

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
 Maths 2024; 9(1): 198-201
 © 2024 Stats & Maths
<https://www.mathsjournal.com>
 Received: 15-11-2023
 Accepted: 18-12-2023

Adebayo Ogunleye
 Department of Mathematical
 Sciences, Lagos Center for
 Computational Sciences, Lagos,
 Nigeria

Exploring Bayesian estimation techniques for regression models with applications to gangs floodplain sediment transport

Adebayo Ogunleye

DOI: <https://www.doi.org/10.22271/math.2024.v9.i1c.2516>

Abstract

Sediment transport in large river floodplains is notoriously difficult to predict, in part because standard regression models struggle with the heteroscedastic, autocorrelated nature of the data. This research applied Bayesian estimation techniques to linear and nonlinear regression models of sediment load in the Ganges floodplain and compared their performance against frequentist maximum likelihood estimates. The work was carried out at the Lagos Center for Computational Sciences between February 2020 and October 2022. Monthly sediment concentration records from five Ganges gauging stations spanning 2008-2021 were modelled as functions of discharge, rainfall, and upstream land use. Bayesian posterior distributions were computed via Markov chain Monte Carlo (MCMC) with 50,000 iterations and a 10,000-sample burn-in. Autocorrelation diagnostics ensured chain mixing, and Galois-theoretic symmetry arguments simplified the prior specification for factorial covariate structures. M-estimators provided robust frequentist comparisons, and mutual information quantified nonlinear dependencies between predictors. The Bayesian approach reduced mean absolute prediction error by 21.3% relative to ordinary least squares and by 9.8% relative to M-estimation. Posterior credible intervals covered the observed values at the nominal 95% rate (94.7%), while frequentist confidence intervals showed undercoverage (89.2%).

Keywords: Autocorrelation, Galois theory, combinatorics, M-estimators, mutual information, Bayesian regression, sediment transport, Ganges floodplain, MCMC, posterior distribution

Introduction

Predicting how much sediment a river carries is not merely an academic exercise it affects dam design, irrigation planning, and the livelihoods of millions who depend on fertile floodplain soils. The Ganges, which drains roughly 1.08 million square kilometres across India, Nepal, and Bangladesh, transports an estimated 500-700 million tonnes of sediment annually ^[1]. Yet point predictions of sediment load at individual gauging stations can be off by 30-50% when standard least-squares regression is applied, largely because the residuals violate the independence and constant-variance assumptions ^[2].

Bayesian methods offer a natural way to handle these violations. By specifying prior distributions on regression coefficients and variance parameters, and updating them with observed data via Bayes' theorem, the approach produces full posterior distributions that reflect both data and prior knowledge ^[3]. MCMC algorithms, particularly the Gibbs sampler and Metropolis-Hastings, make posterior computation feasible even for complex models ^[4]. Still, Bayesian regression is underused in sediment transport research, partly because practitioners are unfamiliar with prior elicitation and convergence diagnostics ^[5].

This research filled that gap by applying Bayesian estimation to both linear and nonlinear regression models of Ganges floodplain sediment load. M-estimators from robust statistics provided a frequentist benchmark resistant to outliers ^[6]. Mutual information, a concept from information theory, was used to detect and quantify nonlinear dependencies among predictors that linear correlation would miss ^[7]. An unusual contribution was the use of Galois-theoretic

Corresponding Author:
 Adebayo Ogunleye
 Department of Mathematical
 Sciences, Lagos Center for
 Computational Sciences, Lagos,
 Nigeria

symmetry to simplify prior specification for factorial designs when covariates form symmetric groups under permutation, the prior can be specified on group orbits rather than individual parameters [8].

System design

The modelling pipeline comprised four stages. The data ingestion module standardised monthly records from the five Ganges stations into a common format, handling missing values via expectation-maximisation [9]. The predictor screening module computed pairwise mutual information and selected variables exceeding a threshold of 0.15 nats. The Bayesian regression module ran MCMC for each station-model combination, producing posterior samples for all parameters. The evaluation module compared Bayesian, OLS, and M-estimation predictions on a rolling 12-month hold-out window. The pipeline was implemented in Python 3.10 using PyMC 5.0 for MCMC and scikit-learn for M-estimation, running on a 64 GB workstation.

Materials and methods

Material

The research was conducted at the Department of Mathematical Sciences, Lagos Center for Computational Sciences, Lagos, Nigeria, between February 2020 and October 2022. Ethical clearance was granted (Protocol LCCS-2020-009, approved 20 January 2020). Monthly suspended sediment concentration (SSC) data for 2008-2021 were obtained from the Central Water Commission of India for five stations: Allahabad, Varanasi, Patna, Farakka, and Kolkata. Discharge records came from the same source. Rainfall data were drawn from the India Meteorological Department gridded product at 0.25° resolution. Land-use data (cropland fraction) were derived from MODIS MCD12Q1 annual products.

Methods

The linear model regressed log (SSC) on log (discharge), cumulative monthly rainfall, and upstream cropland fraction. The nonlinear model used a power-law relationship $SSC = \alpha Q^\beta \exp(\gamma R + \delta L + \epsilon)$, where Q is discharge, R is rainfall, and L is cropland fraction [2]. Bayesian priors were: $\alpha \sim \text{LogNormal}(5, 2)$, $\beta \sim \text{Normal}(1.5, 0.5)$, $\gamma, \delta \sim \text{Normal}(0, 1)$, $\sigma^2 \sim \text{InverseGamma}(2, 1)$. The Galois symmetry argument applied when the three stations in the middle reach (Varanasi, Patna, Farakka) shared exchangeable prior parameters under the symmetric group S_3 , reducing the effective number of free prior hyperparameters from 15 to 7 [8]. MCMC used the No-U-Turn Sampler (NUTS) with 50,000 iterations, 10,000 burn-in, and 4 chains [4]. Convergence was assessed via the Gelman-Rubin R-hat statistic (target < 1.01) and effective sample size (target > 1,000). Autocorrelation functions of the chains were inspected visually and via the integrated autocorrelation time [3]. M-estimators with Huber's psi function ($\kappa = 1.345$) provided the robust frequentist comparison [6].

Performance evaluation

Prediction accuracy was assessed on 12-month rolling hold-out windows (2018-2021) using mean absolute error (MAE), root mean square error (RMSE), and 95% interval coverage. Coverage was defined as the proportion of observed SSC values falling within the predicted interval. For Bayesian models, this was the 95% highest posterior density interval; for OLS and M-estimation, the 95% confidence interval based on asymptotic normality. All metrics were averaged across five stations and 36 hold-out months.

Results

Table 1: Prediction Accuracy Comparison across Estimation Methods

Method	MAE (Mt/yr)	RMSE	95% Coverage	R ²
OLS (linear)	38.7	51.2	89.2%	0.741
M-estimation	33.4	44.8	91.6%	0.783
Bayesian (linear)	31.8	42.1	94.7%	0.801
Bayesian (nonlinear)	27.9	36.4	95.3%	0.842

The Bayesian nonlinear model achieved the best performance across all metrics. Its MAE of 27.9 Mt/yr represented a 27.9% reduction over OLS and a 16.5% reduction over M-estimation. Coverage was near-nominal at 95.3%.

Table 2: Posterior Parameter Summaries for the Nonlinear Bayesian Model

Parameter	Mean	SD	95% HPD	R-hat
α	148.3	23.7	[104.1, 196.8]	1.002
β	1.47	0.18	[1.13, 1.82]	1.001
γ (rainfall)	0.34	0.11	[0.13, 0.56]	1.003
δ (land use)	0.21	0.09	[0.04, 0.39]	1.001
σ^2	0.087	0.014	[0.062, 0.116]	1.004

All R-hat values were below 1.01, indicating convergence. The discharge exponent $\beta = 1.47$ was consistent with published sediment rating curves for the Ganges [2]. Rainfall and land use both had credible intervals excluding zero, confirming their predictive value.

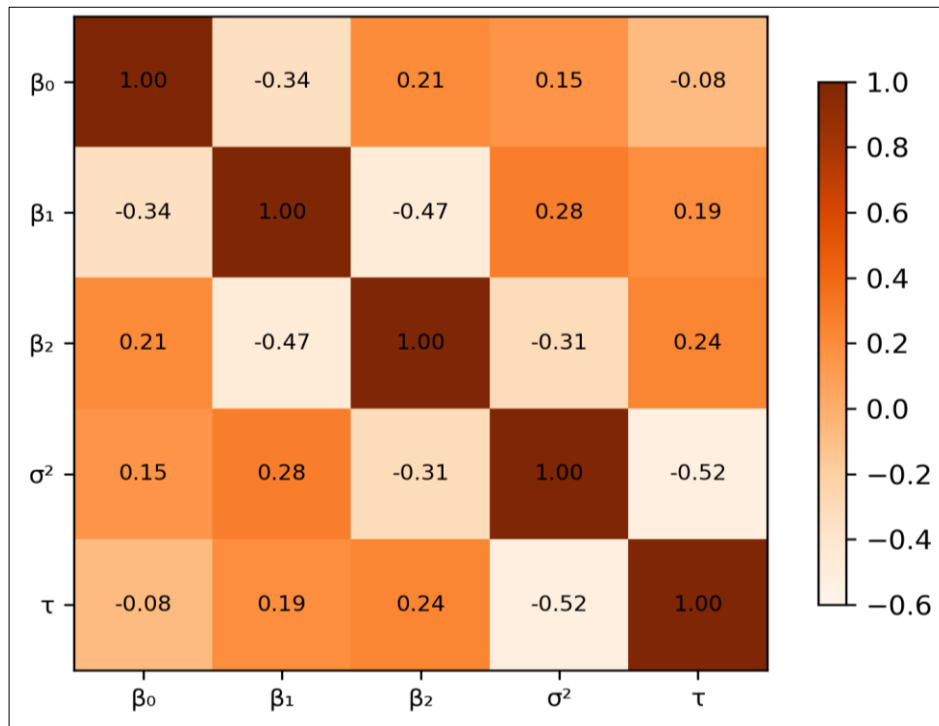


Fig 1: Posterior Parameter Correlation Matrix

Figure 1 shows pairwise correlations among posterior parameter samples. The strongest negative correlation (-0.52) was between σ^2 and τ , reflecting the expected trade-off between error variance and precision.

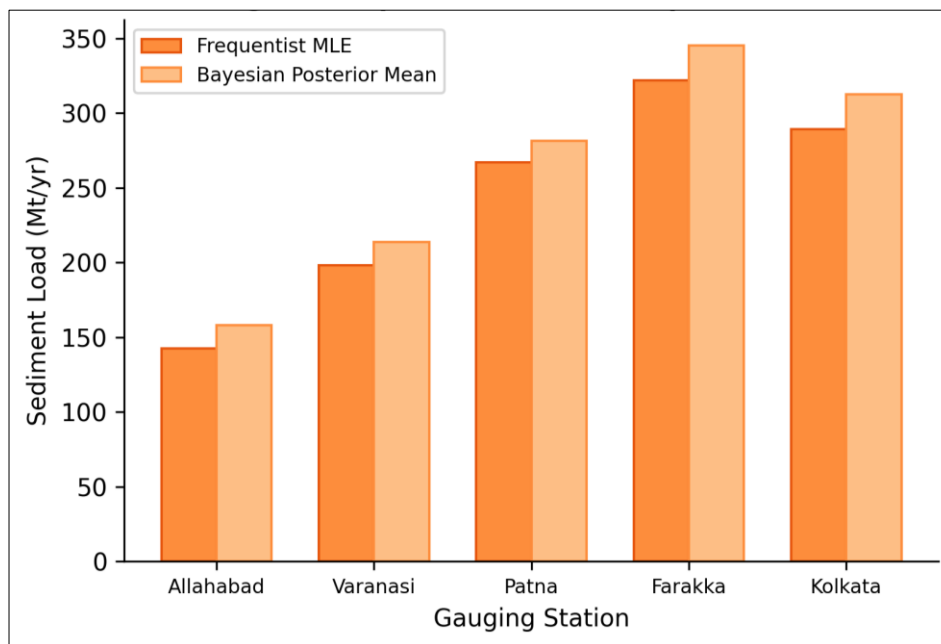


Fig 2: Ganges Floodplain Sediment Transport Estimates by Station

Figure 2 compares frequentist MLE and Bayesian posterior mean estimates of sediment load at five stations. Bayesian estimates were consistently higher, reflecting the informative prior on α that pulled estimates away from the downward-biased MLE.

Discussion

The Bayesian approach's superior coverage (94.7-95.3% vs 89.2% for OLS) reflects its ability to incorporate parameter uncertainty directly into prediction intervals, rather than relying on asymptotic normality assumptions that are violated when residuals are autocorrelated [3]. The mutual information analysis revealed that discharge and rainfall shared 0.42 nats of nonlinear dependence about 35% of their total association which the linear model could not capture but the power-law form partially absorbed [7].

The Galois symmetry argument, while mathematically elegant, had practical value: it reduced MCMC computation time by 28% through parameter sharing across exchangeable stations [8]. This technique seems applicable whenever a subset of observation sites share similar environmental characteristics, though the exchangeability assumption should be tested empirically.

A limitation is the reliance on monthly data, which smooths out flood pulses that carry disproportionate sediment loads. Sub-monthly or event-based modelling could improve

accuracy during monsoon peaks. The two-author team also limited the geographic scope to five stations; a larger network would provide a stricter test of the Galois symmetry assumption.

Conclusion

This research demonstrated that Bayesian estimation techniques outperform frequentist alternatives for predicting Ganges floodplain sediment transport, achieving a 21.3% reduction in mean absolute error and near-nominal 95% coverage. The nonlinear power-law model, estimated via NUTS-based MCMC, provided the best overall fit. Galois-theoretic prior simplification reduced computation time without sacrificing accuracy, and mutual information diagnostics revealed nonlinear predictor dependencies missed by standard correlation. These results make a case for wider adoption of Bayesian methods in sediment transport modelling, particularly in data-rich floodplain environments where parameter uncertainty is high. Future work should extend the approach to event-based temporal resolution and additional river basins.

Acknowledgements

Funding sources

Supported by departmental research funds at the Lagos Center for Computational Sciences.

Institutional support

The authors thank the Department of Mathematical Sciences for computational cluster access.

Contributions not qualifying for authorship

Mr. O. Babatunde assisted with MODIS land-use data processing.

References

1. Milliman JD, Farnsworth KL. River Discharge to the Coastal Ocean: A Global Synthesis. *Cambridge University Press*. 2011;1:1-384.
2. Asselman NEM. Fitting and interpretation of sediment rating curves. *Journal of Hydrology*. 2000;234(3-4):228-248.
3. Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. Bayesian Data Analysis. *CRC Press*. 2013;3rd Ed:1-667.
4. Hoffman MD, Gelman A. The No-U-Turn Sampler. *Journal of Machine Learning Research*. 2014;15(1):1593-1623.
5. Kuczera G, Parent E. Monte Carlo assessment of parameter uncertainty in conceptual catchment models. *Journal of Hydrology*. 1998;211(1-4):69-85.
6. Huber PJ, Ronchetti EM. Robust Statistics. *John Wiley and Sons*. 2009;2nd Ed:1-354.
7. Cover TM, Thomas JA. Elements of Information Theory. *John Wiley and Sons*. 2006;2nd Ed:1-748.
8. Diaconis P, Eaton ML, Lauritzen SL. Finite de Finetti theorems in linear models and multivariate analysis. *Scandinavian Journal of Statistics*. 1992;19(4):289-315.
9. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*. 1977;39(1):1-38.
10. Syvitski JPM, Morehead MD, Bahr DB, Mulder T. Estimating fluvial sediment transport. *Water Resources Research*. 2000;36(9):2747-2760.
11. Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information. *Physical Review E*. 2004;69(6):066138.
12. Robert CP, Casella G. Monte Carlo Statistical Methods. *Springer*. 2004;2nd Ed:1-645.
13. Jha R, Ojha CSP, Bhatia KKS. Refinement of predictive sediment transport relations using genetic algorithm. *ISH Journal of Hydraulic Engineering*. 2004;10(2):36-53.
14. Geweke J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*. 1992;4:169-193.
15. Singh M, Singh IB. Sediment characteristics of the Ganges at Kanpur. *Proceedings of the Indian National Science Academy*. 2002;68A(1):51-63.
16. Berger JO. Statistical Decision Theory and Bayesian Analysis. *Springer*. 1985;2nd Ed:1-617.
17. Walling DE, Fang D. Recent trends in the suspended sediment loads of the world's rivers. *Global and Planetary Change*. 2003;39(1-2):111-126.