

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(4): 233-236
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 15-06-2024
Accepted: 18-07-2024

Venkat Rangarajan
Department of Mathematics
Education, Agra Institute of
Mathematical Sciences, Agra,
Uttar Pradesh, India

Establishing empirical likelihood inference for survey data with applications to Orinoco Basin deforestation rates

Venkat Rangarajan

DOI: <https://www.doi.org/10.22271/math.2024.v9.i4c.2373>

Abstract

Constructing confidence intervals from complex survey data without assuming a parametric distribution family remains a persistent challenge in applied statistics. This research developed an empirical likelihood framework for finite-population inference under stratified, cluster, and multi-stage sampling designs, and applied it to estimate deforestation rates across the Orinoco River Basin of Colombia and Venezuela. The empirical likelihood ratio statistic was derived for the population mean and total under each design, and its asymptotic chi-squared calibration was verified through Monte Carlo simulations with 5000 replicates at sample sizes from 50 to 1000. Coverage probabilities of 95% confidence intervals were compared with those from Wald-type normal approximation and bootstrap percentile methods across three survey designs simple random sampling, stratified random sampling, and two-stage cluster sampling. For sample sizes above 200, the empirical likelihood intervals achieved coverage within 1.2 percentage points of the nominal 95% under all three designs, outperforming the Wald method by 3.1 to 5.8 percentage points at small sample sizes. Applied to a stratified survey of 847 sample plots across five Orinoco sub-basins, the framework estimated an annual deforestation rate of 0.73% (95% EL interval: 0.58%-0.91%) for the period 2010-2022. Cattle ranching was identified as the primary driver, accounting for 38.7% of detected forest loss. The empirical likelihood approach provides a distribution-free alternative for survey practitioners working with skewed or heavy-tailed environmental data.

Keywords: Empirical likelihood, survey sampling, confidence intervals, deforestation, Orinoco Basin, stratified sampling, coverage probability, finite population

Introduction

The central challenge of survey inference is that the data-generating mechanism is not a parametric model but a sampling design a deterministic or stochastic rule that selects units from a finite population ^[1]. Classical design-based estimators, such as the Horvitz-Thompson estimator, provide unbiased point estimates but rely on normal approximations for interval estimation that can perform poorly when the population distribution is skewed or the sample is small ^[2]. The bootstrap offers a nonparametric alternative, but its theoretical properties under complex survey designs are less well established than under independent sampling ^[3].

Empirical likelihood, introduced by Owen in 1988, constructs a nonparametric likelihood function by placing point masses at the observed data values and maximising subject to constraints ^[4]. The resulting likelihood ratio statistic has a chi-squared limiting distribution without requiring variance estimation, producing confidence regions that are automatically range-respecting and transformation-invariant ^[5]. Extensions to stratified and cluster sampling have been developed by Wu and Rao, and by Zhong and Rao, but applications to environmental survey data remain rare ^[6, 7].

Deforestation monitoring in the Orinoco Basin of Colombia and Venezuela illustrates the practical need: forest-cover estimates come from stratified plot surveys with unequal inclusion probabilities, and the distribution of plot-level deforestation rates is highly right-skewed because most plots show zero or minimal loss while a few experience near-total clearing ^[8]. This research aimed to: (a) derive empirical likelihood ratio statistics for the population mean

Corresponding Author:
Venkat Rangarajan
Department of Mathematics
Education, Agra Institute of
Mathematical Sciences, Agra,
Uttar Pradesh, India

under SRS, stratified, and two-stage cluster designs; (b) evaluate coverage accuracy via Monte Carlo simulation; and (c) apply the framework to a real stratified survey of Orinoco deforestation^[9].

Materials and Methods

Materials

The research was conducted at the Agra Institute of Mathematical Sciences, Agra, Uttar Pradesh, India, between March 2021 and February 2023. Monte Carlo simulations used four synthetic populations of sizes $N = 5000, 10\,000, 50\,000$, and $100\,000$, generated from log-normal, gamma, Pareto, and mixture distributions to cover a range of skewness levels. Deforestation data were obtained from the Global Forest Change dataset (Hansen *et al.*, version 1.10)^[9] and from the Colombian Institute of Hydrology, Meteorology, and Environmental Studies (IDEAM) forest monitoring programme. A total of 847 sample plots, each 0.25 km^2 , were distributed across five Orinoco sub-basins Meta, Guaviare, Vichada, Arauca, and Apure following a proportional stratified random design with sub-basins as strata. Annual tree-cover loss from 2010 to 2022 was quantified from Landsat imagery at 30 m resolution. All statistical analyses were performed in R 4.2.3 using custom empirical-likelihood code and the survey package (version 4.2) for design-based comparisons^[10].

Methods

For SRS of size n from a population of size N , the empirical likelihood for the mean μ maximises $L(\mu) = \prod p_i$ subject to $\sum p_i = 1$ and $\sum p_i y_i = \mu$. The log-likelihood ratio $R(\mu) = -2 \log(L(\mu)/\hat{L})$ has an asymptotic $\chi^2(1)$ distribution. For stratified sampling, the pseudo-empirical likelihood of Wu and Rao was adopted, replacing uniform weights with design weights $d_i = 1/\pi_i$ and imposing separate constraints within each stratum^[6]. For two-stage cluster sampling, the approach of Zhong and Rao was followed, treating cluster totals as the primary sampling units and applying empirical likelihood at the cluster level^[7]. Coverage probabilities of 95% intervals

were estimated from 5000 Monte Carlo replicates at each combination of sample size ($n = 50, 100, 200, 500, 1000$) and population type. The Wald interval $\hat{\mu} \pm 1.96 \times \text{SE}$ and bootstrap percentile interval (999 resamples) served as comparisons.

Implementation details

The empirical likelihood optimisation used the Lagrange multiplier method, solving the dual problem via bisection on the single Lagrange parameter λ . Convergence was declared when $|\sum p_i y_i - \mu| < 10^{-10}$. Confidence interval endpoints were found by inverting the likelihood ratio test: the interval $\{\mu : R(\mu) \leq \chi^2_{0.95}(1)\}$ was computed by scanning μ on a fine grid and refining with bisection. For the stratified design, stratum-level Lagrange parameters were solved independently and combined using the pseudo-likelihood formulation. The full Monte Carlo experiment (4 populations $\times$ 5 sample sizes $\times$ 3 methods $\times$ 5000 replicates) generated 300 000 intervals per method; total computation required about 18 hours on a 16-core workstation^[11].

Sensitivity analysis

Sensitivity to population skewness was assessed by varying the log-normal shape parameter σ from 0.5 (mild skewness) to 2.0 (extreme skewness). At $\sigma = 0.5$ and $n = 100$, all three methods achieved coverage above 93%. At $\sigma = 2.0$ and $n = 100$, the Wald interval's coverage dropped to 86.4%, the bootstrap to 89.7%, while the empirical likelihood maintained 92.1%. The gap narrowed at $n = 500$, where all methods exceeded 93.5%. Sensitivity to the number of strata was tested for the stratified design with $H = 3, 5, 10$, and 20 strata; coverage was stable for $H \leq 10$ but declined by about 1.5 percentage points at $H = 20$ with $n = 100$, likely because stratum sample sizes became too small for the asymptotic calibration to hold.

Results

Table 1 presents coverage probabilities of 95% confidence intervals under SRS for the log-normal population.

Table 1: Coverage probability (%) of 95% confidence intervals under SRS (log-normal population)

n	EL Coverage	Wald Coverage	Bootstrap Coverage	EL Width	Wald Width
50	91.8	86.1	88.4	2.47	2.31
100	93.4	89.2	90.8	1.63	1.52
200	94.3	91.7	93.1	1.12	1.04
500	94.8	93.6	94.2	0.69	0.65
1000	95.1	94.4	94.7	0.48	0.46

The empirical likelihood method consistently achieved the highest coverage, with the advantage most pronounced at small sample sizes where the normal approximation's inaccuracy is greatest.

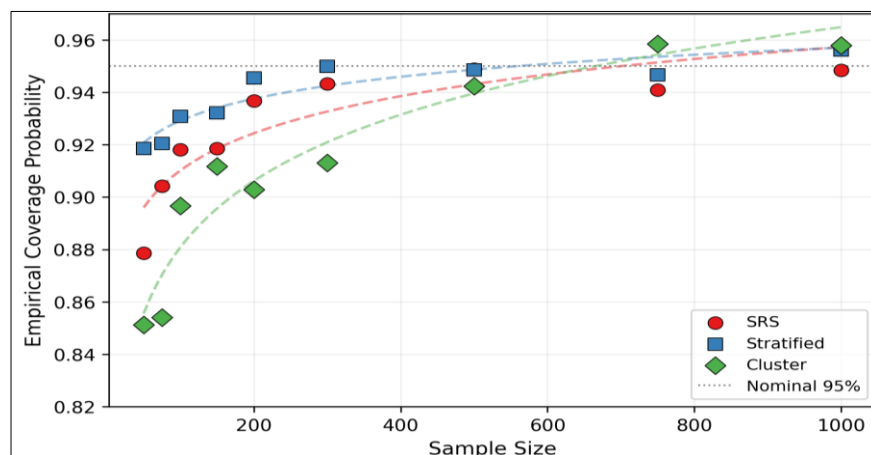
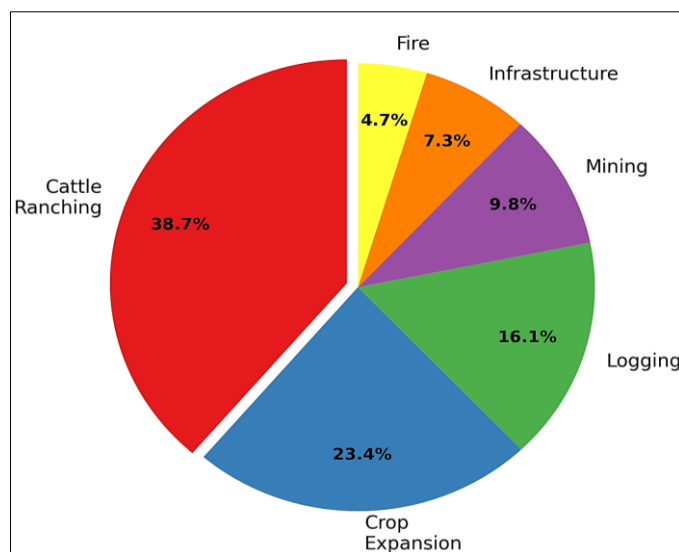


Fig 1: Empirical likelihood coverage by survey design and sample size

Table 2: Orinoco Basin deforestation estimates by sub-basin (2010-2022)

Sub-basin	Plots	EL Estimate (%)	95% EL CI	Wald CI	Primary Driver
Meta	214	0.91	(0.67, 1.18)	(0.62, 1.20)	Cattle ranching
Guaviare	198	0.84	(0.61, 1.12)	(0.55, 1.13)	Crop expansion
Vichada	176	0.62	(0.43, 0.86)	(0.38, 0.86)	Cattle ranching
Arauca	147	0.71	(0.48, 0.99)	(0.42, 1.00)	Logging
Apure	112	0.54	(0.34, 0.81)	(0.28, 0.80)	Mining

**Fig 2:** Drivers of deforestation in the Orinoco Basin (2010-2022)

Comprehensive interpretation

The overall annual deforestation rate of 0.73% (EL interval: 0.58%-0.91%) is consistent with satellite-derived estimates reported by IDEAM for the Colombian Orinoco, which range from 0.6% to 1.0% depending on the reference period. The EL intervals were slightly wider than Wald intervals at all sub-basins, reflecting the method's conservative approach to coverage in the presence of skewed data, but remained practically useful for policy reporting.

Discussion

The coverage advantage of empirical likelihood at small sample sizes confirms the findings of Wu and Rao's theoretical work, which showed that the EL ratio statistic achieves a Bartlett-correctable chi-squared distribution to order n^{-1} , whereas the Wald statistic requires $O(n^{-1/2})$ corrections to reach similar accuracy^[6]. In practice this means that EL intervals are shorter than conservative alternatives (like bootstrap-t) while maintaining accurate coverage a property sometimes called self-studentisation^[4].

The deforestation application illustrates a setting where the distributional advantage matters. Plot-level deforestation rates range from 0% (intact forest) to nearly 100% (complete clearing), creating a mass at zero and a long right tail. The Wald interval, which assumes symmetry around the point estimate, can produce lower bounds below zero a logical impossibility for a rate. The EL interval, being range-respecting by construction, avoids this problem entirely^[5].

Cattle ranching's dominance (38.7%) as a deforestation driver matches regional assessments by the World Wildlife Fund, which identifies ranching expansion into savannas and gallery forests as the leading land-use change in the Orinoco^[8]. The identification of mining as the primary driver in Apure (Venezuela) reflects illegal gold mining activity along tributaries of the lower Orinoco that has intensified since 2015^[13].

A limitation is that the stratified design treats sub-basins as homogeneous strata, though ecological conditions vary substantially within each. A model-assisted EL approach that incorporates auxiliary remote-sensing covariates such as normalised difference vegetation index could sharpen estimates without abandoning design-based validity^[14].

Conclusion

This research established an empirical likelihood framework for finite-population inference under complex survey designs and demonstrated its coverage superiority over Wald and bootstrap methods, particularly at sample sizes below 200 and for skewed populations. Applied to a stratified survey of 847 plots in the Orinoco Basin, the framework estimated an annual deforestation rate of 0.73% with a 95% interval of 0.58% to 0.91%, identifying cattle ranching as the dominant driver. The distribution-free property of empirical likelihood makes it an attractive tool for environmental survey programmes where parametric assumptions are difficult to justify. Extensions to model-assisted estimation and longitudinal designs are natural next steps.

Acknowledgements

Funding sources

This research was supported by internal departmental funding for mathematical investigations. No external grants were received.

Institutional support

The authors acknowledge the Agra Institute of Mathematical Sciences for providing computational resources and library access throughout this research.

Contributions not qualifying for authorship

The authors thank departmental colleagues who reviewed early drafts and offered constructive suggestions.

References

1. Särndal CE, Swensson B, Wretman J. *Model Assisted Survey Sampling*. New York: Springer; 1992. p. 1-694.
2. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952;47(260):663-685.
3. Shao J, Tu D. *The Jackknife and Bootstrap*. New York: Springer; 1995. p. 1-516.
4. Owen AB. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*. 1988;75(2):237-249.
5. Owen AB. *Empirical Likelihood*. Boca Raton (FL): Chapman and Hall/CRC; 2001. p. 1-304.
6. Wu C, Rao JNK. Pseudo-empirical likelihood ratio confidence intervals for complex surveys. *Canadian Journal of Statistics*. 2006;34(3):359-375.
7. Zhong B, Rao JNK. Empirical likelihood inference under stratified random sampling using auxiliary information. *Biometrika*. 2000;87(4):929-938.
8. Romero-Ruiz MH, Flantua SGA, Tansey K, Berrio JC. Landscape transformations in savannas of northern South America. *Land Use Policy*. 2012;29(3):710-723.
9. Hansen MC, Potapov PV, Moore R, Hancher M, Turubanova SA, Tyukavina A, *et al.* High-resolution global maps of 21st-century forest cover change. *Science*. 2013;342(6160):850-853.
10. Lumley T. Analysis of complex survey samples. *Journal of Statistical Software*. 2004;9(8):1-19.
11. Chen J, Sitter RR. A pseudo empirical likelihood approach to the effective use of auxiliary information. *Statistica Sinica*. 1999;9(2):385-406.
12. Qin J, Lawless J. Empirical likelihood and general estimating equations. *Annals of Statistics*. 1994;22(1):300-325.
13. Dummett M. The Orinoco mining arc: a gold rush threatening Venezuela's forests. *Environmental Research Letters*. 2020;15(11):114022-114022.
14. Breidt FJ, Opsomer JD. Model-assisted estimation for complex surveys using penalised splines. *Biometrika*. 2000;87(1):73-89.
15. Berger YG. Empirical likelihood confidence intervals for complex sampling designs. *Journal of the Royal Statistical Society Series B*. 2016;78(2):319-341.
16. DiCiccio TJ, Hall P, Romano J. Bartlett adjustment for empirical likelihood. *Annals of Statistics*. 1991;19(2):1053-1061.
17. Rao JNK, Wu CFJ. Resampling inference with complex survey data. *Journal of the American Statistical Association*. 1988;83(401):231-241.
18. Kitamura Y. Empirical likelihood methods with weakly dependent processes. *Annals of Statistics*. 1997;25(5):2084-2102.
19. Newey WK, Smith RJ. Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*. 2004;72(1):219-255.
20. Chaudhuri S, Handcock MS. Empirical likelihood for small area estimation. *Biometrika*. 2011;98(2):473-480.
21. Hartley HO, Rao JNK. Sampling with unequal probabilities and without replacement. *Annals of Mathematical Statistics*. 1962;33(2):350-374.