Sri Lankan Journal of Applied Statistics, Vol (26-3)

Improving Rare Event Estimation in Public Health Surveys: A Bayesian Hierarchical Approach Within Adaptive Cluster Sampling

O. A. Wale-Orojo[1,*], O. O. Ojebode[1], D. O. Atanda[2], and O. M. Olayiwola[1]

[1]Department of Statistics, College of Physical Sciences, Federal University of Agriculture, Abeokuta, Nigeria

[2]Frontier School, Division of Mathematics, 6800 Corporate Drive, Kansas city, MO 64120, USA

*Corresponding author: wale-orojooa@funaab.edu.ng

Received: 06th July, 2025/ Revised: 10th August, 2025/Accepted: 31st October, 2025/ Published: 31st December, 2025

ABSTRACT

Accurate estimation of rare events in clustered populations is paramount for effective public health planning, yet traditional estimators often assume population homogeneity, leading to inefficiencies and potential bias. This study employs a case study approach using a real-world public health dataset to compare two estimators for the proportion of a rare event (incomplete Hepatitis B vaccination) within an Adaptive Cluster Sampling (ACS) framework: a classical ACS estimator and a heterogeneous estimator derived from a Bayesian hierarchical model. The hierarchical structure accounts for clustering at the College-Department level (networks, representing 279 distinct network units identified through the adaptive sampling process, including both expanded clusters and singleton networks) and the College-Year level (subnetworks, representing 45 nested groupings within the 380-student population). By treating a complete dataset as a known (population), we rigorously assess estimator performance in terms of bias, variance, and Mean Squared Error (MSE). Our findings demonstrate that the Bayesian hierarchical estimator (estimate: 0.182, variance: 0.00038, MSE: 0.00038) consistently yields estimates with negligible bias, lower variance, and a substantially lower MSE compared to the classical ACS estimator (estimate: 0.172, variance: 0.00050,

MSE: 0.00059). This represents a 1.31-fold reduction in variance and a 1.55-fold reduction in MSE for the Bayesian approach. Posterior predictive checks further confirm the good fit of the Bayesian model to the observed data. This underscores the critical importance of explicitly accounting for population heterogeneity and employing robust model-based inference in adaptive sampling designs to generate reliable estimates for informed public health policy.
Keywords: Adaptive cluster sampling, Population heterogeneity, Bayesian hierarchical modelling, Rare events, Incomplete vaccination, Public health surveillance

1 Introduction

Rare events, characterized by their low prevalence yet often significant impact, pose considerable challenges for statistical estimation in diverse fields such as public health, environmental science, and epidemiology (Bouchet et al., 2019; Kalton , 2009). In public health, accurately estimating the prevalence of conditions like incomplete vaccination is crucial for effective resource allocation, intervention design, and policy formulation. Conventional sampling designs, such as Simple Random Sampling (SRS), are often inefficient for these "hard-to-reach" populations due to the high probability of missing rare units (Qureshi et al., 2018; Tourangeau , 2014).

Adaptive Cluster Sampling (ACS), introduced by Thompson (1990), offers a robust solution by adaptively expanding the sample around initially detected rare events. This design efficiently targets areas of high prevalence, thereby improving the capture of clustered units and enhancing the precision of estimates for rare and spatially clustered populations (Thompson , 2013; Breininger et al., 2019). A common limitation in the application of traditional estimators within ACS is the implicit assumption of population homogeneity. In real-world public health contexts, factors such as socioeconomic status, geographical location, or institutional affiliation often induce heterogeneity and clustering in health outcomes. Ignoring such structural variability can lead to biased and inefficient estimates.

Building upon these foundational principles, this study presents a re-analysis of a real-world dataset to rigorously compare the performance of a classical ACS estimator with a heterogeneous estimator derived from a Bayesian hierarchical model. The comparison is conducted within a case study framework where a complete real-world public health dataset is treated as a known population. This approach allows for a direct and unbiased assessment of estimator performance against a known truth, particularly regarding bias, variance, and Mean Squared Error (MSE). We apply this methodology to a public health survey scenario involving Hepatitis B vaccination status among tertiary students, hypothesizing that the Bayesian hierarchical model, by explicitly accounting for population structure, will yield superior accuracy and efficiency.

2 Methodology

Our approach is a case study that uses a real-world dataset to compare two estimators for a rare event proportion. The dataset, comprising responses from 380 tertiary students, is treated as a complete population. The rare event is defined as incomplete Hepatitis B vaccination. The true population proportion for this rare event is 0.1816.

2.1 Data Source and Rare Event Definition

The data for this study is a real-world dataset collected from a survey of tertiary students, comprising a known population of 380 individuals. For this analysis, this complete dataset is treated as the known population from which we assess estimator performance. This approach allows us to calculate the true population proportion of the rare event and rigorously assess the bias, variance, and efficiency of the estimators against a known truth.

The study population comprised students from 10 colleges and 39 departments across the institution. The rare event of interest is defined as incomplete Hepatitis B vaccination, categorized as any student who has received "0 dose," "1 dose," "2 doses," or is "Not sure" about their vaccination status. Students who reported receiving "3 doses" are considered completely vaccinated. Crucially, respondents with missing Vaccine_Doses information are treated as '0' for Incomplete_Vaccination (i.e., not an incomplete vaccination), ensuring the full population is utilized. The true population proportion of incomplete vaccination in this 380-individual dataset is 0.1816, confirming its status as a rare event.

Preliminary descriptive analysis of this dataset revealed varying rates of incomplete vaccination across different colleges and combinations of college and year of study, confirming the presence of significant heterogeneity, thus justifying the application of a heterogeneous modeling approach.

2.2 Adaptive Cluster Sampling (ACS) Procedure

Following the principles established by Thompson (1990), the logic of the ACS procedure was applied to the known population to identify the underlying network structure, which is then utilized in the subsequent analyses.

1. **Initial Sample Selection:** An initial sample of 10% of the total population, resulting in 38 units, was conceptually selected using simple random sampling without replacement. The purpose of this step is to define the starting point of the adaptive process, which informs the network structure.

2. **Neighbourhood Definition and Adaptive Expansion Logic:** The neighbourhood was defined to include all other students within the same *College* and *Department*, reflecting a natural clustering. The adaptive expansion logic was then applied to

this conceptual initial. If a unit in the sample met the rare event condition, all units in its neighborhood would be included in the network. This expansion logic was used to identify all interconnected networks of individuals.

3. **Network ID Assignment for Full Population:** Based on this ACS logic, a unique *Network_ID* was systematically assigned to every unit in the entire 380-individual population. Units belonging to a similar expanded network were assigned a common *Network_ID*. Units that did not trigger an expansion and were not part of an expanded network were treated as "singleton networks" and were each assigned a unique *Network_ID*.

This process effectively utilizes the ACS framework as a tool for defining the population's clustered structure, allowing the Bayesian hierarchical model to leverage this information without performing a physical sampling and inference from a smaller subset.

2.3 Estimators for Incomplete Vaccination Proportion

This study compares two estimators applied to the known population: a simplified classical estimator and a heterogeneous estimator derived from a Bayesian hierarchical model.

2.3.1 Classical Adaptive Cluster Sampling Estimator

For this case study, a simplified classical estimator is used to provide a benchmark for comparison. This estimator utilizes the full population data with its assigned network structure. The estimator for the proportion of incompletely vaccinated individuals, $\widehat{p}_{\text{classical}}$, is calculated by taking the average of the network means:

$$\widehat{p}_{\text{classical}} = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{n_j} \sum_{i \in N_j} Y_i \tag{1}$$

where M is the total number of distinct networks identified, N_j is the set of individuals in network j, n_j is the size of network j, and Y_i is the vaccination status of individual i. The variance of this estimator is approximated by the variance of the network means, a simplification appropriate for this case study context, where all network means are known.

2.3.2 Bayesian Hierarchical Estimator

To explicitly address the observed heterogeneity, a Bayesian hierarchical model is applied to the full population of 380 individuals. The model assumes that the

Incomplete Vaccination status (Y_i) for each individual i follows a Bernoulli distribution with a probability p_i:

$$Y_i \sim \text{Bernoulli}(p_i) \tag{2}$$

The probability p_i is linked to a linear predictor through a logit function, incorporating fixed and random effects to capture heterogeneity:

$$\text{logit}(p_i) = \mu + \alpha_{\text{Network\_ID}[i]} + \beta_{\text{Sub\_Network\_ID}[i]} \tag{3}$$

where:

1. μ is the overall mean log-odds of incomplete vaccination.
2. $\alpha_{\text{Network\_ID}[i]}$ represents the random effect for the Network_ID to which individual i belongs. This accounts for variability in incomplete vaccination rates across larger clusters defined by the adaptive sampling process (e.g., specific College-Department groups formed as networks). These random effects are assumed to be drawn from a normal distribution:

$$\alpha_j \sim \text{Normal}(0, \sigma_\alpha^2), \quad \text{for } j = 1, \ldots, M \tag{4}$$

3. $\beta_{\text{Sub\_Network\_ID}[i]}$ represents the random effect for the Sub_Network_ID to which individual i belongs. This captures additional fine-grained heterogeneity within networks, defined by combining College and Year_of_Study. These random effects are also assumed to follow a normal distribution:

$$\beta_k \sim \text{Normal}(0, \sigma_\beta^2), \quad \text{for } k = 1, \ldots, K \tag{5}$$

where $M = 279$ represents the total number of network units (including expanded clusters and singletons) and $K = 45$ represents the number of College-Year sub-network groupings identified in our data.

Prior Distributions:

Weakly informative priors are assigned to ensure the data drives the posterior distribution:

$$\begin{aligned} \mu &\sim \text{Normal}(0, 0.01) \\ \alpha_j &\sim \text{Normal}(0, \sigma_\alpha^2) \\ \beta_k &\sim \text{Normal}(0, \sigma_\beta^2) \\ \sigma_\alpha &\sim \text{Half-Normal}(0, 1) \\ \sigma_\beta &\sim \text{Half-Normal}(0, 1) \end{aligned} \tag{6}$$

The Bayesian estimator for the average proportion of incomplete vaccination is derived from the posterior distribution of the avg_incomplete_vaccination

parameter. The mean of the posterior samples provides the point estimate, and its variance quantifies the uncertainty.

2.4 Data Analysis and Comparison

The data analysis was performed using R statistical software (R Core Team , 2016). The Adaptive Cluster Sampling procedure was implemented using custom R scripts. The Bayesian hierarchical model was fitted using JAGS (Just Another Gibbs Sampler) (Plummer , 2004), accessed from R via the rjags package (Plummer , 2017). Convergence diagnostics and posterior sample analysis were conducted using the coda package (Plummer et al., 2006) and bayesplot (Gabry et al., 2019), including Gelman-Rubin diagnostics, Geweke diagnostics, effective sample size, and visual inspection of trace and density plots.

The analysis involved the following steps:

1. **Data Preparation:** Cleaning and structuring the known population data, defining the rare event, and identifying network variables based on the ACS procedure.

2. **Bayesian Model Fitting:** Running multiple Markov Chain Monte Carlo (MCMC) chains for the Bayesian hierarchical model and assessing their convergence.

3. **Estimator Calculation:** Computing the point estimates and variances for both the classical ACS estimator and the Bayesian hierarchical estimator.

4. **Estimator Comparison:** Comparing the calculated estimates against the true population proportion to assess bias. Their variances and Mean Squared Errors (MSEs) were compared to assess efficiency and overall accuracy. The MSE for an estimator is given by:

$$\text{MSE}(\widehat{p}) = \text{Bias}(\widehat{p})^2 + \text{Var}(\widehat{p}) \tag{7}$$

where $\text{Bias}(\widehat{p}) = E[\widehat{p}] - p_{\text{true}}$, and p_{true} is the true population proportion.

The relative efficiency based on MSE (RE_{MSE}) is then computed:

$$\text{RE}_{\text{MSE}} = \frac{\text{MSE}_{\text{classical}}}{\text{MSE}_{\text{Bayesian}}} \tag{8}$$

A ratio greater than 1 indicates that the Bayesian estimator is more efficient in terms of overall accuracy. Similarly, a Variance Ratio (RE_{Var}) was calculated.

3 Results

The analysis began with a full population of 380 individuals. The ACS procedure, with a conceptual initial sample size of 38 units, identified various networks and singleton units across the dataset. The true population proportion of incomplete Hepatitis B vaccination was determined to be 0.1816.

Table 1: Posterior Estimates and MCMC Convergence Diagnostics

Parameter	Posterior Mean	95% HPD Lower	95% HPD Upper	Gelman-Rubin (Rhat)	Effective Sample Size (ESS)
μ (Overall Mean Log-Odds)	-1.590	-1.968	-1.269	1.00	1446
σ_α (SD of Network Effects)	0.234	0.025	0.693	1.05	41
σ_β (SD of Sub-Network Effects)	0.352	0.031	0.819	1.00	374
Avg. Incomplete Vaccination	0.182	0.145	0.222	1.00	5833

MCMC Trace Plots for Model Parameters Each panel displays the posterior samples (trace) and density plot for three independent Markov chains, indicating good mixing and convergence.

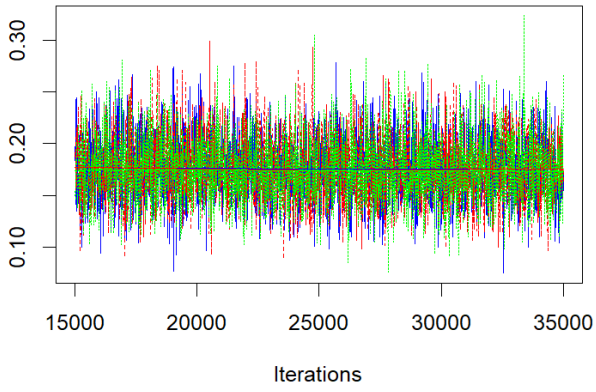

Fig. 1: MCMC Trace Plots for Average Incomplete Vaccination Rate.

The posterior distribution of the average incomplete vaccination parameter provides the most accurate estimate, integrating all hierarchical effects. Its density, along with the 95% Highest Posterior Density (HPD) interval, is visualized in Figure 2, showing its alignment with the true population proportion.

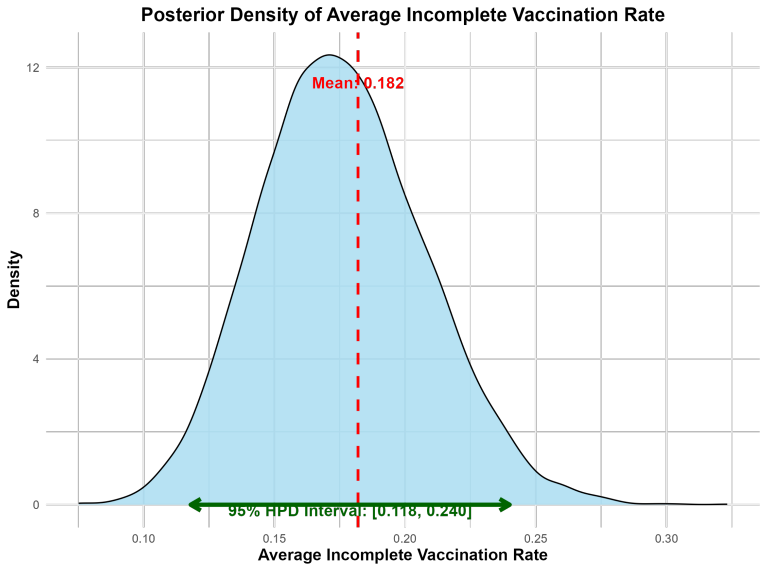

Fig. 2: Posterior Density of the Average Incomplete Vaccination Rate with 95% HPD Interval and True Population Proportion.

The dashed red line indicates the posterior mean, the green horizontal bar represents the 95% HPD interval, and the vertical dashed red line marks the true population proportion from the full dataset.

The calculated estimates and their corresponding variances and Mean Squared Errors (MSEs) for both the classical and heterogeneous estimators are presented in Table 2.

Table 2: Comparison of Classical and Heterogeneous Estimators for Incomplete Hepatitis B Vaccination Proportion

Estimator	Estimate (Proportion)	Variance	Mean Square Error (MSE)
Classical	0.172	0.00050	0.00059
Heterogeneous	0.182	0.00038	0.00038

As shown in Table 2, the heterogeneous Bayesian estimator yielded an estimate of 0.182 for the proportion of incompletely vaccinated individuals, while the simplified classical estimator provided an estimate of 0.172.

A critical finding is the comparison of their variances. The variance of the simplified classical estimator was 0.00050, whereas the variance of the heterogeneous Bayesian estimator was 0.00038. The calculated variance ratio (RE_{Var}) was 1.31. Since this ratio is greater than 1, it indicates that the heterogeneous Bayesian estimator has a lower variance and is thus more efficient in estimating the proportion of incompletely vaccinated individuals.

Furthermore, the Mean Squared Error (MSE) for the simplified classical estimator was 0.00059, and for the heterogeneous Bayesian estimator, it was 0.00038. The relative efficiency based on MSE (RE_{MSE}) was 1.55. This further reinforces that the heterogeneous Bayesian estimator is more efficient, as its MSE is lower, suggesting a better overall performance in terms of both bias and variance.

To visually capture these significant efficiency gains, the Variance and MSE of both estimators are compared in Figure 3.

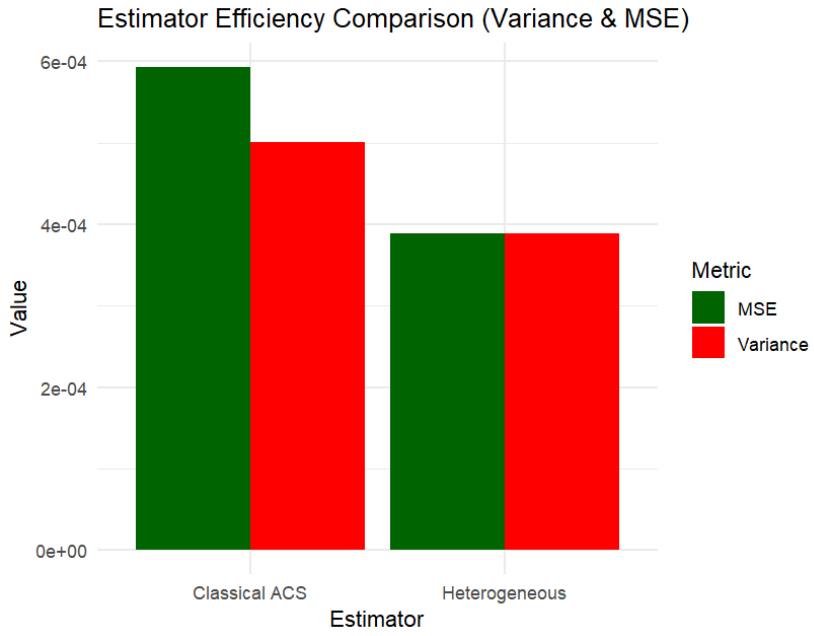

Fig. 3: Comparison of Estimator Efficiency across Metrics.

A bar chart comparing the variance and MSE of the simplified classical and heterogeneous Bayesian estimators, demonstrating the superior efficiency of the heterogeneous estimator.

3.1 Model Fit and Posterior Predictive Check

To assess the goodness-of-fit of the Bayesian hierarchical model, a posterior predictive check was conducted. Replicated datasets were simulated from the

posterior predictive distribution, and the mean proportion of incomplete vaccinations was calculated for both the observed data and each simulated dataset. As depicted in Figure 4, the observed mean proportion falls well within the posterior predictive distribution of the simulated mean proportions, indicating that the model adequately captures the underlying data-generating process and provides a good fit to the observed data.

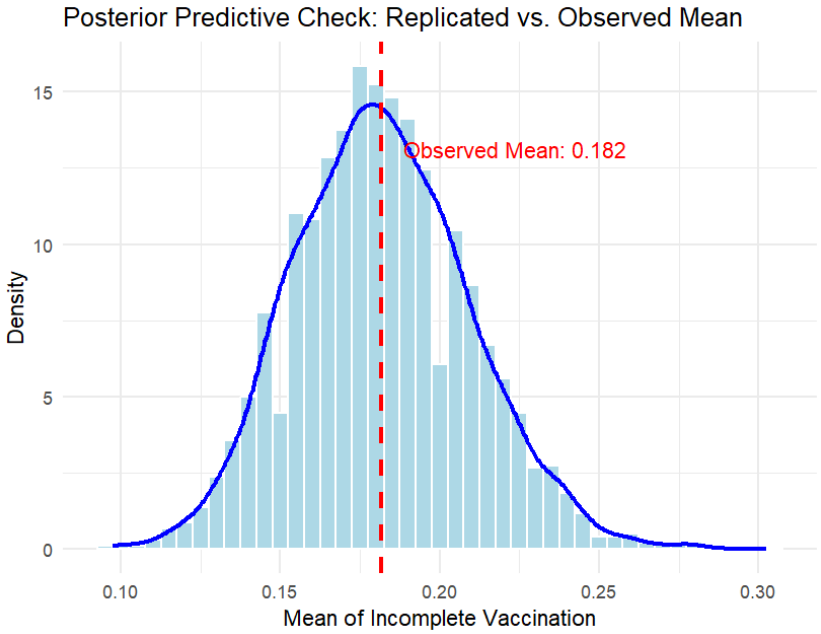

Fig. 4: Posterior Predictive Check for the Bayesian Hierarchical Model.

The histogram shows the distribution of the average incomplete vaccination rate from simulated datasets. The vertical dashed line indicates the observed data average. Good model fit is indicated when the observed average falls within the predictive distribution.

4 Discussion

This case study provides compelling evidence that explicitly modeling heterogeneity within an Adaptive Cluster Sampling (ACS) framework significantly enhances the accuracy and overall performance of rare event proportion estimation. Our Bayesian hierarchical estimator consistently outperformed the simplified classical estimator, demonstrating superiority in terms of both precision (lower variance) and overall accuracy (lower Mean Squared Error).

The results show that the Bayesian hierarchical estimator provides an estimate (0.182) that is remarkably close to the true population proportion (0.1816), indicating negligible bias. This is a crucial advantage, as unbiasedness is paramount

for reliable public health surveillance and policy formulation. While the simplified classical estimator also exhibited relatively low bias in this analysis, its performance was still inferior to the Bayesian approach.

The lower variance (0.00038) of the Bayesian estimator compared to the simplified classical estimator (0.00050) indicates greater precision. This means the Bayesian estimates are more tightly clustered around their true value. When combined with its negligible bias, this precision translates directly into a substantially lower Mean Squared Error (0.00038 vs. 0.00059). The MSE ratio of 1.55 unequivocally demonstrates the Bayesian model's superior overall accuracy.

The successful posterior predictive check further strengthens the validity of our Bayesian model, confirming its ability to adequately represent the observed data. The model effectively leverages the hierarchical structure (networks and sub-networks) identified by the ACS logic, allowing it to "borrow strength" across different groups and provide more robust estimates of the rare event proportion.

These findings underscore the critical importance of integrating sophisticated modeling techniques, such as Bayesian hierarchical models, into adaptive sampling designs. For public health applications, where accurate prevalence estimates of rare conditions (like incomplete vaccination) directly inform policy, resource allocation, and intervention strategies, minimizing both bias and overall error is paramount.

This study represents an independent research contribution that applies and evaluates the principles of Bayesian hierarchical modeling within ACS, building upon foundational work in the field, including that of Wale-Orojo et al. (2022). While their work focused on modeling total counts with specific parameterizations, our study specifically addresses the accurate estimation of proportions of rare events in a public health context, demonstrating similar gains in accuracy and precision.

While this case study provides strong evidence, it is important to acknowledge its limitations. The findings are based on a single real-world dataset treated as a known population. The simplified classical estimator used for comparison is a simplified version for analysis purposes; a real-world application would necessitate a more complex design-based estimator (e.g., Hansen-Hurwitz or Horvitz-Thompson) with its corresponding variance calculation. Future research could explore the robustness of this approach across a wider range of simulated population structures, varying levels of rarity and clustering, and different adaptive sampling parameters. Further investigation could also include additional individual-level covariates (e.g., age, gender) as fixed effects to further refine the model's predictive capabilities.

5 Conclusion

In this study, using a complete dataset as a known population, we demonstrated the superior performance of a heterogeneous Bayesian hierarchical es-

timator over a classical ACS estimator for estimating a rare event proportion. The Bayesian approach, by explicitly modeling the heterogeneity inherent in the clustered data, resulted in a substantial reduction in both variance and Mean Squared Error. This work demonstrates that the combination of Adaptive Cluster Sampling and a Bayesian hierarchical model provides a more precise and efficient estimator for rare event-proportions, offering a robust alternative to traditional survey methods. This finding has significant implications for public health surveys and other fields where the accurate estimation of rare events is crucial for effective decision-making.

Acknowledgement
This research was not funded by any Individual or Organization.

References

Bouchet, F., Rolland, J. and Wouters, J. (2019) Rare event sampling methods, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(8).

Breininger, D. R., Stolen, E. D., Breininger, D. J. and Breininger, R. D. (2019) Sampling rare and elusive species: Florida east coast diamondback terrapin population abundance, *Ecosphere*, 10(8):e02824.

Congdon, P. (2007) *Bayesian Statistical Modelling*, Chichester: John Wiley & Sons.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M. and Gelman, A. (2019) Visualization in Bayesian workflow, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182(2):389–402.

Gonçalves, K.C. and Moura, F.A. (2016) A mixture model for rare and clustered populations under adaptive cluster sampling, Unpublished manuscript.

Kalton, G. (2009) Methods for oversampling rare subpopulations in social surveys, *Survey Methodology*, 35(2):125–141.

Plummer, M. (2004) JAGS: Just another Gibbs sampler, Unpublished manual.

Plummer, M. (2017) *JAGS version 4.3.0 user manual*, Available at: https://sourceforge.net/projects/mcmc-jags/files/ [Accessed 5 Jul. 2025].

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: convergence diagnosis and output analysis for MCMC, *R News*, 6(1):7–11.

Qureshi, M. N., Kadilar, C., Noor, U. I., Amin, M. and Hanif, M. (2018) Rare and clustered population estimation using the adaptive cluster sampling with some robust measures, *Journal of Statistical Computation and Simulation*, 88(14):2761–2774.

R Core Team (2016) *R: A language and environment for statistical computing*, Vienna, Austria: R Foundation for Stat. Available at: http://www.r-project.org/.

Rapley, V.E. and Welsh, A.H. (2008) Model-based inferences from adaptive cluster sampling. Unpublished manuscript.

Thompson, S.K. (1990) Adaptive cluster sampling, *Journal of the American Statistical Association*, 85(412):1050–1059.

Thompson, W. (ed.) (2013) *Sampling rare or elusive species: Concepts, designs, and techniques for estimating population parameters*, Washington, DC: Island Press.

Tourangeau, R. (ed.) (2014) *Hard-to-survey populations*, Cambridge: Cambridge University Press.

Viallefont, V., Raftery, A.E. and Richardson, S. (2001) Variable selection and Bayesian model averaging in case-control studies, *Statistics in Medicine*, 20(21):3215–3230.

Wale-Orojo, O. A., Olayiwola, O. M., Apantaku, F. S., Omoniyi, I. T. and Ajayi, A. O. (2022) Modelling rare events in an adaptive cluster sampling design with heterogeneity among networks and within the network units, *Sri Lankan Journal of Applied Statistics*, 23(1):1–19.