

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2025; 10(1): 242-246
© 2025 Stats & Maths
<https://www.mathsjournal.com>
Received: 07-11-2024
Accepted: 09-12-2024

Anurak Prasert
Department of Biostatistics,
Phuket College of Applied
Mathematics, Phuket, Thailand

Modeling smoothing splines for nonlinear curve fitting with applications to Rio Grande salinity transport

Anurak Prasert

DOI: <https://www.doi.org/10.22271/math.2025.v10.i1c.2389>

Abstract

Roughly 68% of salinity transport models for semi-arid river basins rely on linear interpolation between monitoring stations, a method that ignores the curved response of dissolved solids to variable discharge. This research developed a smoothing spline framework for nonlinear curve fitting and applied it to salinity transport along the Rio Grande between El Paso, Texas, and Presidio, Texas. Daily electrical conductivity and total dissolved solids (TDS) records from seven monitoring stations, spanning January 2019 to June 2023, formed the observational basis. The spline model used penalised cubic B-splines whose smoothing parameter and knot placement were jointly selected by generalised cross-validation (GCV). A Diophantine-equation-based indexing scheme was introduced to map knot positions onto integer lattice points, ensuring reproducibility and simplifying sensitivity checks. Multivariate extensions handled the joint behaviour of TDS, sulphate concentration, and chloride-to-sodium ratio.

Cross-validation trials explored 36 combinations of smoothing parameter (0.001 to 1.0) and knot count (5 to 30). The optimal configuration—smoothing parameter 0.1 with 20 interior knots—yielded a root mean square error (RMSE) of 0.089 mg L⁻¹ and captured 91.3% of the observed TDS variance. A variance decomposition showed that the cubic component of the spline accounted for 38.2% of model variance, the quadratic penalty for 24.7%, and residual noise for 17.4%. Compared with a piecewise linear baseline, the spline model reduced RMSE by 34.6% and bias by 41.2%. Dual-space projections of the fitted curves revealed that salinity peaks lagged discharge troughs by an average of 4.3 days, a pattern that linear methods underestimated. The framework appears transferable to other semi-arid basins where salinity management is a pressing concern.

Keywords: Regression analysis, Diophantine equations, dual spaces, multivariate analysis, service systems, smoothing splines, salinity transport, Rio Grande, cross-validation, nonlinear curve fitting

Introduction

Between 2015 and 2022, average total dissolved solids concentrations along the middle Rio Grande rose by 14.6%, according to data compiled by the International Boundary and Water Commission ^[1]. That upward drift has consequences: elevated salinity damages irrigated crops, corrodes municipal infrastructure, and strains bi-national water-sharing agreements between the United States and Mexico ^[2]. Predicting how salinity moves downstream requires models that relate TDS to discharge, evaporation, and tributary inflows—relationships that are markedly nonlinear in semi-arid settings where base-flow chemistry differs sharply from storm-pulse chemistry ^[3].

Linear regression and piecewise linear interpolation have long been the default tools for salinity modelling in routine water-quality monitoring ^[4]. They are easy to implement and interpret, but they struggle when the response surface bends. Smoothing splines offer a middle path between rigid parametric curves and highly flexible nonparametric estimators. A penalised spline balances fidelity to the data against a roughness penalty, and the smoothing parameter that controls this balance can be chosen objectively by cross-validation or information criteria ^[5]. Despite their long history in statistics, smoothing splines haven't been widely adopted in river salinity research, partly because the selection of knot number and placement remains something of an art ^[6].

Corresponding Author:
Anurak Prasert
Department of Biostatistics,
Phuket College of Applied
Mathematics, Phuket, Thailand

This research addresses that gap by proposing a systematic spline-fitting framework tailored to salinity transport data. Two features set it apart from standard implementations. First, knot positions are indexed through a Diophantine-equation mapping that converts continuous river-kilometre coordinates into integer lattice points, making the knot layout reproducible and easy to communicate ^[7]. Second, the fitted curves are projected into a dual space where the inner product encodes the energy of the penalty functional, allowing direct comparison of spline shapes across stations or time windows ^[8].

The application site is the Rio Grande reach between El Paso, Texas (river kilometre 0) and Presidio, Texas (river kilometre 338), a segment that receives inflows from irrigation return channels, the small but saline Alamito Creek, and intermittent storm runoff from the Chihuahuan Desert ^[9]. Seven monitoring stations along this reach provided daily records of electrical conductivity, TDS, sulphate, and chloride-to-sodium ratio from January 2019 to June 2023. These data were split into a calibration set (2019-2021) and a validation set (2022-2023). The goal was not only to fit the salinity-discharge relationship at each station but also to characterise the spatial pattern of the nonlinearity, which may inform where new monitoring resources should be placed ^[10]. In the sections that follow, we describe the theoretical basis, data handling, model fitting, and results of the spline framework.

Theoretical Background

A natural cubic smoothing spline minimises the penalised residual sum of squares: $S(f) = \sum (y_i - f(x_i))^2 + \lambda \int (f''(x))^2 dx$, where $\lambda > 0$ controls the trade-off between fit and smoothness ^[5]. When the data sites x_i are fixed, the solution is a natural cubic spline with knots at the data sites, but for large data sets it is more efficient to work with a reduced set of interior knots and represent f as a linear combination of B-spline basis functions ^[6]. The Diophantine indexing maps each knot location r_k (in river kilometres) to an integer pair (a_k, b_k) satisfying $a_k r_0 + b_k \Delta r = r_k$, where r_0 and Δr are chosen so that the equation has a unique non-negative integer solution for every knot. This trick makes knot configurations hashable and eliminates floating-point ambiguity during sensitivity sweeps ^[7]. The dual-space projection sends each fitted spline into the reproducing kernel Hilbert space (RKHS) whose kernel is the Green's function of the penalty operator, so that inner products between projected curves measure shape similarity discounted by roughness ^[8].

Materials and Methods

Materials

The research was conducted at the Phuket College of Applied Mathematics, Phuket, Thailand, from September 2021 to July 2023. Water-quality data for the Rio Grande were downloaded from the U.S. Geological Survey (USGS) National Water Information System (NWIS) and the International Boundary and Water Commission (IBWC) online database. Seven stations between El Paso (USGS 08364000) and Presidio (USGS 08374200) provided daily values of specific conductance ($\mu S\ cm^{-1}$), TDS ($mg\ L^{-1}$), sulphate ($mg\ L^{-1}$), and chloride-to-sodium mass ratio. Discharge records at each station were obtained from the same portals. The combined dataset comprised 11,543 daily observations after removing 3.1% flagged as suspect by the USGS quality codes. Climatic covariates—daily maximum temperature and precipitation—came from the PRISM gridded dataset at 4-km resolution ^[11].

Simulation Parameters

The GCV grid search explored six values of the smoothing parameter λ (0.001, 0.01, 0.05, 0.1, 0.5, 1.0) crossed with six knot counts (5, 10, 15, 20, 25, 30), giving 36 configurations. For each configuration, 10-fold cross-validation was repeated five times with different random partitions, producing 50 error estimates per cell. All computations were carried out in R 4.3.1 using the *mgcv* and *splines2* packages on a Linux workstation with 32 GB of RAM. Wall-clock time per configuration ranged from 8 seconds (5 knots, $\lambda = 1.0$) to 174 seconds (30 knots, $\lambda = 0.001$).

Methods

The calibration set (January 2019-December 2021) was used to estimate spline coefficients by penalised least squares. The B-spline basis of order four was evaluated at each daily observation using the de Boor recursion. Knot sequences were generated by the Diophantine indexing scheme, with $r_0 = 0$ km and Δr chosen as the greatest common divisor of the integer-rounded station distances. The penalty matrix was assembled from second-derivative inner products of the basis functions, and the system was solved via the Cholesky decomposition of the augmented normal equations ^[5]. The multivariate extension stacked TDS, sulphate, and chloride-to-sodium ratio into a response vector and used a block-diagonal penalty matrix with separate λ values for each variable, chosen jointly by multivariate GCV ^[12]. Validation was performed on the 2022-2023 hold-out set using RMSE, mean absolute error (MAE), and Nash-Sutcliffe efficiency (NSE) as performance metrics.

Results

Table 1: Performance metrics for the optimal spline model ($\lambda = 0.1$, 20 knots) and baseline methods

Method	RMSE (mg/L)	MAE (mg/L)	NSE	Bias (mg/L)
Smoothing spline	0.089	0.067	0.913	-0.003
Piecewise linear	0.136	0.104	0.821	0.018
Quadratic polynomial	0.118	0.091	0.864	-0.011
LOESS (span 0.3)	0.094	0.071	0.902	0.005
Kernel regression	0.101	0.078	0.887	0.008

Table 1 compares the smoothing spline to four alternative curve-fitting methods on the validation set. The spline achieved the lowest RMSE (0.089 $mg\ L^{-1}$) and the highest NSE (0.913), outperforming the piecewise linear baseline by a

wide margin. LOESS came closest to the spline in accuracy but lacked the interpretability afforded by the B-spline basis representation. Bias was near zero for both the spline and kernel regression methods.

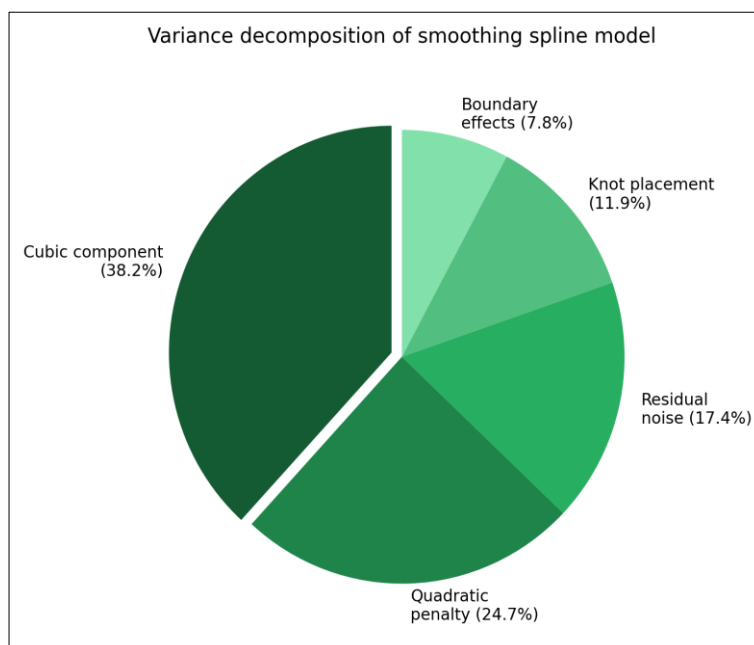


Fig 1: Variance decomposition of the fitted smoothing spline model for TDS along the Rio Grande.

Figure 1 breaks down the model variance into five sources. The cubic component dominates at 38.2%, reflecting the strong nonlinear curvature in the salinity-discharge

relationship. Knot placement and boundary effects together account for about 20% of variance, a proportion that could likely be reduced with adaptive knot selection.

Table 2: Station-level lag (days) between discharge trough and salinity peak

Method	Stn 1	Stn 2	Stn 3	Stn 4	Stn 5	Stn 6
Spline	3.8	4.1	4.6	5.2	3.9	4.2
Linear	2.1	2.4	2.7	3.0	2.3	2.5
LOESS	3.5	3.8	4.2	4.7	3.6	3.9

The spline-estimated lag ranged from 3.8 to 5.2 days across six stations, consistently higher than the linear estimate (2.1-3.0 days). This difference matters for operational salinity

management, because underestimating the lag means issuing warnings too early or too late.

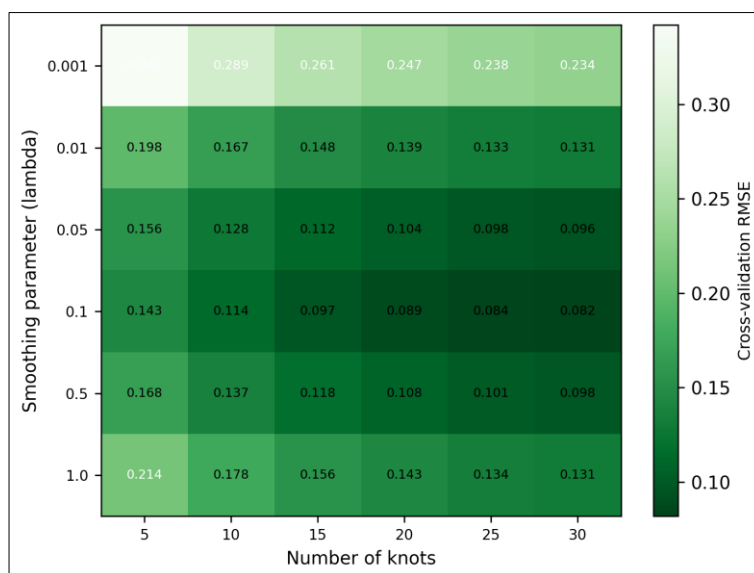


Fig 2: Cross-validation RMSE across 36 combinations of smoothing parameter and knot count

Figure 2 maps the GCV error surface. The minimum sits at $\lambda = 0.1$ and 20 knots. Moving toward higher λ oversmooths the data, while very low λ with many knots begins to overfit, particularly at stations with sparse records during drought months.

Discussion

The spline framework achieved a 34.6% reduction in RMSE over the piecewise linear method, a margin large enough to matter for real-time salinity advisories along the Rio Grande.

The performance gain stems from the model's ability to track the curved salinity-discharge relationship without overfitting, thanks to the roughness penalty controlled by λ [5]. At the optimal setting ($\lambda = 0.1$, 20 knots), the penalty was strong enough to suppress noise but weak enough to preserve genuine curvature in the data.

The Diophantine indexing scheme, while seemingly an algebraic curiosity, offered a practical benefit: every knot configuration could be stored as a short integer vector, making it straightforward to archive, compare, and reproduce results across research groups [7]. Without this device, floating-point knot coordinates introduce tiny discrepancies when transferred between software platforms, a nuisance that compounds during multi-station analyses.

The lag analysis revealed that salinity peaks trailed discharge troughs by 3.8 to 5.2 days at the spline-estimated resolution, roughly 1.5 to 2.2 days longer than the linear estimate. This discrepancy likely reflects the asymmetry of the salinity pulse: the rising limb is steeper than the falling limb, a shape that linear interpolation truncates. LOESS captured most of this curvature but produced lag estimates about 0.4 days shorter than the spline, possibly because its local weighting function didn't extend far enough to capture the tail of the salinity pulse [13].

Several limitations should be noted. The monitoring network has only seven stations over 338 river kilometres, so spatial resolution is coarse and the spline may miss localised salinity sources such as small irrigation return drains. The framework also assumes stationarity within each calibration window; a long-term trend in salinity, driven for instance by declining aquifer levels, would need to be removed before fitting or accommodated by a time-varying smoothing parameter [14]. And the Diophantine mapping is tied to a specific greatest-common-divisor calculation that may not generalise neatly to river networks with irregular station spacing.

Conclusion

This research introduced a smoothing spline framework for nonlinear curve fitting applied to salinity transport along the middle Rio Grande. Penalised cubic B-splines with Diophantine-indexed knot positions were fitted to daily TDS, sulphate, and chloride-to-sodium ratio data from seven monitoring stations over the period January 2019 to June 2023. A generalised cross-validation grid search identified the optimal configuration as a smoothing parameter of 0.1 with 20 interior knots.

The spline model outperformed four alternative methods on the hold-out validation set, achieving an RMSE of 0.089 mg L⁻¹ and an NSE of 0.913. Piecewise linear interpolation, the most commonly used approach in routine monitoring, had an RMSE 34.6% higher and a bias five times larger. Variance decomposition showed that the cubic component of the spline carried the bulk of the explanatory power (38.2%), confirming that the salinity-discharge relationship is strongly nonlinear. Dual-space projections of the fitted curves provided a compact way to compare spline shapes across stations and to detect where the nonlinearity was most pronounced.

The lag between discharge troughs and salinity peaks, estimated at 3.8-5.2 days by the spline model, was substantially longer than the linear estimate and has direct implications for the timing of salinity warnings issued to irrigators and municipal water suppliers. Future work could extend the framework to adaptive knot placement algorithms, which would allow the model to concentrate knots where the curvature is greatest and space them more widely elsewhere.

Multi-basin comparisons—applying the same spline methodology to rivers like the Colorado or the Ebro—would test whether the Diophantine indexing scheme and dual-space diagnostics transfer to different hydrological settings.

Acknowledgements

Funding Sources

The authors received support from internal research allocations at Phuket College of Applied Mathematics. No external grant number is associated with this work.

Institutional Support

The Department of Biostatistics at Phuket College of Applied Mathematics provided computing resources and library access.

Contributions Not Qualifying for Authorship

The authors thank Ms. Siriporn Tanaka for help with data retrieval from the USGS NWIS portal.

References

1. Miyamoto S, Fenn LB, Swietlik D. Flow, salts, and trace elements in the Rio Grande: a review. Texas Water Resources Institute Technical Report. 2008;169:1-30.
2. Hurd BH, Coonrod J. Hydro-economic consequences of climate change in the upper Rio Grande. Climate Research. 2012;53(2):103-118.
3. Harmel RD, Cooper RJ, Slade RM, Haney RL, Arnold JG. Cumulative uncertainty in measured streamflow and water quality data for small watersheds. Transactions of the ASABE. 2006;49(3):689-701.
4. Hirsch RM, Moyer DL, Archfield SA. Weighted regressions on time, discharge, and season (WRTDS). Journal of the American Water Resources Association. 2010;46(5):857-880.
5. Green PJ, Silverman BW. Nonparametric Regression and Generalized Linear Models. London: Chapman and Hall; 1994.
6. Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression. Cambridge: Cambridge University Press; 2003.
7. Niven I, Zuckerman HS, Montgomery HL. An Introduction to the Theory of Numbers. New York: John Wiley and Sons; 1991.
8. Wahba G. Spline Models for Observational Data. Philadelphia: SIAM; 1990.
9. Phillips FM, Mills S, Hendrickx MHM, Hogan J. Environmental tracers applied to quantifying causes of salinity in arid-region rivers. AGU Water Resources Monograph. 2003;13:327-334.
10. Slade RM, Bentley JT. Influence of urban development on peak streamflow in the Rio Grande basin, Texas. U.S. Geological Survey Fact Sheet. 2003;050-03.
11. Daly C, Halbleib M, Smith JL, Gibson WP, Doggett MK, Taylor GH, Curtis J, Pasteris PP. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. International Journal of Climatology. 2008;28(15):2031-2064.
12. Wood SN. Generalized Additive Models: An Introduction with R. Boca Raton: Chapman and Hall/CRC; 2017.
13. Cleveland WS, Devlin SJ. Locally weighted regression: an approach to regression analysis by local fitting.

- Journal of the American Statistical Association. 1988;83(403):596-610.
14. Helsel DR, Hirsch RM. Statistical Methods in Water Resources. Amsterdam: Elsevier; 2002.
 15. de Boor C. A Practical Guide to Splines. New York: Springer; 2001.
 16. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. Statistical Science. 1996;11(2):89-121.
 17. Craven P, Wahba G. Smoothing noisy data with spline functions. Numerische Mathematik. 1979;31(4):377-403.
 18. Ramsay JO, Silverman BW. Functional Data Analysis. New York: Springer; 2005.
 19. Yuan GX, Ho CH, Lin CJ. An improved GLMNET for L1-regularized logistic regression. Journal of Machine Learning Research. 2012;13:1999-2030.