Journal of Transfer and Generalization
Volume 1 · Number 1 · 2026

Later-Period Gas-Sensor Calibration: A Small Robust-Fitting Gain Without a Reliable Disagreement Alarm

Andrew Gardner[1] · Luke Bailey[2] · Ryan King[3]

[1] Serbia · [2] Montenegro · [3] Romania

Original publication: November 1, 2026 | Article 622

Abstract

Calibration can improve while a proposed warning signal remains uninformative. Using an Italian multisensor recording, we fitted ordinary and Huber regressions before January 2005 and compared their later carbon-monoxide errors. Across 88 adequately observed test days, Huber fitting reduced equal-day mean absolute error by 0.0105 mg/m3, with a seven-lag HAC interval of 0.0035 to 0.0174. A disagreement score constructed from training-period sensor correlations did not reliably locate that advantage: its slope interval included zero. The result supports a modest retrospective fitting benefit on this recording, while leaving the proposed operational alarm unvalidated.

Study-level research draft based on public measurements. Author verification and submission review remain pending.

Later-Period Gas-Sensor Calibration: A Small Robust-Fitting Gain Without a Reliable Disagreement Alarm

Study-level research draft based on public measurements. Author verification and submission review remain pending.

Abstract

Calibration can improve while a proposed warning signal remains uninformative. Using an Italian multisensor recording, we fitted ordinary and Huber regressions before January 2005 and compared their later carbon-monoxide errors. Across 88 adequately observed test days, Huber fitting reduced equal-day mean absolute error by 0.0105 mg/m3, with a seven-lag HAC interval of 0.0035 to 0.0174. A disagreement score constructed from training-period sensor correlations did not reliably locate that advantage: its slope interval included zero. The result supports a modest retrospective fitting benefit on this recording, while leaving the proposed operational alarm unvalidated.

1. The calibration decision and the warning decision

An environmental instrument can generate several internally coherent electrical responses while its concentration estimate is inaccurate. Conversely, a disagreement among channels can accompany a change in pollutant mixture rather than a malfunction. These possibilities make two practical decisions distinct. One concerns the loss function used to fit a calibration. The other concerns the statistic that should warn an operator that the selected calibration is becoming relatively disadvantageous. A favorable average comparison does not automatically supply such a warning statistic.

The changing monitoring landscape described by Snyder et al., 2013 makes this distinction useful: more accessible measurements increase the need to characterize their quality. The caution advanced by Lewis and Edwards, 2016 also concerns validation rather than the mere availability of a sensor reading. Our analysis addresses a limited version of that problem with an existing public recording. It asks whether a robust calibration has a later-period error advantage and whether a sensor-network disagreement summary identifies days on which that advantage is larger.

The UCI Air Quality release supplies simultaneous chemical-sensor responses and reference-analyzer measurements (Vito, 2008). Its associated calibration study examined benzene estimation and the persistence of calibration performance (De Vito et al., 2008). We use the recorded carbon-monoxide reference instead. That choice does not recreate the original neural calibration or transfer its conclusions between chemical species. It provides a transparent setting in which to separate a fitted concentration estimate from a proposed diagnostic signal.

Our research question is therefore two-part: does Huber fitting improve later daily calibration error relative to ordinary least squares, and does training-defined graph disagreement explain the size of that relative improvement? The second question is deliberately about the difference between two errors. A detector of generally difficult hours need not be a detector of when one fitting rule is preferable. Confusing those objectives would make an attractive diagnostic plot answer the wrong operational question.

2. Why a robust fit need not create a useful alarm

Ordinary least squares gives large residuals substantial influence during fitting. Huber estimation limits the residual contribution beyond its transition point (Huber, 1964). Influence-function analysis explains why sensitivity to unusual observations depends on the estimating rule (Hampel, 1974), while the broader robust-regression treatment in Rousseeuw, 1984 emphasizes that robustness is a property of a specified procedure and contamination setting. These principles motivate a comparison; they do not establish that this particular recording contains a known contamination process.

Graph representations add another choice. The graph-signal accounts in Shuman et al., 2012 and Ortega et al., 2017 relate multivariate observations to an explicitly defined network. Here the channels, rather than monitoring locations, form the nodes. Their training-period correlation determines a small signed graph. An inverse response can be consistent with the network once its sign is respected. A large unsigned difference between two standardized channels would otherwise label an expected inverse relationship as disagreement.

The target contribution, Yan et al., 2021, places normalized least-mean-p estimation within graph signal processing. It motivates attention to both the error criterion and the representation of multichannel observations. Our study addresses the empirical decision that precedes adoption of a more specialized estimator: whether a simple robust fit transfers better, and whether a cheap graph summary can identify its relative advantage. Huber regression with standardized predictors is a different procedure from the named adaptive graph algorithm. The present calculations establish no performance result for that algorithm.

Long-term calibration work such as Malings et al., 2019 makes temporal performance a concrete evaluation concern. We consequently define P1 as the mean later-day advantage of Huber over ordinary fitting, measured by ordinary-minus-Huber absolute error. P2 concerns the slope of this advantage on training-defined disagreement. A positive P1 with an uncertain P2 would support a small average benefit while withholding support from a switching alarm. Figure 1 connects these propositions to their measured variables.

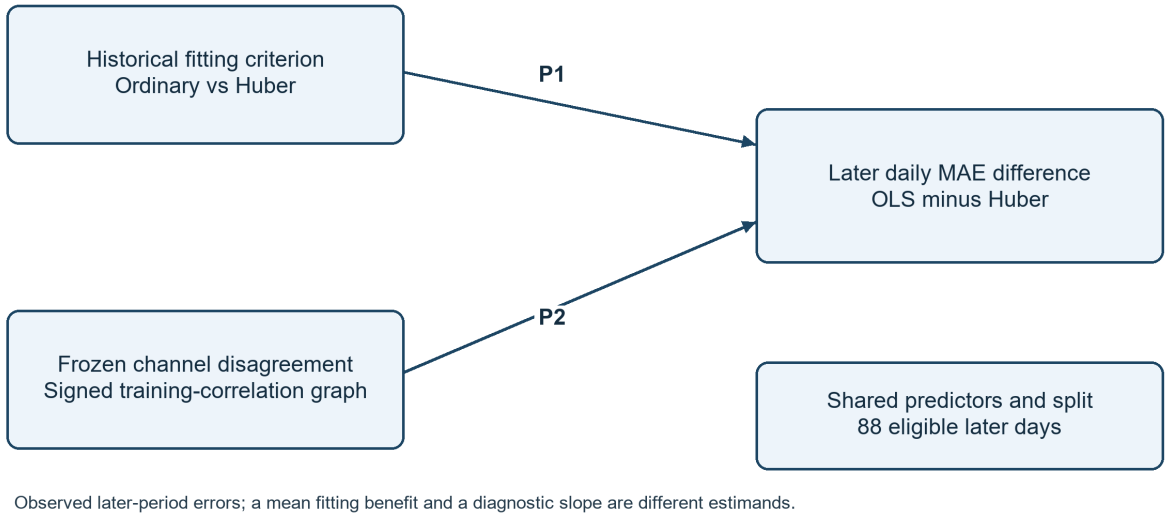


Figure 1. The two calibrations share training rows and seven predictors. P1 summarizes their later daily loss difference; P2 relates that difference to frozen sensor disagreement. Neither path identifies a causal sensor-drift effect.

3. Reading the released file before defining the sample

The CSV contains 9,471 physical rows, including trailing records without dates. We retain 9,357 dated hourly records, spanning 10 March 2004 through 4 April 2005. This observed coverage is reported directly from the file. It should not be silently replaced by a rounded date range from a repository description. The recorded hourly timestamps are unique, and the source archive is retained with its download record and checksum.

The five electrical-response columns are PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), and PT08.S5(O3). Temperature and relative humidity complete the seven predictors. The dependent calibration variable is CO(GT), whose unit follows the repository variable description, mg/m3. We convert the source's minus-200 missing-value code to missing before calculating any correlations, means, or regressions. Treating that code as an extreme measurement would manufacture precisely the large residuals that a robust method might appear to handle well.

Requiring the reference and all seven predictors leaves 7,344 usable hours. Missing-data guidance makes the resulting target population conditional on this selection (Little and Rubin, 2002). We do not impute a reference concentration or evaluate predictions against an invented target. Measurement-error considerations also remain relevant after complete-case selection, because the retained predictors and reference are recorded instruments rather than error-free latent quantities (Carroll et al., 2006).

The calendar boundary is fixed at 1 January 2005. There are 5,332 complete training hours and 2,012 later test hours. Training observations end on 28 December because the intervening source records do not meet the complete-case rule. Test observations extend from 1 January through 4 April. Both calibrations use exactly the same eligible training rows, and no test response determines predictor scaling or graph construction.

4. Two calibrations and one frozen network descriptor

We standardize each predictor using its training mean and population standard deviation. Both regressions include an intercept. Ordinary least squares minimizes squared reference residuals. The robust fit uses the Huber transition constant 1.345 and converges after 26 iterations. The robust procedure changes the fitting criterion while keeping the measured inputs, sample, and linear predictor specification fixed. We fit no nonlinear calibration, select no sensor subset using test errors, and apply no artificial spikes.

To construct the channel graph, each sensor selects its two other channels with the largest absolute training correlations. The undirected union has seven edges. Each edge keeps the sign and absolute magnitude of its training correlation. For an hourly observation, we compare the two standardized endpoint responses after aligning the second response to that correlation sign, square their difference, and take the correlation-weighted average across edges. The resulting roughness is then standardized against its training distribution.

This construction is reproducible but does not reveal a physical causal network. A directed-acyclic interpretation would require assumptions beyond a correlation graph (Greenland et al., 1999). Similarly, transport arguments concern whether an analysis remains applicable under a changed population, not whether a calendar split alone establishes exchangeability (Stuart, 2010). The graph can summarize an observed departure from a historical channel relationship without identifying its chemical or electronic cause.

At each test hour, we retain the actual reference and calculate both absolute errors. Their difference is positive when Huber is better. We then average the difference and standardized roughness within day.

Days need at least 12 complete hours. This leaves 88 of 90 observed test days, representing 2,008 hours. The four hours on two sparsely observed days remain in the supplementary hourly file but not the primary day-level regressions. Equal day weights define the primary estimand.

5. Estimation and interpretation of uncertainty

P1 uses an intercept-only regression of daily error difference. P2 adds the day's mean standardized graph roughness. Both use heteroskedasticity and autocorrelation consistent covariance with seven retained-day lags and a finite-sample adjustment. The covariance approach follows the logic of Newey and West, 1987, extending the concern for unequal residual variability developed in White, 1980. Missing calendar days mean that a retained-observation lag need not always equal the same number of consecutive calendar days.

The uncertainty concerns a hypothetical series of comparable days conditional on the fitted historical calibrations. It does not include every source of training-sample variation or uncertainty about alternative calibration choices. The forecast-evaluation perspective in West, 1996 helps distinguish a realized loss contrast from a complete account of estimated-model uncertainty. With 88 days from one deployment, this remains a narrow retrospective comparison.

There is an additional distinction between an average improvement and an operational decision rule. An operator considering whether to switch calibration needs a prediction of relative future loss, together with the cost of switching or recalibration. The P2 outcome is chosen to address that relative-loss problem. Regressing either method's absolute error on roughness instead would ask whether a day is generally difficult. Both calibrations could become less accurate on a difficult day while their difference stays unchanged. A good predictor of difficulty would then provide little guidance about which calibration to use.

The channel graph also contains substantive information that a generic variance statistic would miss. For example, the training correlation between S2 and S3 is about -0.817, whereas that between S2 and S5 is about 0.895. Both can represent stable associations despite having opposite signs. Sign alignment places them on a comparable disagreement footing before squaring their residual difference. Updating those signs or weights with later observations would produce a different descriptor and a different evaluation protocol. Retaining the historical graph therefore makes the present diagnostic question specific, while exposing its dependence on the earlier correlation regime. It does not imply that keeping an old graph is the best deployment policy.

Absolute-error summaries answer a particular loss question. Forecast-accuracy guidance in Hyndman and Koehler, 2006 discourages treating an unnamed error measure as universal performance. Proper-score arguments likewise require the evaluation target to be specified (Gneiting and Raftery, 2007); our point-prediction comparison supplies no predictive distribution or calibrated concentration interval. We report the retained-day average separately from the average over all complete test hours because these weighting rules differ.

The executed estimates appear in Table 1. An intercept-only regression has essentially zero R-squared by construction. That value does not make its mean contrast uninformative, and a nonzero R-squared in P2 does not validate a diagnostic alarm. The models estimate different features of the same saved daily observations.

Table 1. Daily later-period contrasts. Outcome: ordinary-minus-Huber MAE in mg/m3. Both models include an intercept; P2 adds mean training-standardized graph roughness. HAC covariance uses seven retained-day lags and t reference intervals. The primary sample contains 88 days with at least 12 complete hours.

Model	Term	Estimate (SE)	95% CI	N
Mean held-out OLS minus Huber MAE	Mean OLS minus Huber MAE: P1	0.0105 (0.0035)	[0.0035, 0.0174]	88
Mean held-out OLS minus Huber MAE: fit	R-squared = 0.0000	-	-	88
Loss contrast versus graph disagreement	Training-standardized disagreement: P2	0.0107 (0.0087)	[-0.0065, 0.0279]	88
Loss contrast versus graph disagreement: fit	R-squared = 0.0367	-	-	88

6. A positive average contrast and an uncertain localization

The equal-day mean absolute error is 0.3743 mg/m3 for ordinary fitting and 0.3639 for Huber. Their difference is 0.01045, with standard error 0.00351 and a 95% interval from 0.00347 to 0.01744. Averaging over every complete test hour instead gives 0.3666 and 0.3566, respectively. These are compatible descriptions of different weighting schemes rather than contradictory estimates of one denominator.

The P2 roughness coefficient is 0.01071, with standard error 0.00867 and interval from -0.00653 to 0.02794. Its sign is positive, but its uncertainty does not establish that higher disagreement reliably identifies a larger robust-fitting advantage. The model explains approximately 3.67% of the observed variation in the daily contrast. Neither this fit nor the mean benefit supplies a usable threshold, false-alarm rate, sensitivity, or economic switching rule.

Agreement analysis provides a useful caution here: related measurements need not be interchangeable (Bland and Altman, 1986). A genuine repeatability study would also need an appropriate replication structure before invoking reliability measures such as those reviewed by Koo and Li, 2016. Our graph roughness has neither repeated independently calibrated channel networks nor operator-confirmed fault labels. It should therefore be described as a candidate descriptor rather than an established instrument-health measure.

7. What a follow-up validation would need

Temporal validation retains the ordering of deployment conditions, consistent with the structured-validation concerns in Roberts et al., 2017. It nevertheless samples one historical transition. Repeatedly searching calendar boundaries or diagnostic thresholds on these same test errors would make a favorable result easier to obtain; the model-selection analysis in Varma and Simon, 2006 explains why such reuse affects estimated performance. The bootstrap treatment in Bühlmann, 2002 suggests ways to respect dependent observations in a further uncertainty analysis, but no bootstrap distribution is claimed here.

Unusual days might influence the disagreement slope. The regression-diagnostic logic in Cook, 1977 motivates examining them during a new, explicitly exploratory audit rather than discarding days merely because they weaken the desired relationship. If a later study screens many chemical channels and warning definitions, the testing family should be declared. The distinct dependence assumptions behind Benjamini and Hochberg, 1995 and Benjamini and Yekutieli, 2001 matter for that extension; our current

two propositions are reported with their effect estimates and intervals, without a battery of selectively emphasized significance tests.

The source, variable rules, temporal boundary, graph edges, hourly errors, and retained-day table are saved locally. Documentation principles in Gebru et al., 2021 and the findability and reuse framework in Wilkinson et al., 2016 guide that record. The repository currently presents both a general attribution license and a research-use restriction in its descriptive text; both are retained for author review before broader reuse. Scientific reproducibility does not settle a conflicting source-use statement.

8. Conclusion

For this recorded later period, Huber fitting has a small positive average advantage over an otherwise matched ordinary linear calibration. The proposed graph-disagreement descriptor does not reliably localize that advantage at the day level. These findings answer different decisions. They justify preserving the observed robust-fitting benefit while leaving an operational alarm, any causal account of drift, and the performance of a specialized adaptive graph estimator open to further investigation.

References

1. Vito, Saverio. (2008). Air Quality [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C59K5F.
2. De Vito, S.; Massera, E.; Piga, M.; Martinotto, L.; Di Francia, G.. (2008). On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. Sensors and Actuators B: Chemical. https://doi.org/10.1016/j.snb.2007.09.060.
3. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
4. Yan, Yi; Adel, Radwa; Kuruoglu, Ercan E.. (2021). Adaptive Normalized LMP Estimation for Graph Signal Processing. 2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP), 1-6. https://doi.org/10.1109/mlsp52302.2021.9596181.
5. 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.
6. Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.
7. Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.
8. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.
9. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
10. Gebru, Timnit, et al.. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
11. 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.
12. Benjamini, Yoav and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.

13. 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.

14. Lewis, Alastair; Edwards, Peter. (2016). Validate personal air-pollution sensors. Nature. https://doi.org/10.1038/535029a.

15. 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.

16. 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.

17. Snyder, Emily G.; Watkins, Timothy H.; Solomon, Paul A.; Thoma, Eben D.; Williams, Ronald W.; Hagler, Gayle S. W.; Shelow, David; Hindin, David A.; Kilaru, Vasu J.; Preuss, Peter W.. (2013). The Changing Paradigm of Air Pollution Monitoring. Environmental Science & Technology. https://doi.org/10.1021/es4022602.

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

20. 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.

21. 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.

22. 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.

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

24. 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.

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

26. 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.

27. Malings, Carl; Tanzer, Rebecca; Hauryliuk, Aliaksei; Kumar, Sriniwasa P. N.; Zimmerman, Naomi; Kara, Levent B.; Presto, Albert A.; , . (2019). Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. Atmospheric Measurement Techniques. https://doi.org/10.5194/amt-12-903-2019.

28. 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.

29. Bühlmann, Peter. (2002). Bootstraps for Time Series. Statistical Science. https://doi.org/10.1214/ss/1023798998.

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