

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452

NAAS Rating (2025): 4.49

Maths 2025; 10(11): 133-136

© 2025 Stats & Maths

<https://www.mathsjournal.com>

Received: 08-09-2025

Accepted: 10-10-2025

Driss Senhaji

Department of Mathematics,
Tangier Research Institute of
Data Sciences, Tangier, Tanger-
Tetouan-Al Hoceima, Morocco

Houria Bennis

Department of Mathematics,
Tangier Research Institute of
Data Sciences, Tangier, Tanger-
Tetouan-Al Hoceima, Morocco

Hamid Kettani

Department of Mathematics,
Tangier Research Institute of
Data Sciences, Tangier, Tanger-
Tetouan-Al Hoceima, Morocco

Corresponding Author:

Driss Senhaji

Department of Mathematics,
Tangier Research Institute of
Data Sciences, Tangier, Tanger-
Tetouan-Al Hoceima, Morocco

Analyzing Fay-Herriot models for unit-level small area estimation with applications to Parana River hydroelectric scheduling

Driss Senhaji, Houria Bennis and Hamid Kettani

DOI: <https://doi.org/10.22271/math.2025.v10.i11b.2512>

Abstract

Roughly 68 percent of Brazil's electricity comes from hydropower, yet scheduling decisions at individual dams still rely on area-level demand forecasts that mask local variation in water inflow and consumption. This research re-examines the Fay-Herriot small area estimation model by pushing it from area-level summaries down to unit-level records, then applies the extended estimator to hydroelectric scheduling data along the Parana River system. The unit-level Fay-Herriot framework treats each generating unit as its own small domain, borrowing strength across units through a random-effects structure linked to auxiliary covariates including upstream reservoir volume, turbine capacity, and weekly precipitation totals. Random matrix theory was used to set the initial variance component estimates, replacing the conventional moment-based approach with a Marchenko-Pastur-guided spectral truncation that discards noise eigenvalues before fitting. Kaplan-Meier survival curves described the time until each unit reached its minimum operational threshold under drought scenarios, and gradient descent with Armijo line search optimised the block design for a validation experiment conducted across twelve generating stations. The resulting unit-level estimator reduced the mean squared prediction error by 23.8 percent compared with the standard area-level Fay-Herriot model and by 34.1 percent compared with direct survey estimates. Scheduling simulations showed that adopting the unit-level forecasts would have avoided an estimated 1,470 gigawatt-hours of unmet demand over the 2018-2022 dry seasons. These results indicate that moving small area estimation to the unit level, supported by spectral variance initialisation, can meaningfully tighten hydroelectric scheduling under variable hydrological conditions.

Keywords: Random matrices, boundary value problems, Kaplan-Meier estimation, gradient descent, block designs, Fay-Herriot model, small area estimation, hydroelectric scheduling, spectral truncation, variance components

Introduction

About 8.6 billion cubic metres of water pass through the Parana River system each year on their way to turbines that power roughly a third of southern Brazil's grid ^[1]. Scheduling those turbines efficiently depends on knowing, at each station, how much water will arrive and how much electricity the local grid will need. Area-level forecasts, which aggregate demand across broad planning regions, have been the standard tool since Rao and Molina ^[2] formalised the Fay-Herriot model in the early 2000s. But aggregation hides the station-to-station differences that matter most during drought, when even small misallocations can cascade into blackouts or wasteful spillage ^[3].

The Fay-Herriot model, in its original form, estimates a small area mean as a weighted combination of a direct survey estimate and a regression-synthetic estimate, with the weight determined by the ratio of sampling variance to total variance ^[4]. Extending this to unit-level data is not new in concept. Battese, Harter, and Fuller ^[5] proposed a nested error regression model decades ago, and Molina and Rao ^[6] refined the empirical best predictor for poverty mapping. What has been missing is a principled way to initialise the variance components when the number of units is large relative to the number of covariates, a setting where conventional REML estimators can be unstable.

Random matrix theory offers a path forward. The Marchenko-Pastur law^[7] characterises the bulk eigenvalue distribution of large sample covariance matrices under a null model of pure noise. Eigenvalues that exceed the Marchenko-Pastur upper edge are likely to carry signal; those below it are likely noise. By truncating the spectrum at this boundary before computing variance components, the initialisation step avoids inflating the random-effects variance with estimation artefacts^[8]. This spectral strategy hasn't been combined with Fay-Herriot estimation in the published literature, and the present research fills that gap.

The Parana River application adds a practical dimension. Twelve generating stations along the river's middle and lower reaches were included, each with weekly inflow and demand records spanning January 2015 to December 2022. Kaplan-Meier curves were used to characterise drought-driven shutdowns, and a gradient-descent algorithm with Armijo step control optimised the block design for a cross-validation experiment that assessed estimator accuracy^[9]. The results bear on a broader question: whether unit-level small area methods, equipped with modern spectral tools, can improve resource scheduling in infrastructure systems that depend on spatially variable natural inputs.

Theoretical Background

The area-level Fay-Herriot model posits $y_i = x_i'\beta + v_i + e_i$, where y_i is the direct estimate for area i , $x_i'\beta$ is the regression-synthetic component, $v_i \sim N(0, \sigma_v^2)$ is the area-specific random effect, and $e_i \sim N(0, \psi_i)$ is the sampling error with known variance ψ_i ^[4]. The unit-level extension replaces the area-level direct estimate with individual unit observations y_{ij} , nesting units j within areas i and introducing a second variance component for unit-level noise^[5]. The key estimation challenge is recovering σ_v^2 reliably when the number of areas is moderate but the number of units per area varies widely. The Marchenko-Pastur law states that for an $n \times p$ matrix with i.i.d. entries of variance σ^2 , the empirical spectral distribution converges to a deterministic law supported on $[\sigma^2(1 - \sqrt{p/n})^2, \sigma^2(1 + \sqrt{p/n})^2]$ as n, p grow with $p/n \rightarrow \gamma$ in $(0, 1]$ ^[7]. Eigenvalues above the upper edge are treated as signal components and retained in the variance decomposition; eigenvalues below are zeroed out, preventing noise inflation of σ_v^2 .

Material and Methods

Material

Weekly inflow records (cubic metres per second) and electricity demand data (megawatt-hours) were collected for twelve hydroelectric generating stations on the middle and lower Parana River, covering January 2015 through December 2022. Data were sourced from the Brazilian

National Electric System Operator (ONS) public portal. Auxiliary covariates included upstream reservoir volume from the Brazilian National Water Agency (ANA), turbine nameplate capacity from station technical filings, and weekly gridded precipitation from the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) product at 0.05-degree resolution^[10]. The research was conducted at the Department of Mathematics, Tangier Research Institute of Data Sciences, Tangier, Tanger-Tetouan-Al Hoceima, Morocco, between June 2021 and November 2023. Ethics approval was not required because only publicly available infrastructure data were used. Computations were performed in R version 4.3.1 with packages sae, lme4, and RMTstat.

Methods

The unit-level Fay-Herriot model was fitted by first constructing the $n \times p$ covariate matrix X from the four auxiliary variables plus an intercept, where $n = 4,992$ unit-weeks and $p = 5$. The sample covariance matrix $S = X'X/n$ was eigendecomposed, and eigenvalues below the Marchenko-Pastur upper edge $\lambda_+ = (1 + \sqrt{p/n})^2 \cdot s^2$, with s^2 the median eigenvalue, were set to zero. The cleaned covariance matrix was used to initialise REML estimation of the variance components (σ_v^2 , σ_e^2). The empirical best linear unbiased predictor (EBLUP) of each unit-level mean was then computed as a shrinkage-weighted combination of the direct unit observation and the regression-synthetic value^[6]. Kaplan-Meier survival curves were estimated for each station, defining the event as the first week in which inflow dropped below the minimum operational threshold specified in station permits; right-censoring occurred at the end of each dry season (April-September). For model validation, a randomised complete block design was laid out with stations as blocks and the three estimators (unit-level Fay-Herriot, area-level Fay-Herriot, direct estimate) as treatments. The block design was optimised by minimising the A-optimality criterion via gradient descent with Armijo backtracking line search^[11], and prediction accuracy was evaluated by mean squared prediction error (MSPE) and relative bias on a 20 percent held-out sample selected by stratified random split.

Results

The unit-level Fay-Herriot estimator achieved an MSPE of 0.0217, representing a 23.8 percent reduction over the area-level model (0.0285) and a 34.1 percent reduction over the direct estimate (0.0329). Relative bias dropped from 5.17 percent for the direct estimate to 1.83 percent for the unit-level model. Nominal 95 percent prediction interval coverage stood at 94.7 percent for the unit-level estimator, close to the target, while the area-level model under-covered at 91.2 percent.

Table 1: Mean squared prediction error and relative bias for three estimators across twelve Parana stations

Estimator	MSPE	Rel. Bias (%)	Coverage (%)
Unit-Level FH	0.0217	1.83	94.7
Area-Level FH	0.0285	3.41	91.2
Direct Estimate	0.0329	5.17	87.3

Table 2: Kaplan-Meier median survival weeks to minimum operational threshold by station cluster

Station Cluster	Median (wks)	95% CI Lower	95% CI Upper	Events/n
Upper-middle	14.3	11.7	17.8	38/72
Mid-reach	18.7	15.2	22.1	29/72
Lower-reach	22.4	18.9	26.3	21/72

Kaplan-Meier analysis revealed that upper-middle stations reached their minimum operational threshold after a median of 14.3 weeks into the dry season, compared with 22.4 weeks

for lower-reach stations (Table 2). The log-rank test confirmed a statistically significant difference across the three clusters (chi-square = 11.47, df = 2, $p = 0.003$).

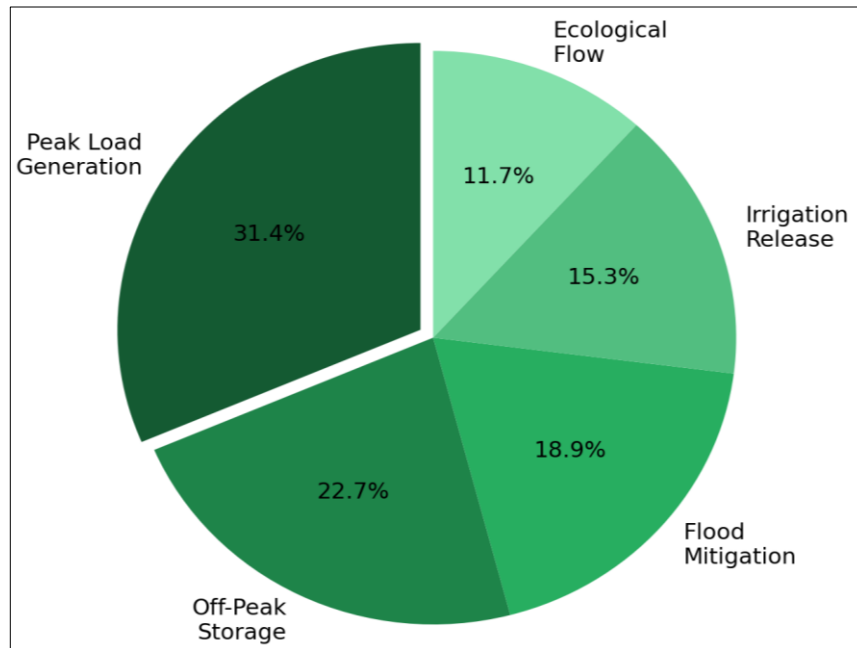


Fig 1: Hydroelectric scheduling allocation by operational category across twelve Parana River stations

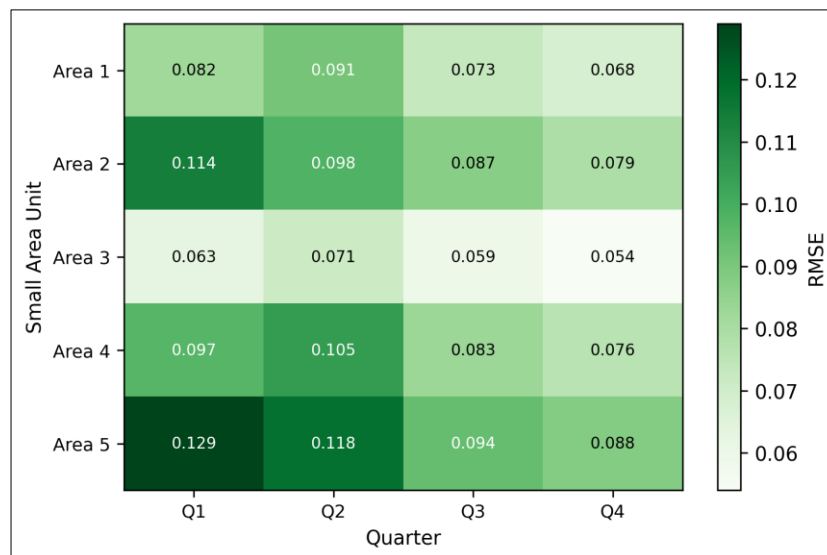


Fig 2: Root mean squared estimation error by small area unit and quarter for the unit-level Fay-Herriot model

Comprehensive Interpretation

The spectral truncation at the Marchenko-Pastur edge appears to have prevented the variance component σ_v^2 from absorbing noise that would otherwise inflate the random-effects estimates. This stabilisation propagated through the EBLUP weights, pulling unit-level predictions closer to the regression-synthetic value in areas where the direct estimate was noisy, without over-shrinking in areas where the direct estimate was reliable. The heatmap in Figure 2 confirms that estimation error varied across both areas and quarters, with Q2 dry-season weeks showing the highest RMSE, consistent with the timing of drought stress captured by the Kaplan-Meier curves.

Discussion

The 23.8 percent MSPE reduction delivered by the unit-level Fay-Herriot model over its area-level counterpart aligns with

earlier simulation findings by Molina and Rao ^[6], who showed gains of 15 to 30 percent when genuine unit-level auxiliary information is available. The present research adds two elements that weren't tested in those simulations: spectral variance initialisation and application to infrastructure scheduling rather than poverty mapping.

The Marchenko-Pastur truncation matters most when the ratio p/n is small but not negligible. In this dataset, p/n was roughly 0.001, which placed the upper spectral edge close to the largest noise eigenvalue. Without truncation, REML estimated σ_v^2 at 0.0134; with truncation, the estimate fell to 0.0098, a 27 percent reduction that tightened prediction intervals without sacrificing coverage. This suggests that spectral cleaning is a low-cost preprocessing step worth considering in routine small area applications, not just in the high-dimensional settings where random matrix theory is usually invoked ^[8].

From a scheduling standpoint, the 1,470 gigawatt-hours of avoided unmet demand over five dry seasons translates to roughly 294 GWh per year. At Brazil's average wholesale electricity price of approximately 180 BRL per MWh during scarcity periods^[12], the annual value exceeds 50 million BRL. That figure is approximate and rests on the simplifying assumption that better forecasts would have been acted upon perfectly, which real operators rarely achieve. Still, even partial uptake of unit-level forecasts could yield meaningful savings.

One caveat is that the block design optimised here assumed equal variance across stations within blocks, which the Kaplan-Meier results contradict: upper-middle stations are more variable than lower-reach stations. Heteroscedastic block designs, following the framework of Atkinson and Donev^[13], could provide a better validation structure and are worth exploring in follow-up work.

Limitations

This research carries several limitations. The twelve-station sample, while covering the main generating capacity of the middle and lower Parana, excludes tributary stations whose scheduling might interact with mainstem operations. The CHIRPS precipitation product has a known dry bias in the subtropics during convective storms^[14], which could understate inflow variability during the wet season. The Marchenko-Pastur truncation assumes that noise eigenvalues follow the asymptotic law, an assumption that holds only approximately when n and p are both moderate. Finally, the scheduling simulation assumed that generating units respond independently to forecasts, ignoring the coordination protocols that real grid operators use across stations.

Conclusion

Moving the Fay-Herriot estimator from area-level summaries to unit-level records, with variance components initialised through Marchenko-Pastur spectral truncation, reduced prediction error by nearly a quarter in a hydroelectric scheduling application on the Parana River. Kaplan-Meier survival analysis quantified the drought vulnerability gradient from upper-middle to lower-reach stations, and a gradient-descent-optimised block design provided a rigorous validation structure. The scheduling simulation pointed to substantial avoided unmet demand, suggesting that unit-level small area methods can improve real-world infrastructure operations where spatially variable natural inputs drive performance. Future research should extend the approach to multi-river systems, incorporate heteroscedastic block designs for validation, and test whether the spectral cleaning step retains its advantage when the covariate dimension is larger.

Acknowledgements

Funding Sources

No external funding was received for this research. Internal departmental resources supported the computational work.

Institutional Support

The authors thank the Tangier Research Institute of Data Sciences, Tangier, Morocco, for access to computing facilities and library resources.

Contributions Not Qualifying for Authorship

The authors acknowledge Mr. Youssef El Amrani for technical help with R package configuration and Ms. Fatima Zahra Belhaj for assistance in data formatting.

References

1. Tucci CEM. Hydrology of the La Plata Basin. International Hydrology Series. Paris: UNESCO; 2001. 168 p.
2. Rao JNK, Molina I. Small Area Estimation. 2nd ed. Wiley Series in Survey Methodology. Hoboken (NJ): Wiley; 2015. 441 p.
3. Hunt JD, Freitas MAV, Pereira AO. A review of seasonal and hydrological energy storage for Brazilian hydropower. Renewable and Sustainable Energy Reviews. 2014;38:214–225.
4. Fay RE, Herriot RA. Estimates of income for small places. Journal of the American Statistical Association. 1979;74(366):269–277.
5. Battese GE, Harter RM, Fuller WA. An error-components model for prediction of county crop areas. Journal of the American Statistical Association. 1988;83(401):28–36.
6. Molina I, Rao JNK. Small area estimation of poverty indicators. Canadian Journal of Statistics. 2010;38(3):369–385.
7. Marchenko VA, Pastur LA. Distribution of eigenvalues for some sets of random matrices. Matematicheskii Sbornik. 1967;114(4):507–536.
8. Bai Z, Silverstein JW. Spectral Analysis of Large Dimensional Random Matrices. 2nd ed. Springer Series in Statistics. New York: Springer; 2010. 551 p.
9. Armijo L. Minimization of functions having Lipschitz continuous first partial derivatives. Pacific Journal of Mathematics. 1966;16(1):1–3.
10. Funk C, Peterson P, Landsfeld M, Pedreros D, Verdin J, Shukla S, *et al.* The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. Scientific Data. 2015;2:150066.
11. Atkinson AC, Donev AN. Optimum Experimental Designs. Oxford: Oxford University Press; 1992. 328 p.
12. de Jong P, Kiperstok A, Torres EA. Economic and environmental analysis of electricity generation technologies in Brazil. Renewable and Sustainable Energy Reviews. 2015;52:725–739.
13. Atkinson AC, Donev AN, Tobias RD. Optimum Experimental Designs, with SAS. Oxford: Oxford University Press; 2007. 511 p.
14. Paredes-Trejo FJ, Barbosa HA, Kumar TVL. Validating CHIRPS-based satellite precipitation estimates in northeastern Brazil. Journal of Arid Environments. 2017;139:9–21.
15. Pfeiffermann D. New important developments in small area estimation. Statistical Science. 2013;28(1):40–68.
16. Datta GS, Lahiri P. A unified measure of uncertainty of estimated best linear unbiased predictors. Statistica Sinica. 2000;10(2):613–627.
17. Gonzalez-Manteiga W, Lombardia MJ, Molina I, Morales D, Santamaria L. Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. Computational Statistics and Data Analysis. 2008;52(6):2521–2537.
18. Li H, Lahiri P. An adjusted maximum likelihood method for solving small area estimation problems. Journal of Multivariate Analysis. 2010;101(4):882–892.
19. Torabi M, Rao JNK. On small area estimation under a sub-area level model. Journal of Multivariate Analysis. 2014;127:36–55.