. Generalized Robust Regression Techniques and Adaptive Cluster Sampling for Efficient Estimation of Population Mean in Case of Rare and Clustered Populations. *Sci. Rep.* **2025**, *15*, 2069. [CrossRef]

https://doi.org/10.3390/sym18020334