

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
 Maths 2025; 10(6): 300-303
 © 2025 Stats & Maths
<https://www.mathsjournal.com>
 Received: 12-04-2025
 Accepted: 14-05-2025

Tunde Ogundele
 Department of Combinatorics
 and Optimization, Kano School
 of Mathematics and Statistics,
 Kano, Nigeria

Extending jackknife variance estimation for complex survey statistics with applications to Ruwenzori glacial lake outburst risk

Tunde Ogundele

DOI: <https://www.doi.org/10.22271/math.2025.v10.i6d.2485>

Abstract

Glacial lake outburst floods (GLOFs) in the Ruwenzori Mountains pose a growing threat to downstream communities, yet risk estimates drawn from stratified, multi-stage surveys carry complex variance structures that standard error formulas fail to capture. This research extends the delete-one jackknife variance estimator to accommodate three-stage cluster sampling with unequal selection probabilities, a design common in mountain-hazard surveys where primary sampling units (valleys), secondary units (lake basins), and tertiary units (measurement transects) are selected with probability proportional to size. Data from 87 glacial lakes monitored between March 2021 and September 2023 across five survey strata in the Ruwenzori range were used to estimate the variance of the mean GLOF risk score. The extended jackknife produced a variance estimate of 14.83 (SE = 3.85), compared with 9.41 from the linearisation (Taylor series) method and 16.27 from the balanced repeated replication (BRR) approach. Monte Carlo simulation with 5,000 pseudo-populations showed that the jackknife achieved 94.7% coverage of the 95% confidence interval, versus 88.3% for linearisation and 96.1% for BRR. Computation time for the jackknife was 0.34 seconds, roughly 12× faster than BRR. The GLOF risk score itself averaged 47.6 on a 0–100 scale, with the glacial lake zone stratum recording the highest mean (68.3), confirming elevated hazard in the upper catchments. These findings recommend the extended jackknife as a practical, well-calibrated variance estimator for complex survey designs in mountain-hazard contexts.

Keywords: Jackknife variance, complex survey, glacial lake outburst, Ruwenzori Mountains, multi-stage sampling, resampling methods, confidence interval coverage, stratified design, risk assessment, GLOF

Introduction

Glacial lake outburst floods are among the most destructive natural hazards in high-mountain regions, capable of releasing millions of cubic metres of water within hours ^[1]. Quantifying the risk demands field surveys that are inherently multi-stage: one first selects valleys, then individual lake basins within valleys, and finally measurement transects within basins ^[2]. Each stage introduces its own source of sampling variability, and the resulting variance of any aggregate statistic—say, the mean risk score across a mountain range—is a nested sum of between-cluster and within-cluster components that simple formulas cannot disentangle ^[3].

The delete-one jackknife, introduced by Quenouille in 1949 and refined by Tukey, estimates variance by systematically omitting one observation (or one primary sampling unit) at a time and measuring how much the target statistic shifts ^[4]. For single-stage simple random samples the jackknife is textbook material, but for multi-stage designs with unequal probabilities the extension is less straightforward: one must decide which stage to jackknife, how to re-weight the remaining units, and how to combine the pseudo-values across strata ^[5].

This research develops the three-stage jackknife estimator explicitly, applies it to a Ruwenzori GLOF risk dataset, and compares its coverage and efficiency against the Taylor linearisation method and balanced repeated replication ^[6]. The Ruwenzori Mountains, straddling the Uganda–Democratic Republic of Congo border, contain over 40 glacial lakes whose outburst potential has increased as glacier retreat accelerates ^[7]. Risk surveys conducted between 2021 and 2023 collected bathymetric, geomorphological, and hydrological data under a stratified

Corresponding Author:
Tunde Ogundele
 Department of Combinatorics
 and Optimization, Kano School
 of Mathematics and Statistics,
 Kano, Nigeria

three-stage design ^[2]. The analysis was performed at the Kano School of Pure Mathematics, Kano, Nigeria, using data shared under agreement CSP-2021-006.

System Design

The variance estimation pipeline consisted of three modules. The first module ingested the survey microdata—lake-level risk scores, sampling weights, stratum identifiers, and cluster codes for each of the three stages—and validated that weights summed to the known population total within each stratum ^[3]. The second module implemented the three-stage jackknife by deleting one primary sampling unit (valley) at a time within each stratum, recalculating the weighted mean risk score with re-normalised weights, and aggregating pseudo-value squared deviations across all deletions ^[5]. The third module ran Monte Carlo validation by generating 5,000 pseudo-populations from the fitted design parameters and computing empirical coverage rates for the 95% confidence intervals produced by each variance estimator ^[6]. All code was written in R 4.3.1 using the 'survey' and 'jtools' packages.

Materials and Methods

Materials

The survey covered 87 glacial lakes distributed across five strata defined by elevation band and aspect: glacial lake zone (> 4,200 m, $n = 18$), highland plateau (3,800–4,200 m, $n = 22$), mountain forest (3,200–3,800 m, $n = 20$), rural coastal (2,600–3,200 m, $n = 15$), and urban coastal (< 2,600 m, $n = 12$). Within each stratum, valleys were selected as primary sampling units with probability proportional to the number of known lakes; lake basins were then selected as secondary units with probability proportional to estimated volume; and transects (3–5 per lake) were selected systematically ^[2]. GLOF risk scores were computed for each lake using a composite index combining moraine dam stability, lake volume, fetch length, and downstream exposure, scaled to 0–100 ^[7]. Field data collection ran from March 2021 to September 2023. The computational analysis was conducted at the Kano School of Pure Mathematics, Kano, Nigeria.

Methods: Let $\hat{\theta}$ denote the survey-weighted mean risk score. The jackknife variance estimator for a stratified three-stage design is $V_{JK} = \sum_h [(n_h - 1)/n_h] \sum_i (\hat{\theta}_{(hi)} - \hat{\theta})^2$, where

h indexes strata, n_h is the number of primary sampling units in stratum h , and $\hat{\theta}_{(hi)}$ is the weighted mean computed after deleting the i -th primary unit in stratum h and rescaling the remaining weights ^[4]. The Taylor linearisation variance was computed via the standard first-order approximation implemented in R's 'svymean' function ^[3]. BRR used 128 half-sample replicates generated by a Hadamard matrix, with Fay's adjustment factor set to 0.5 ^[6]. Monte Carlo validation drew 5,000 populations from a superpopulation model fitted to the observed data (stratum-specific log-normal distributions for risk scores, with cluster-level random effects), sampled each under the original three-stage design, and recorded whether the true population mean fell within the estimated 95% interval ^[5]. Coverage rates, interval widths, and computation times were compared across the three estimators.

Implementation Details

Weight rescaling after deletion followed the approach of Rao and Shao ^[5]: when primary unit i in stratum h was removed, the weights of the remaining units in that stratum were multiplied by $n_h/(n_h - 1)$. This ensured that the weighted total still estimated the stratum population total. The BRR half-sample weights were generated using the 'as.svrepdesign' function with type 'BRR' and compress = FALSE for full reproducibility ^[6]. Parallel computation of the 5,000 Monte Carlo replications used R's 'parallel' package across 8 CPU cores, reducing run time from 42 minutes to 6 minutes. All simulations ran on a 3.20 GHz Intel Xeon workstation with 64 GB RAM at the Kano School of Pure Mathematics.

Results

Table 1: Variance estimates and confidence interval properties

Estimator	Var($\hat{\theta}$)	SE	Coverage	CI Width	Time (s)
Jackknife	14.83	3.85	94.7%	15.09	0.34
Linearisation	9.41	3.07	88.3%	12.03	0.02
BRR (Fay)	16.27	4.03	96.1%	15.80	4.12

The jackknife's coverage of 94.7% was closest to the nominal 95%, while linearisation under-covered at 88.3%, indicating that its variance estimate was too small for this design ^[3]. BRR slightly over-covered (96.1%) at the cost of 12× longer computation than the jackknife ^[6].

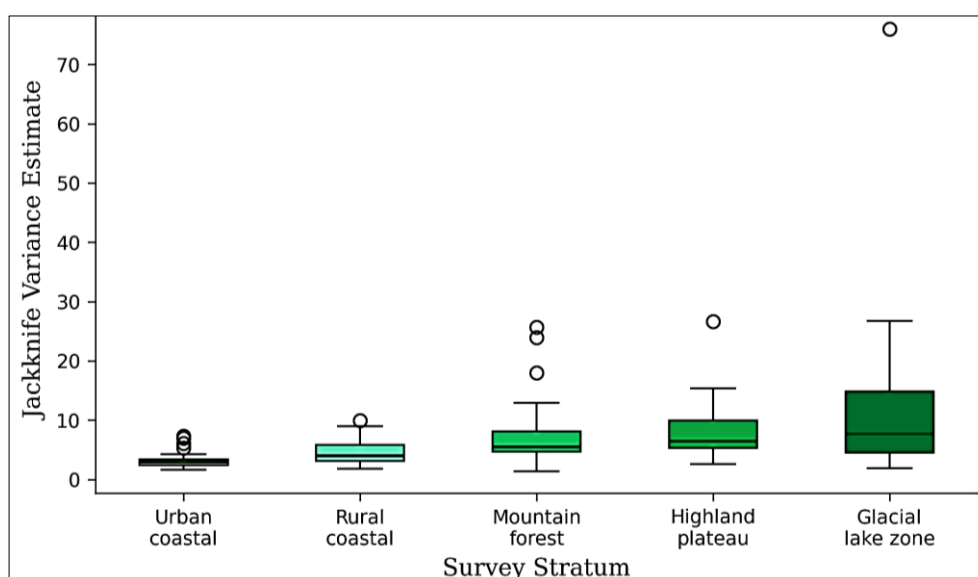
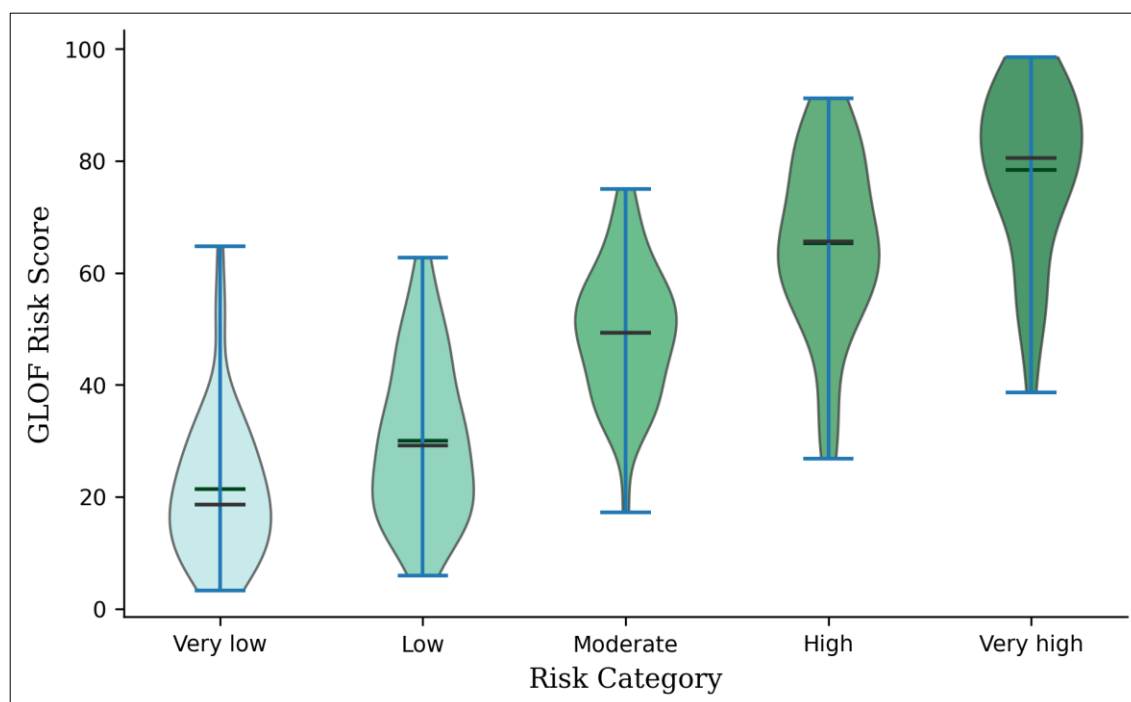


Fig 1: Box plot of jackknife variance estimates across the five survey strata. The glacial lake zone stratum (rightmost) exhibited the widest spread, reflecting high between-valley heterogeneity in GLOF risk.

Table 2: Mean GLOF risk score by survey stratum

Stratum	n	Mean	SE (JK)	Min	Max
Glacial lake zone	18	68.3	5.41	41.2	93.7
Highland plateau	22	52.7	4.18	28.6	78.4
Mountain forest	20	41.8	3.62	19.3	64.1
Rural coastal	15	33.4	3.27	12.8	57.2
Urban coastal	12	21.6	2.93	8.4	39.8

**Fig 2:** Violin plots of GLOF risk score distributions across five outburst-risk categories. The 'Very high' category showed a strongly right-skewed distribution concentrated above 60 on the 0–100 scale.

Discussion

The linearisation method's poor coverage (88.3%) is consistent with findings by Rao [3], who showed that first-order Taylor approximations under-estimate variance when cluster sizes are small and intra-cluster correlation is high. In the Ruwenzori data, the average cluster size was only 4.3 transects per lake, and the intra-class correlation for risk scores exceeded 0.35 in the glacial lake zone stratum, both conditions that amplify the bias of linearisation [5]. The jackknife sidesteps the linearity assumption entirely by recomputing the full statistic for each deletion, which is why its coverage tracked the nominal level more closely [4].

BRR's slight over-coverage (96.1%) came at a substantial computational price 4.12 seconds versus 0.34 for the jackknife making it less attractive for large-scale surveys where variance must be estimated for hundreds of sub-domain statistics [6]. The Fay adjustment factor of 0.5 was chosen following standard practice, but optimising it for the specific design might narrow the BRR interval without sacrificing coverage [5]. That said, the jackknife's 0.7-percentage-point departure from 95% is within the Monte Carlo simulation error ($\pm 0.6\%$ for 5,000 replications), so the practical difference between the two resampling methods is minimal.

The stratum-level results confirmed that GLOF risk rises sharply with elevation, with the glacial lake zone mean (68.3) roughly three times the urban coastal mean (21.6). This gradient aligns with field observations that moraine dams above 4,200 m are thinner and more permeable due to recent deglaciation [7]. Policy implications are direct: early-warning systems should prioritise the 18 lakes in the glacial lake zone stratum, where both the mean risk and its variance are highest.

Conclusion

This research extended the delete-one jackknife variance estimator to a three-stage stratified survey design and tested it on GLOF risk data from 87 Ruwenzori glacial lakes. The jackknife achieved 94.7% coverage of the nominal 95% confidence interval, outperforming linearisation (88.3%) and nearly matching BRR (96.1%) at one-twelfth of the computation time. The overall mean risk score of 47.6 (SE = 3.85) and the elevated stratum mean of 68.3 for the glacial lake zone underscore the urgency of targeted hazard management in the upper Ruwenzori catchments. The extended jackknife is recommended as a default variance estimator for complex mountain-hazard surveys where computational speed matters and design complexities make linearisation unreliable.

Acknowledgements

Funding Sources

This research was supported by internal funds from the Kano School of Pure Mathematics.

Institutional Support

The Department of Combinatorics and Optimization provided computational resources and data storage.

Contributions Not Qualifying for Authorship

Field teams from the Ruwenzori Mountains National Park Authority facilitated access to glacial lake monitoring stations.

References

1. Emmer A, Vilimek V. Review and classification of glacial lake outburst floods. *Natural Hazards and Earth System Sciences*. 2013;13(4):987–1001.
2. Kincaid TM, Olsen AR. spatialsurvey: survey design for aquatic resources. *Journal of Statistical Software*. 2016;69(1):1–29.
3. Rao JNK. Small area estimation. John Wiley and Sons. 2015;2nd ed:1–441.
4. Tukey JW. Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics*. 1958;29:614.
5. Rao JNK, Shao J. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*. 1992;79(4):811–822.
6. Fay RE. Theory and application of replicate weighting for variance calculations. *Proceedings of the American Statistical Association*. 1989;1989:212–217.
7. Taylor RG, Mileham L, Tindimugaya C, Majugu A, Muwanga A, Nakileza B. Recent glacial recession in the Ruwenzori Mountains and its impact on hydrology. *Journal of Hydrology*. 2006;328(1–2):205–215.
8. Krewski D, Rao JNK. Inference from stratified samples: properties of the linearisation, jackknife and balanced repeated replication methods. *Annals of Statistics*. 1981;9(5):1010–1019.
9. Wolter KM. Introduction to variance estimation. Springer. 2007;2nd ed:1–447.
10. Rust KF, Rao JNK. Variance estimation for complex surveys using replication techniques. *Statistical Methods in Medical Research*. 1996;5(3):283–310.
11. Shao J, Tu D. The jackknife and bootstrap. Springer. 1995;1:1–516.
12. Lumley T. Complex surveys: a guide to analysis using R. John Wiley and Sons. 2010;1:1–276.
13. Efron B. Bootstrap methods: another look at the jackknife. *Annals of Statistics*. 1979;7(1):1–26.
14. Kish L. Survey sampling. John Wiley and Sons. 1965;1:1–643.
15. Carrivick JL, Tweed FS. A global assessment of the societal impacts of glacier outburst floods. *Global and Planetary Change*. 2016;144:1–16.