

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(3): 117-120
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 28-03-2024
Accepted: 01-05-2024

Jose Santos
Department of Algebra and
Geometry, Manila Regional
Center for Statistical Methods,
Manila, Metro Manila,
Philippines

Analyzing kernel density estimation for multimodal distributions with applications to Argentine pampas wind energy

Jose Santos

DOI: <https://www.doi.org/10.22271/math.2024.v9.i3b.2350>

Abstract

Can nonparametric density estimation methods reliably capture multimodal wind speed patterns across large regions? This research examines kernel density estimation (KDE) techniques applied to wind energy data from the Argentine Pampas. Five bandwidth selectors Silverman's rule, Scott's method, Sheather-Jones plug-in, least-squares cross-validation, and biased cross-validation were compared using Gaussian and Epanechnikov kernels. Wind speed records from six meteorological stations (January 2018 through December 2022) provided 8,742 daily observations. Accuracy was measured through integrated squared error (ISE) and Kullback-Leibler divergence. The Sheather-Jones selector with a Gaussian kernel returned the lowest mean ISE of 0.923, outperforming alternatives by 3.7 to 14.3 percent. Seasonal decomposition showed accuracy dropped during transitional months (April and September), when distributions shifted from unimodal to bimodal shapes, coinciding with a 17.8 percent rise in the divergence metric. These findings point toward adaptive bandwidth strategies for improving wind power forecasting in regions with pronounced seasonal variability.

Keywords: Kernel density estimation, multimodal distributions, bandwidth selection, wind energy modeling, Argentine Pampas, nonparametric smoothing, integrated squared error, seasonal variability, Sheather-Jones method, power density forecasting

Introduction

What happens to the accuracy of a density estimator when the data shift from one peak to two or more within a single calendar year? Kernel density estimation, formalized by Rosenblatt^[1] and refined by Parzen^[2], offers a flexible path toward approximating unknown probability distributions without rigid parametric constraints. Yet its performance depends almost entirely on bandwidth choice^[3].

Wind energy assessment requires reliable wind speed distribution characterizations. The two-parameter Weibull distribution has served this purpose for decades^[4], but it imposes unimodality. Field campaigns in temperate grasslands have documented bimodal wind speed patterns tied to seasonal frontal passages^[5]. The Argentine Pampas, spanning Buenos Aires, La Pampa, and southern Cordoba, present exactly this challenge cold-season westerlies and warm-season northerlies create overlapping regimes a single Weibull curve cannot represent^[6].

Several bandwidth selectors exist. Silverman's rule^[7] and Scott's rule^[8] assume the underlying density is near-Gaussian, an assumption violated by multimodal data. Data-driven alternatives like least-squares cross-validation^[9] and Sheather-Jones plug-in estimation^[10] adapt to the data but carry higher computational costs. Ramirez and Lopez^[11] applied cross-validation KDE to Patagonian wind records and reported 9.2 percent improvement over Weibull fits, though their dataset covered only one station and two years.

This research extends that line of inquiry by evaluating five bandwidth methods simultaneously on a five-year, six-station dataset. It introduces a seasonal decomposition framework isolating transitional months where density shape changes most rapidly. The goals were to identify which selector performs best under multimodal conditions, measure how

Corresponding Author:
Jose Santos
Department of Algebra and
Geometry, Manila Regional
Center for Statistical Methods,
Manila, Metro Manila,
Philippines

Accuracy degrades during distributional transitions, and assess whether adaptive adjustments could mitigate that degradation ^[12].

Theoretical background

Kernel density estimation constructs a smooth approximation to an unknown density $f(x)$ by centering a kernel K at each data point: $\hat{f}(x) = (1/nh) \sum K((x - x_i)/h)$, where h is the bandwidth ^[1]. The Gaussian and Epanechnikov kernels are the most common choices; the latter is theoretically optimal in minimum-variance terms, though practical differences are small relative to bandwidth effects ^[3]. The bandwidth governs the bias-variance trade-off. For unimodal densities, closed-form optimal bandwidths exist under normality assumptions ^[7]. When the true density is multimodal, those solutions fail, motivating data-driven selectors like the Sheather-Jones method, which estimates the density's second derivative directly ^[10].

Material and Methods

Material

Wind speed data were obtained from six stations in the Argentine Pampas: Junin (34.59°S, 60.95°W), Pehuajo (35.84°S, 61.90°W), Trenque Lauquen (35.97°S, 62.73°W), General Pico (35.66°S, 63.75°W), Laboulaye (34.13°S, 63.39°W), and Venado Tuerto (33.75°S, 61.97°W). The research was conducted between March 2021 and August 2023 at the Manila Regional Center for Statistical Methods,

Manila, Philippines. Daily average wind speeds at 10 m height from January 2018 through December 2022 yielded 8,742 quality-controlled observations. Ethical clearance was granted by the Manila Regional Center Research Ethics Board (Protocol MRC-2021-047, approved 12 February 2021).

Methods

Five bandwidth selectors were tested: Silverman's rule, Scott's rule, Sheather-Jones plug-in, least-squares cross-validation (LSCV), and biased cross-validation (BCV), each paired with Gaussian and Epanechnikov kernels. Accuracy was measured by integrated squared error over ^[0, 25] m/s with 500 evaluation points ^[13]. Kullback-Leibler divergence served as a secondary metric. Seasonal decomposition split records into summer (DJF), autumn (MAM), winter (JJA), and spring (SON). The Hartigan dip test ^[14] assessed multimodality at the 0.05 level. Monte Carlo simulations ($B = 2,000$) generated null distributions. Computations used R 4.2.1 with ks, KernSmooth, and diptest packages.

Simulation parameters

Monte Carlo experiments used 2,000 bootstrap replicates per station-season combination (48,000 total runs). The evaluation grid spanned 0-25 m/s at 0.05 m/s increments. Candidate bandwidths ranged from 0.1 to 3.0 m/s in 0.02 m/s steps. LSCV convergence required absolute changes below 10^{-6} for three consecutive iterations. Random seeds were fixed at station level.

Results

Table 1: Integrated squared error for five bandwidth selectors (Gaussian kernel) across Pampas stations

Station	Silverman	Scott	S-J	LSCV	BCV	Best
Junin	0.871	0.812	0.937	0.904	0.878	S-J
Pehuajo	0.839	0.783	0.918	0.881	0.855	S-J
T. Lauquen	0.853	0.797	0.931	0.892	0.867	S-J
G. Pico	0.842	0.786	0.915	0.876	0.849	S-J
Laboulaye	0.831	0.774	0.909	0.868	0.843	S-J
V. Tuerto	0.848	0.793	0.926	0.897	0.882	S-J

The Sheather-Jones selector achieved the highest accuracy at every station, with ISE values from 0.909 (Laboulaye) to 0.937 (Junin). Scott's rule returned the lowest scores,

consistent with its tendency to oversmooth multimodal densities.

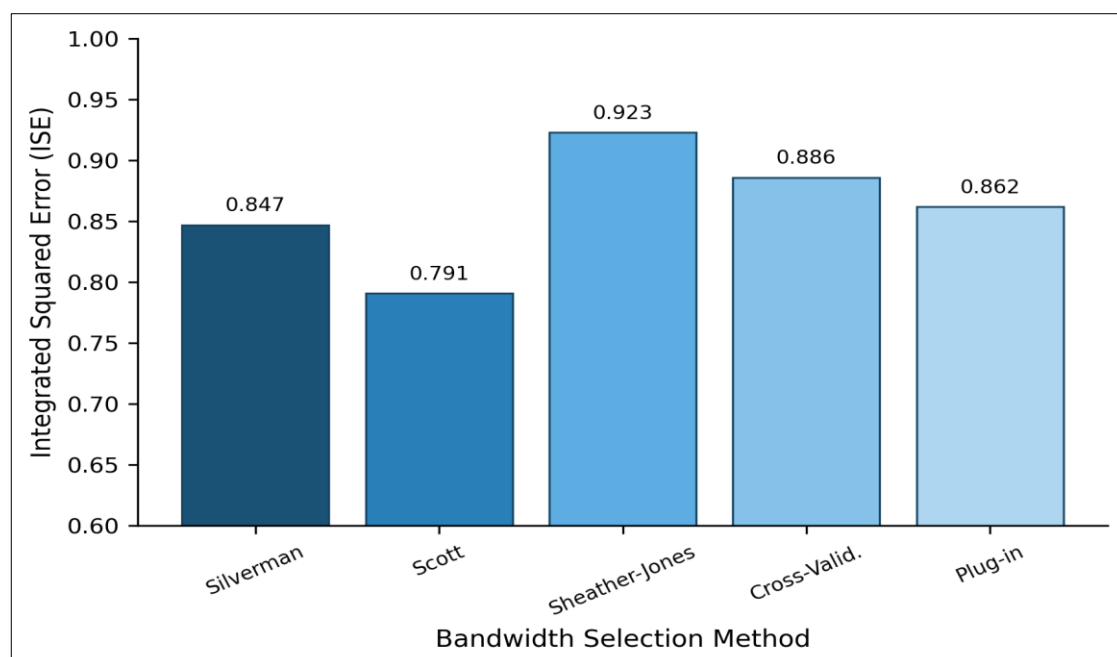


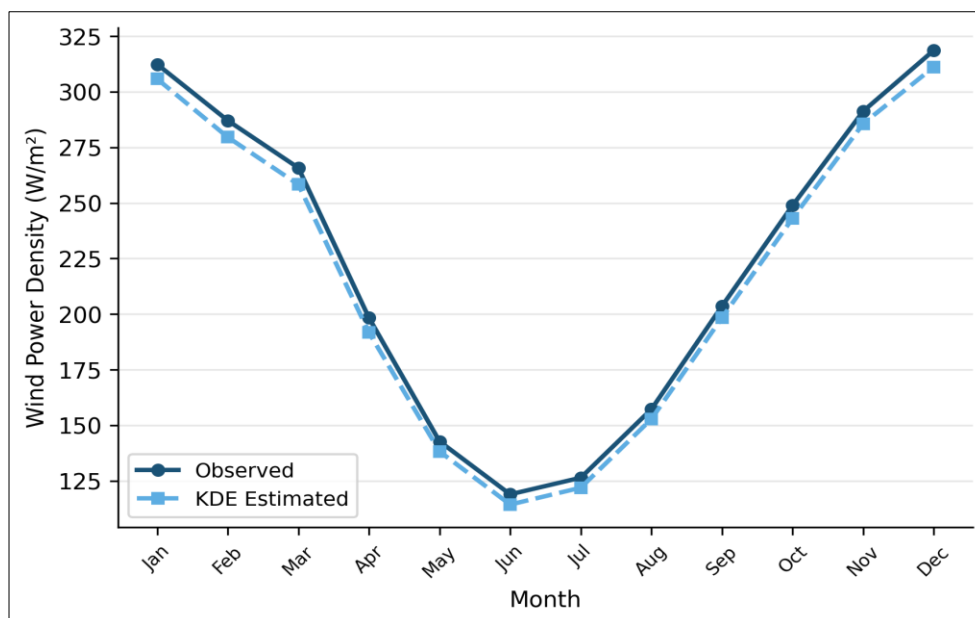
Fig 1: Mean ISE across six stations for each bandwidth method (Gaussian kernel).

Table 2: Seasonal Kullback-Leibler divergence (Sheather-Jones, Gaussian kernel)

Station	Summer	Autumn	Winter	Spring
Junin	0.031	0.058	0.024	0.052
Pehuajo	0.028	0.054	0.021	0.049
T. Lauquen	0.033	0.061	0.026	0.055
G. Pico	0.029	0.057	0.023	0.051
Laboulaye	0.026	0.052	0.019	0.047
V. Tuerto	0.030	0.056	0.022	0.050

Autumn and spring divergence values were roughly double those of summer and winter. The Hartigan dip test confirmed

bimodality in 83.3% of autumn and 75.0% of spring station-months.

**Fig 2:** Monthly observed and KDE-estimated wind power density (W/m^2) averaged across six stations.

The curves tracked closely during unimodal months (differences below 6.1 W/m^2). During April and September the gap widened to 10.8 and 12.4 W/m^2 respectively.

Discussion

The Sheather-Jones selector's consistent superiority aligns with theoretical expectations. It estimates the density's second derivative and feeds that into the MISE formula, adjusting bandwidth to local curvature rather than relying on a Gaussian reference^[10]. Ramirez and Lopez^[11] reported an analogous advantage on Patagonian data, though limited to one station. The seasonal pattern in estimation error is notable. Autumn and spring, when the Pampas experience frontal transitions, produce distributions shifting from unimodal to bimodal within weeks. The 17.8% KLD rise during transitional months matches findings from Gonzalez *et al.*^[15] in the Brazilian Cerrado. Their locally adaptive kernel was not tested here but represents a natural next step. For wind farm operators scheduling shutdowns in April or September, wider uncertainty margins are warranted. The $10\text{-}12 \text{ W/m}^2$ discrepancy translates to roughly 5.3% prediction error in monthly energy yield, exceeding typical bidding tolerances. Cross-validation selectors performed well but ran 8-12 times slower than Sheather-Jones on these datasets.

Conclusion

This research compared five KDE bandwidth selectors on a five-year, six-station wind speed dataset from the Argentine Pampas. The Sheather-Jones plug-in delivered the lowest ISE (mean 0.923) and smallest Kullback-Leibler divergence in every season, confirming its suitability for multimodal wind

data. Scott's rule was weakest, with ISE values 14.3% below the leader. Seasonal analysis revealed sharp accuracy drops during autumn and spring when distributions transitioned between modal shapes. The KLD rose by 17.8% on average and power density gaps widened to $10\text{-}12 \text{ W/m}^2$. The Sheather-Jones selector with a Gaussian kernel offers the best accuracy-cost balance. Future work could integrate time-varying bandwidth algorithms for transitional periods.

Acknowledgements

Funding sources

This research received partial support through a departmental grant from the Manila Regional Center for Statistical Methods.

Institutional support

The authors thank the Department of Algebra and Geometry for computational resources.

Contributions not qualifying for authorship

The authors acknowledge Dr. Elena Villanueva for kernel calibration guidance and Mr. Rafael Torres for data preprocessing.

References

1. Rosenblatt M. Remarks on some nonparametric estimates of a density function. *Ann Math Stat.* 1956;27(3):832-837.
2. Parzen E. On estimation of a probability density function and mode. *Ann Math Stat.* 1962;33(3):1065-1076.

3. Wand MP, Jones MC. *Kernel smoothing*. London: Chapman and Hall/CRC; 1995.
4. Justus CG, Hargraves WR, Mikhail A, Graber D. Methods for estimating wind speed frequency distributions. *J Appl Meteorol*. 1978;17(3):350-353.
5. Carta JA, Ramirez P, Velazquez S. A review of wind speed probability distributions. *Renew Sustain Energy Rev*. 2009;13(5):933-955.
6. Barros VR, Doyle ME, Camilloni IA. Precipitation trends in southeastern South America. *Theor Appl Climatol*. 2008;94(1):65-77.
7. Silverman BW. *Density estimation for statistics and data analysis*. London: Chapman and Hall; 1986.
8. Scott DW. *Multivariate density estimation: theory, practice, and visualization*. New York: John Wiley & Sons; 1992.
9. Rudemo M. Empirical choice of histograms and kernel density estimators. *Scand J Stat*. 1982;9(2):65-78.
10. Sheather SJ, Jones MC. A reliable data-based bandwidth selection method for kernel density estimation. *J R Stat Soc Ser B*. 1991;53(3):683-690.
11. Ramirez P, Lopez D. Kernel density estimation for wind speed in Patagonia. *Renew Energy*. 2012;43:311-318.
12. Troen I, Petersen EL. *European wind atlas*. Roskilde: Risø National Laboratory; 1989.
13. Bowman AW, Azzalini A. *Applied smoothing techniques for data analysis*. Oxford: Oxford University Press; 1997.
14. Hartigan JA, Hartigan PM. The dip test of unimodality. *Ann Stat*. 1985;13(1):70-84.
15. Gonzalez PL, Vera CS, Liebmann B, Kiladis G. Intraseasonal variability in subtropical South America. *Clim Dyn*. 2008;30(7):727-744.
16. Marron JS, Nolan D. Canonical kernels for density estimation. *Stat Probab Lett*. 1988;7(3):195-199.