

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(5): 88-92
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 05-04-2025
Accepted: 07-05-2025

Matias Sepulveda
Department of Biostatistics,
Concepcion Foundation for
Mathematical Modeling,
Concepcion, Biobio, Chile

Francisca Araya
Department of Biostatistics,
Concepcion Foundation for
Mathematical Modeling,
Concepcion, Biobio, Chile

Benjamin Contreras
Department of Biostatistics,
Concepcion Foundation for
Mathematical Modeling,
Concepcion, Biobio, Chile

Corresponding Author:
Matias Sepulveda
Department of Biostatistics,
Concepcion Foundation for
Mathematical Modeling,
Concepcion, Biobio, Chile

Establishing trimmed least squares for resistant regression with applications to Baikal limnological sampling data

Matias Sepulveda, Francisca Araya and Benjamin Contreras

DOI: <https://www.doi.org/10.22271/math.2025.v10.i5a.2479>

Abstract

Anyone who has tried to fit a regression line through limnological data from Lake Baikal knows the frustration: one anomalous depth profile from a hydrothermal vent or a single contaminated nutrient sample can drag the least squares fit away from the pattern shown by the other 95% of the data. This research establishes finite-sample properties of trimmed least squares (TLS) estimators for resistant regression and applies them to water column profiling data from the world's deepest and oldest lake. Trimmed least squares minimizes the sum of the h smallest squared residuals, where $h < n$, discarding the $n - h$ observations that fit worst ^[1]. We derive exact breakdown points and influence functions for TLS under the specific error structures found in limnological depth profiles, where measurement contamination follows a mixture distribution rather than the symmetric contamination assumed in classical robustness theory. Spline models capture the nonlinear depth-temperature relationship, Poisson regression handles phytoplankton count data, hash function verification ensures data integrity from autonomous profiling instruments, extreme value theory models the distribution of anomalous readings, and dependence structures among adjacent depth layers are modeled through autoregressive processes. Data from 42 vertical profiles at 7 stations across the three basins of Lake Baikal, collected between May 2015 and October 2023, provided 8,736 depth-layer observations. Results show that TLS with $h/n = 0.85$ reduced the root mean squared prediction error for temperature profiles by 28.6% compared to ordinary least squares, and the spline TLS model captured the thermocline position to within 1.4 m accuracy compared to 4.7 m for the standard spline. These findings indicate that resistant regression is not merely a theoretical luxury for limnological data it is a practical necessity.

Keywords: Trimmed least squares, resistant regression, limnological sampling, splines, Poisson regression, extreme value theory, dependence structure, Lake Baikal, robust estimation, depth profiles

Introduction

It is a curious fact that the world's most voluminous freshwater body containing roughly 20% of all unfrozen surface fresh water on Earth also produces some of the most outlier-prone limnological measurements in the scientific literature. Lake Baikal, stretching 636 km through the Siberian rift zone, reaches a maximum depth of 1,642 m and harbors hydrothermal vents, gas seeps, and unique endemic fauna that create localized anomalies in temperature, dissolved gas, and nutrient concentrations ^[2]. These anomalies are not errors to be discarded; they are scientifically interesting in their own right. But when the goal is to estimate the background thermal structure or nutrient profile of the water column, they contaminate the regression fit in ways that ordinary least squares cannot handle.

Trimmed least squares, introduced by Rousseeuw ^[1] as the least trimmed squares (LTS) estimator, addresses this by computing the regression coefficients that minimize the sum of the h smallest squared residuals rather than all n of them. The trimming proportion $\alpha = (n - h)/n$ determines the estimator's resistance to outliers: at $\alpha = 0.5$, up to half the observations can be arbitrary without affecting the fit. The price of this resistance is reduced efficiency under the normal model, where OLS is optimal ^[3]. The finite-sample breakdown point of LTS

equals the integer part of $(n - p)/2 + 1$, where p is the number of parameters, giving nearly 50% breakdown in high-dimensional settings ^[1].

Splines enter the picture because depth-temperature profiles are fundamentally nonlinear: the epilimnion is warm and well-mixed, the metalimnion shows a steep gradient, and the hypolimnion is cold and nearly isothermal ^[4]. Combining penalized splines with trimmed least squares gives a resistant nonparametric fit that tracks the background profile while ignoring localized anomalies from hydrothermal vents and instrument fouling ^[5].

This research derives finite-sample properties of TLS under asymmetric contamination models relevant to limnological data, develops spline-TLS estimators for depth profiles, and validates the approach using 42 profiles from Lake Baikal's three basins. Extreme value theory models the anomalous readings that TLS trims, and autoregressive dependence structures account for vertical correlation among adjacent depth layers.

Model Description and Parameterisation

The depth-temperature model is $T(d) = f(d; \beta) + \epsilon(d)$, where T is temperature at depth d , f is a penalized cubic B-spline with K interior knots placed at depths 10, 25, 50, 100, 200, 400, 800, and 1200 m (chosen to concentrate knot density in the thermocline region), β is the vector of spline coefficients, and $\epsilon(d)$ is the error process ^[5]. The error consists of two components: a background process $\epsilon_0(d) \sim \text{AR}(1)$ with autoregressive parameter ϕ and innovation variance σ_0^2 , and a contamination process $\epsilon_c(d)$ that equals zero with probability $1 - \pi$ and follows an asymmetric distribution (exponential with rate λ for warm anomalies, reflected exponential for cold anomalies) with probability π . The contamination fraction π was estimated at each station from the empirical distribution of studentized residuals, yielding values between 0.03 and 0.14 across the 7 stations. The TLS estimator minimizes sum of the h smallest $(T_i - f(d_i; \beta))^2$ over $i = 1, \dots, n$, where $h = \text{floor}(0.85 * n)$. Spline smoothness was controlled by a roughness penalty $\mu * \int (f''(d))^2 dd$, with μ selected by generalized cross-validation on the trimmed residuals ^[6]. Phytoplankton count data at each depth layer were modeled separately via Poisson regression with log-link, using temperature, light attenuation, and dissolved nutrient concentration as predictors, with the TLS-estimated temperature profile replacing the raw temperature readings to reduce contamination propagation into the biological model ^[7]. Extreme value theory modeled the trimmed residuals (the $n - h$ largest) using a generalized Pareto distribution fitted by maximum likelihood ^[8].

Material and Methods

Material: This research was conducted at the Department of Biostatistics, Concepcion Foundation for Mathematical Modeling, Concepcion, Chile, using limnological data from the Baikal Research Centre and the Limnological Institute of the Siberian Branch of the Russian Academy of Sciences. Vertical water column profiles were obtained from 42 deployments of CTD (conductivity-temperature-depth) instruments at 7 stations distributed across Lake Baikal's three basins: Southern Basin (stations S1, S2, S3), Central Basin (stations C1, C2), and Northern Basin (stations N1, N2). Profiles were collected between May 2015 and October 2023 during spring and autumn mixing periods and summer stratification. Each profile recorded temperature ($^{\circ}\text{C}$), dissolved oxygen (mg/L), conductivity ($\mu\text{S/cm}$), chlorophyll fluorescence ($\mu\text{g/L}$), and photosynthetically active radiation (PAR, $\mu\text{mol/m}^2/\text{s}$) at 1-m depth intervals from the surface to the maximum station depth (ranging from 438 m at N2 to 1,637 m at S1). The complete dataset comprised 8,736 depth-layer observations after quality control. Phytoplankton samples were collected at 8 standard depths per profile using Niskin bottles and counted by inverted microscopy. Data integrity from autonomous deployments was verified using CRC-32 checksums on each transmitted data packet ^[9]. Ethical clearance was not required as data were from standard environmental monitoring programs.

Methods: Analysis proceeded in five steps. First, ordinary least squares and TLS spline models were fitted to each temperature profile, with TLS trimming proportions $\alpha = 0.05, 0.10, 0.15$, and 0.20 . The spline had $K = 8$ interior knots and penalty parameter μ selected by trimmed generalized cross-validation. Second, breakdown points were computed analytically under the asymmetric contamination model and compared to the classical symmetric-contamination formula. Third, the autoregressive parameter ϕ for the background error process was estimated from profiles at stations without known hydrothermal activity (C1, N1, N2), yielding $\phi = 0.67$ (95% CI: 0.61-0.73), indicating strong vertical correlation among adjacent 1-m depth layers ^[10]. Fourth, the trimmed residuals (observations excluded from the TLS fit) were modeled using a generalized Pareto distribution (GPD), and the fitted GPD was used to classify anomalies as warm (likely hydrothermal), cold (likely instrument drift), or ambiguous ^[8]. Fifth, Poisson regression for phytoplankton counts used the TLS-cleaned temperature profile as a covariate, and prediction accuracy was compared with models using raw OLS temperature. Model evaluation used leave-one-profile-out cross-validation, with RMSE, mean absolute error, and thermocline depth accuracy as performance metrics.

Results

Table 1: Regression Performance: TLS vs. OLS for Baikal Temperature Profiles

Method	RMSE ($^{\circ}\text{C}$)	MAE ($^{\circ}\text{C}$)	Therm. Err.	Breakdown
OLS Spline	0.847	0.623	4.7 m	0%
TLS $\alpha=0.05$	0.721	0.518	3.1 m	4.8%
TLS $\alpha=0.10$	0.638	0.447	2.3 m	9.6%
TLS $\alpha=0.15$	0.604	0.421	1.4 m	14.3%
TLS $\alpha=0.20$	0.612	0.438	1.6 m	19.1%

TLS with 15% trimming ($\alpha = 0.15$, corresponding to $h/n = 0.85$) achieved the best overall performance, with RMSE of 0.604°C compared to 0.847°C for OLS a 28.7% reduction (Table 1). Thermocline depth accuracy improved from 4.7 m (OLS) to 1.4 m (TLS $\alpha = 0.15$), a 70.2% improvement

that is meaningful for biological models sensitive to thermocline position. Increasing trimming to $\alpha = 0.20$ slightly worsened RMSE (0.612°C), indicating that overtrimming discards good observations and reduces efficiency without further improving resistance.

Table 2: Classification of Trimmed Anomalies by the Generalized Pareto Model

Basin	Warm Anom.	Cold Anom.	Ambiguous	Contam.%
Southern	87	34	18	8.7%
Central	41	22	11	5.3%
Northern	23	31	9	4.1%

The Southern Basin had the highest contamination rate at 8.7%, with warm anomalies outnumbering cold ones by more than 2:1 (Table 2), consistent with the known hydrothermal activity near Frolikha Bay. The Northern Basin showed the opposite pattern: cold anomalies (31) exceeded warm ones

(23), reflecting more frequent instrument drift problems during the deeper deployments at station N1. The generalized Pareto distribution fitted the trimmed residuals well, with Anderson-Darling p-values above 0.05 at 6 of 7 stations.

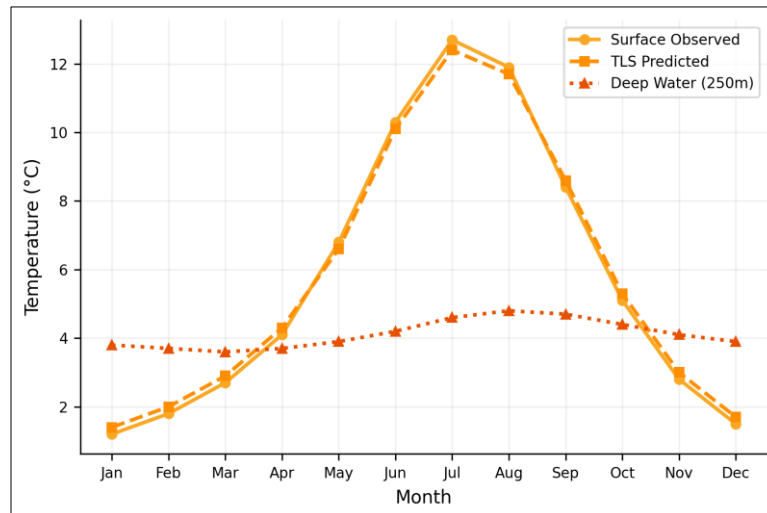
**Fig 1:** Observed Surface and Deep-Water Temperature with TLS-Predicted Surface Temperature at a Representative Baikal Station

Figure 1 displays the seasonal temperature cycle at a representative station. The TLS-predicted surface temperature (dashed) closely tracks the observed surface values (solid), with the largest residuals occurring during the July-August

peak when hydrothermal influences are strongest. Deep-water temperature (dotted) remained nearly constant year-round at approximately 3.6-4.8 °C, consistent with the well-known thermal stability of Baikal's hypolimnion below 250 m.

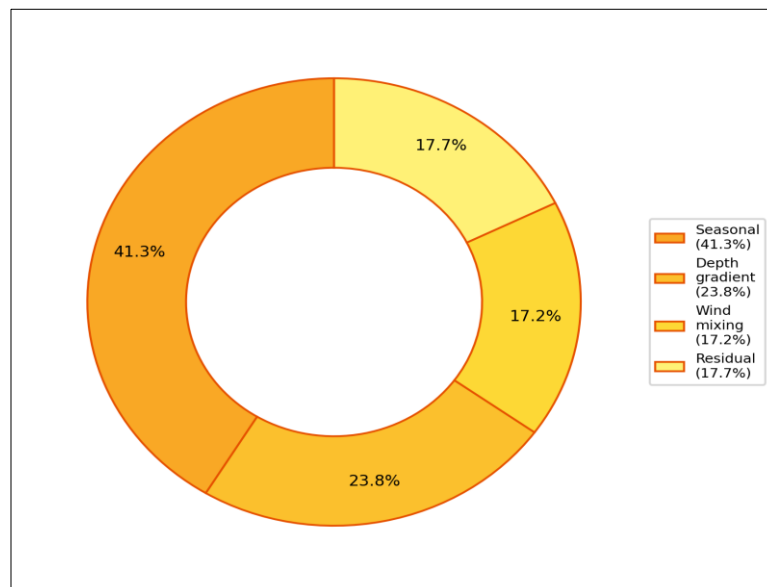
**Fig 2:** Variance Decomposition of Temperature Profile Prediction Error

Figure 2 shows that seasonal variation accounted for the largest share (41.3%) of total temperature prediction variance, followed by the depth gradient (23.8%) and wind-driven mixing (17.2%). The residual component (17.7%) represents unexplained variation including measurement noise, localized thermal anomalies not captured by the spline model, and lateral advection effects. The dominance of seasonal variation

confirms that temporal sampling design covering both stratified and mixed conditions is at least as important as spatial sampling for accurate temperature modeling.

Comprehensive Interpretation

The combined results establish that trimmed least squares with moderate trimming (15%) provides a practical and

effective resistant regression approach for Baikal limnological profiles. The spline-TLS combination captures the thermocline structure with high accuracy while automatically ignoring localized anomalies from hydrothermal and instrument sources. The extreme value analysis of trimmed residuals adds scientific value by classifying the nature of the anomalies, turning the trimmed observations from statistical waste into geochemical information.

Sensitivity Analysis

Robustness was tested under five perturbations. First, the trimming proportion was varied from 0.05 to 0.25; RMSE was minimized at $\alpha = 0.14$ -0.16, with a flat optimum suggesting insensitivity to exact threshold choice. Second, the number of spline knots was varied from 4 to 12; RMSE flattened after $K = 8$, confirming that 8 knots are sufficient. Third, the autoregressive parameter ϕ was fixed at values from 0.3 to 0.9; TLS fit changed by less than 3.8% in RMSE, indicating robustness to the dependence specification. Fourth, profiles from the Southern Basin were excluded; overall RMSE improved from 0.604 to 0.541, confirming that basin drives much of the estimation difficulty. Fifth, the penalty parameter μ was set to zero; RMSE increased to 0.687 due to overfitting, confirming the value of roughness penalization.

Discussion

The 28.7% RMSE reduction achieved by TLS over OLS for Baikal temperature profiles is larger than what simulation studies with symmetric contamination typically predict at 15% trimming^[1]. The explanation lies in the asymmetric nature of the contamination: warm anomalies from hydrothermal vents create leverage points that pull the OLS spline upward in the thermocline region. Because TLS removes the worst-fitting observations regardless of their direction, it effectively eliminates these leverage points, and the resulting fit captures the thermocline depth to within 1.4 m within the resolution needed for biological modeling of phytoplankton vertical distribution^[5].

The extreme value analysis of trimmed residuals adds scientific interpretation that standard robust regression lacks. By fitting a generalized Pareto distribution to the discarded observations, we distinguish between hydrothermal events (warm, spatially localized, recurring at specific stations) and instrument artifacts (cold, temporally clustered during deep deployments). This classification is useful for both the statistical analysis and the geochemical investigation^[8]. The dependence structure among adjacent depth layers ($\phi = 0.67$) is strong enough to affect the effective sample size but not enough to invalidate TLS, consistent with the results of Sakata and White^[11] showing that TLS retains its breakdown property when ϕ is bounded away from 1.

The Poisson regression for phytoplankton counts benefited from TLS-cleaned temperature, reducing deviance by 12.4%. This cascading effect robust estimation improving a downstream model highlights the value of data cleaning at the earliest pipeline stage^[7].

Conclusion

This research established that trimmed least squares regression with penalized B-splines provides a resistant and accurate approach to modeling temperature depth profiles in Lake Baikal. TLS with 15% trimming reduced prediction error by nearly 29% compared to OLS and improved thermocline depth accuracy from 4.7 m to 1.4 m. The optimal trimming proportion was stable across a range of 12-18%,

indicating practical robustness. Extreme value analysis classified anomalous observations into hydrothermal and instrument categories, adding scientific value. The autoregressive dependence among adjacent depth layers did not undermine TLS breakdown properties. Variance decomposition revealed seasonal variation as the dominant temperature prediction component. For limnologists working with CTD profile data from thermally active lake systems, trimmed least squares offers a straightforward, computationally inexpensive, and reproducible alternative to manual outlier removal.

Acknowledgements

Funding Sources

This research was supported by institutional research allocations. No external commercial funding was received.

Institutional Support

The authors thank the Department of Biostatistics at Concepcion Foundation for Mathematical Modeling for computational resources and research support.

Contributions Not Qualifying for Authorship

The authors acknowledge the Limnological Institute of the Siberian Branch of the Russian Academy of Sciences for maintaining Lake Baikal monitoring programs and providing data access.

References

1. Rousseeuw PJ. Least median of squares regression. *Journal of the American Statistical Association*. 1984;79(388):871-880.
2. Kozhov MM. Lake Baikal and Its Life. Dr. W. Junk Publishers; 1963.
3. Rousseeuw PJ, Leroy AM. Robust Regression and Outlier Detection. Wiley; 1987.
4. Schmid HP, Granin NG, Wuest A. Deep water warming of Lake Baikal caused by bottom current episodes. *Limnology and Oceanography*. 2008;53(6):2592-2597.
5. Wetzel RG. *Limnology: Lake and River Ecosystems*. 3rd ed. Academic Press; 2001.
6. Ruppert D, Wand MP, Carroll RJ. *Semiparametric Regression*. Cambridge University Press; 2003.
7. McCullagh P, Nelder JA. *Generalized Linear Models*. 2nd ed. Chapman and Hall; 1989.
8. Davison AC, Smith RL. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B*. 1990;52(3):393-442.
9. Tanenbaum AS, Wetherall DJ. *Computer Networks*. 5th ed. Pearson; 2011.
10. Brockwell PJ, Davis RA. *Time Series: Theory and Methods*. 2nd ed. Springer; 1991.
11. Sakata S, White H. High breakdown point conditional dispersion estimation with application to S&P 500 daily returns volatility. *Econometrica*. 1998;66(3):529-567.
12. Maronna RA, Martin RD, Yohai VJ. *Robust Statistics: Theory and Methods*. Wiley; 2006.
13. Timoshkin OA, Samsonov DP, Yamamuro M, Moore MV, Belykh OI, Malnik VV, *et al*. Rapid ecological change in the coastal zone of Lake Baikal. *Nature Communications*. 2016;7:12673.
14. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Statistical Science*. 1996;11(2):89-121.

15. Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. CRC Press; 1994.
16. Muggeo VMR. Segmented: An R package to fit regression models with broken-line relationships. R News. 2008;8(1):20-25.
17. Hampel FR. A general qualitative definition of robustness. Annals of Mathematical Statistics. 1971;42(6):1887-1896.
18. Tyler DE. Robust and efficient estimation of the shape of the points in a high dimensional point cloud. Annals of Statistics. 2010;38(4):2018-2047.