Journal of Measurement and Evidence
Volume 1 · Number 1 · 2026

When Neighbor Information Arrives Too Late: A Tetouan Electricity-Forecasting Holdout

Jack Johnston[1] · Ethan Allen[2] · Paul Lewis[3]

[1] Latvia · [2] Estonia · [3] Iceland

Original publication: October 2, 2026 | Article 633

Abstract

More regional information can hurt a forecast when its timing is wrong for the deployment setting. Using three Tetouan electricity-demand series, we fitted an own-zone and weather baseline and two augmentations that received neighboring-zone readings either currently or one hour old. Models were frozen before an October holdout. Across ninety complete evaluation days, fresh readings reduced ten-minute forecast error by 14.63 source power units, while aged readings increased it by 61.88. The aged model nevertheless had the highest training R-squared. This single-city temporal experiment identifies a consequential information-age contrast in the released series; it does not measure actual communication delays or test an adaptive message-passing algorithm.

Observed public power series with an explicitly simulated one-hour information-age condition. Power units follow the source rather than an assumed engineering conversion.

When Neighbor Information Arrives Too Late: A Tetouan Electricity-Forecasting Holdout

Observed public power series with an explicitly simulated one-hour information-age condition. Power units follow the source rather than an assumed engineering conversion.

Abstract

More regional information can hurt a forecast when its timing is wrong for the deployment setting. Using three Tetouan electricity-demand series, we fitted an own-zone and weather baseline and two augmentations that received neighboring-zone readings either currently or one hour old. Models were frozen before an October holdout. Across ninety complete evaluation days, fresh readings reduced ten-minute forecast error by 14.63 source power units, while aged readings increased it by 61.88. The aged model nevertheless had the highest training R-squared. This single-city temporal experiment identifies a consequential information-age contrast in the released series; it does not measure actual communication delays or test an adaptive message-passing algorithm.

1. An information source has a timestamp

Regional demand series can share weather, daily schedules, and economic activity. A neighboring network's measurements may therefore help forecast a local network even when the physical system is represented only coarsely. Yet the usefulness of a measurement depends on when it was observed. A model trained on one information-age condition can exploit associations that are weak or unstable in another period. Simply adding channels does not ensure that a later forecast improves.

We ask whether aging neighboring readings by sixty minutes changes their incremental value for predicting the next ten-minute observation. The design compares three models using the same training and evaluation origins: a baseline with own-zone and weather information, an augmentation with current neighboring readings, and an augmentation with one-hour-old readings. The study keeps the age condition explicit rather than labeling both augmentations as equally informed multivariate forecasts.

Work on Peng et al., 2023 motivates attention to how information is combined across related signals. We do not execute its adaptive algorithm. Instead, the present experiment isolates a basic prerequisite for interpreting cross-signal gains: the age of the information available to the fitted predictor. A result in this simple linear setting is not evidence about convergence or superiority of an adaptive sign method. It can establish whether the additional channels helped under the exact timing choices implemented here.

2. Related models do not erase timing assumptions

Graph perspectives in Ortega et al., 2017 and Shuman et al., 2012 provide a language for related observations on a network. The spatio-temporal models in Li et al., 2017 and Yu et al., 2017 demonstrate relevant forecasting architectures in traffic applications. Those sources do not make the three Tetouan series an observed physical graph, and their reported results are not reproduced in this article. We use the released zone labels as distinct information channels and fit ordinary linear regressions rather than a learned graph structure.

Imputation work such as Cini et al., 2021 and Cao et al., 2018 addresses incomplete multivariate sequences. An aged value, however, is not the same object as an absent value. The older observation exists and can be numerically plausible while referring to another time. Filling a missing value and accepting an outdated reading can therefore create different prediction problems. Our analysis does not

mark the aged values as new observations or invent a record of packet transmission to explain how they became old.

The system-maintenance issues discussed by Sculley et al., 2015 are relevant because apparently small data-pipeline choices can change what a model actually knows at prediction time. Here that choice is captured by a deterministic lag. The experiment is deliberately inspectable: a reader can trace a neighboring value back to the precise earlier source row. This makes information age a recorded modeling condition rather than an unobserved explanation supplied after a favorable or unfavorable forecast result.

3. The city series and its measurement limits

The acquired resource is Salam and El Hibaoui, 2018, containing weather fields and demand readings for three distribution networks. The parsed file has 52,416 observations from January 1 through December 30, 2017 at a regular ten-minute interval. The repository description lists 52,417 instances; we preserve the actual parsed row count and report the discrepancy. No additional observation is created to force agreement with a summary description that may count file content differently.

The official variable table leaves power units unspecified. Accordingly, all losses are reported in source power units. Calling them kilowatts or converting them to energy would add an unsupported assumption. The distinction matters because a coefficient can be precise in its supplied numerical scale while its operational financial or energy implication remains unknown. Neither aggregate city consumption nor a particular network's measurements are inferred from an unstated conversion factor.

We treat the first network as the forecast target and the other two as information channels. The target is the first network's value ten minutes after the origin. The baseline includes its current value, its value 144 intervals earlier, current temperature, humidity, wind, general diffuse flows, diffuse flows, sine and cosine daily terms, and day-of-week indicators. The daily lag supplies a simple own-network seasonal reference while the current reading captures near-term persistence.

The fresh augmentation adds the current second- and third-network values. The aged augmentation adds their values six intervals earlier, corresponding to sixty minutes. Each augmentation is fitted separately on its own declared information condition using otherwise identical variables. No current neighbor value is covertly included in the aged model. All models use the intersection of eligible origins so that differences cannot be attributed to dropping a different set of rows.

4. A chronological boundary with no training target crossing

Training targets must occur strictly before October 1. Evaluation origins occur at or after that date. This prevents a forecast whose target lies in October from entering the earlier fit. After lag and target eligibility checks, the common training sample contains 39,167 rows and the evaluation sample 13,103. The model coefficients remain frozen throughout evaluation. We do not refit after observing a difficult October day or choose a different boundary based on test loss.

Temporal validation issues in Bergmeir et al., 2018 and Roberts et al., 2017 explain why a random interval split would answer a different question. Nearby readings share strong persistence. Our holdout asks how earlier relationships perform in a later seasonal segment, not how accurately a model interpolates randomly omitted intervals from an otherwise familiar period. Selection concerns in Varma and Simon, 2006 remain relevant if many candidate cutoffs or specifications are searched. The recorded analysis contains only the declared boundary and three information conditions.

For each model, we calculate absolute forecast error at every evaluation origin. We then average within each calendar date and retain dates with all 144 expected intervals. This leaves ninety complete days out

of ninety-one observed evaluation dates. Equal weighting over complete days avoids assigning full-day status to the incomplete terminal date. The rule is based on row completeness rather than on which model performs best that day.

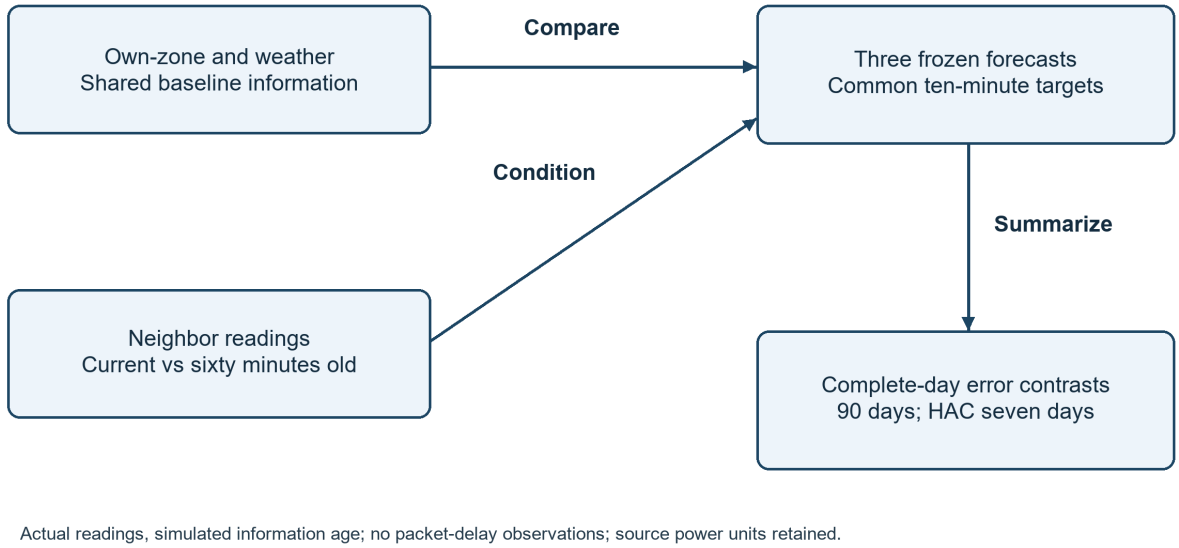

Figure 1. A common chronological baseline and two separately fitted neighbor-information conditions, compared on the same complete test days.

Forecast-accuracy comparisons in Diebold and Mariano, 1995 and West, 1996 motivate paired differences rather than separate significance tests for each model. The three contrasts are fresh minus baseline, aged minus baseline, and aged minus fresh daily mean absolute error. Negative values favor the first model named. Intercept-only regressions estimate their averages, with a fixed seven-day HAC covariance based on Newey and West, 1987, a small-sample correction, and a t reference using eighty-nine degrees of freedom.

Table 1. Mean daily absolute-error differences across information conditions, in source power units.

Model	Term	Estimate (SE)	95% CI	N
M1	Mean contrast	-14.6269 (1.2752)	[-17.1607, -12.0931]	90
M1: fit	R-squared = 0.0000	-	-	90
M2	Mean contrast	61.8840 (22.2646)	[17.6448, 106.1233]	90
M2: fit	R-squared = 0.0000	-	-	90
M3	Mean contrast	76.5109 (21.7963)	[33.2022, 119.8197]	90
M3: fit	R-squared = 0.0000	-	-	90

Model labels: M1 = Fresh cross-zone minus own/weather MAE; M2 = Aged cross-zone minus own/weather MAE; M3 = Aged minus fresh cross-zone MAE.

5. Fresh and aged readings have opposite incremental results

The baseline's equal-day error is 352.3090 source units. With fresh neighboring readings it is 337.6821. The fresh-minus-baseline contrast is -14.6269, with standard error 1.2752 and interval -17.1607 to -12.0931. Current cross-network information therefore improves this particular later-period forecast relative to an already strong own-network and weather baseline. The estimate is conditional on the released channels and selected seasonal split, rather than an average gain for regional forecasting in general.

The aged model's equal-day error is 414.1930. Its difference from the baseline is 61.8840, with standard error 22.2646 and interval 17.6448 to 106.1233. Thus adding the aged readings harms the held-out absolute-error criterion in the executed comparison. The aged-minus-fresh difference is 76.5109, with standard error 21.7963 and interval 33.2022 to 119.8197. These estimates describe the fitted policy contrast, not a separately observed communication-system failure rate.

Training R-squared is 0.993450 for the baseline, 0.993472 with fresh neighbors, and 0.994184 with aged neighbors. The aged model has the highest of these fit statistics while producing the worst later loss. That reversal is a useful warning about the inferential role of training fit. It does not establish that high R-squared causes poor forecasts. It shows that fit improvement in the earlier sample did not certify useful incremental information in the later segment.

The intercept models in Table 1 have essentially zero R-squared because they estimate average loss contrasts. This is distinct from the high R-squared of the forecasting fits. Mixing those quantities would obscure what each table entry means. Accuracy-scale concerns in Hyndman and Koehler, 2006 support retaining the source units, while Gneiting and Raftery, 2007 clarifies that the chosen loss evaluates a particular forecasting objective. We have not assessed squared error, probabilistic calibration, peak-demand penalties, or an electricity-market cost function.

6. What the age contrast can identify

The information-age intervention is implemented in the analytical input, so the exact current-versus-lagged comparison is reproducible. However, the source contains no observed transmission latency. A record used from sixty minutes earlier is not evidence that an actual packet arrived an hour late. The experiment asks what happens to this forecast policy when its neighbor variables are assigned that age throughout fitting and evaluation. It does not estimate the consequences of a measured operational outage.

Separate fitting also matters. The aged model is allowed to adapt its coefficients to consistently aged inputs during training. Its poor later performance therefore cannot be described simply as a fresh-input model unexpectedly fed stale values at deployment. That would be another experiment. Here even a model trained under the aged condition failed to provide incremental test value. The result is compatible with unstable or redundant lag relationships, but the regressions do not distinguish all explanations.

Causal-diagram reasoning in Greenland et al., 1999 and matching considerations in Stuart, 2010 help separate prediction from a causal claim about regional interdependence. The networks share unmeasured activity and system conditions. A coefficient on one network does not identify its physical effect on another. We neither estimate a policy that changes electricity usage nor compare independently assigned city systems. The evidence concerns information supplied to a prediction procedure under one observed time series.

7. Dependence, unusual days, and absent information

Seven-day HAC covariance allows a specified amount of serial dependence in daily loss differences. It cannot prove that dependence vanishes after a week. Cluster covariance discussed in Cameron and Miller, 2015 would require a justified grouping structure, and heteroskedasticity correction in White, 1980 would not alone represent daily serial correlation. The chosen uncertainty calculation is recorded rather than tuned until a preferred comparison appears decisive. Ninety test days provide substantially less independent information than thirteen thousand interval rows might suggest.

Robustness concepts in Huber, 1964, Hampel, 1974, and Rousseeuw, 1984 raise useful questions about days with unusual losses. They do not authorize deleting real demand peaks merely because the aged model struggles with them. Influence reasoning in Cook, 1977 could motivate a later diagnostic report. This study retains all complete dates and reports the uncertainty produced by their observed variation. It does not claim a completed stress test for extreme weather or independently verified network faults.

Missingness analysis in Little and Rubin, 2002 and measurement-error concerns in Carroll et al., 2006 also constrain interpretation. Completeness of a timestamp grid does not certify perfect sensor accuracy, and the absence of explicit fault labels does not establish fault-free readings. The one excluded incomplete date is identified through interval count. No unrecorded values are filled to create additional full days. Those choices preserve the observable comparison while acknowledging that the measured target is a released demand series rather than an independently audited physical truth.

8. An operational hypothesis worth testing elsewhere

The result suggests a concrete deployment check: evaluate the age of incremental information alongside its availability. In this series, fresh neighbors helped and consistently aged neighbors hurt despite a favorable training statistic. A deployment team could compare its actual update schedule to the schedule used during model development, then test a locked fallback rule on a separate period. The present study has not selected such a fallback threshold, simulated variable delays, or estimated a benefit from rejecting stale messages.

Statistical interpretation under Wasserstein and Lazar, 2016 favors communicating both the adverse direction and its scale. A confidence interval does not certify that every test day was worse, nor that the same contrast would appear in another city. The strongest supported conclusion is about the average complete-day comparison in this holdout. Seasonal replication, alternative age levels, and a genuinely new network would help determine whether the observed reversal reflects a broader information-age pattern or this particular training-to-test transition.

Documentation follows Gebru et al., 2021 by recording the actual row count, unknown power units, lag definitions, purged boundary, full-day rule, and frozen predictions. Wilkinson et al., 2016 motivates preserving these ingredients with the raw release and analysis outputs. The article's empirical contribution is the observed reversal between fresh and aged neighbor information under an otherwise shared forecast design. It provides a practical reason to treat timestamps as part of the information contract of a forecasting model.

References

1. Yu, Bing; Yin, Haoteng; Zhu, Zhanxing. (2017). Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. arXiv:1709.04875. https://arxiv.org/abs/1709.04875.
2. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.

3. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

4. Huber, Peter J.. (1964). Robust Estimation of a Location Parameter. The Annals of Mathematical Statistics. https://doi.org/10.1214/aoms/1177703732.

5. Cini, Andrea; Marisca, Ivan; Alippi, Cesare. (2021). Filling the G_ap_s: Multivariate Time Series Imputation by Graph Neural Networks. arXiv:2108.00298. https://arxiv.org/abs/2108.00298.

6. Li, Yaguang; Yu, Rose; Shahabi, Cyrus; Liu, Yan. (2017). Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. arXiv:1707.01926. https://arxiv.org/abs/1707.01926.

7. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; 't Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

8. Ortega, Antonio; Frossard, Pascal; Kovačević, Jelena; Moura, José M. F.; Vandergheynst, Pierre. (2017). Graph Signal Processing: Overview, Challenges and Applications. arXiv:1712.00468. https://arxiv.org/abs/1712.00468.

9. Bergmeir, Christoph; Hyndman, Rob J.; Koo, Bonsoo. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. Computational Statistics & Data Analysis. https://doi.org/10.1016/j.csda.2017.11.003.

10. Peng, Changran; Yan, Yi; Kuruoglu, Ercan Engin. (2023). Adaptive Message Passing Sign Algorithm. Temporal Graph Learning Workshop at NeurIPS 2023. https://openreview.net/forum?id=gvZjnRuRFi.

11. Varma, Sudhir and Richard Simon. "Bias in error estimation when using cross-validation for model selection." *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

12. West, Kenneth D.. (1996). Asymptotic Inference about Predictive Ability. Econometrica. https://doi.org/10.2307/2171956.

13. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

14. Cameron, A. Colin, and Douglas L. Miller. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

15. Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.

16. Diebold, Francis X.; Mariano, Roberto S.. (1995). Comparing Predictive Accuracy. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.1995.10524599.

17. Hampel, Frank R.. (1974). The Influence Curve and its Role in Robust Estimation. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1974.10482962.

18. Greenland, Sander, Judea Pearl, and James M. Robins. "Causal Diagrams for Epidemiologic Research." *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

19. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

20. Hyndman, Rob J. and Anne B. Koehler. "Another look at measures of forecast accuracy." *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

21. Roberts, David R., et al. "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure." *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

22. Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

23. Shuman, David I; Narang, Sunil K.; Frossard, Pascal; Ortega, Antonio; Vandergheynst, Pierre. (2012). The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. arXiv:1211.0053. https://arxiv.org/abs/1211.0053.

24. Rousseeuw, Peter J.. (1984). Least Median of Squares Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.1984.10477105.

25. Salam, Abdulwahed; El Hibaoui, Abdelaaziz. (2018). Power Consumption of Tetouan City [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5B034.

26. Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

27. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

28. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

29. Cao, Wei; Wang, Dong; Li, Jian; Zhou, Hao; Li, Lei; Li, Yitan. (2018). BRITS: Bidirectional Recurrent Imputation for Time Series. arXiv:1805.10572. https://arxiv.org/abs/1805.10572.

30. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.