

Validation Protocol and Reported Forecast Accuracy: Holdout, Time-Series Cross-Validation and Walk-Forward Compared

Ofieroher Ufuoma Earnest

Department of Finance and Banking, Faculty of Management Sciences, University of Port Harcourt, Nigeria

ORCID: <https://orcid.org/0009-0002-7468-8082>

Corresponding Author

Abstract: Reported forecast accuracy depends on the validation protocol used to compute it, and papers rarely state which they used. This paper holds the model, features, hyperparameter grid and dataset fixed and varies only the protocol. A support vector regression forecasting Nigerian FDI one quarter ahead, evaluated on 13 held-out quarters from 2021Q3 to 2024Q3, returns RMSE 902.72 under a single holdout with ordinary three-fold cross-validation for tuning, 901.95 under the same holdout with time-series cross-validation, and 653.34 under expanding-window walk-forward evaluation. Out-of-sample R-squared moves from -1.90 to -0.52. The direction here runs opposite to the standard warning: the naive protocol understated accuracy here, it did not overstate it. Switching the tuning procedure from ordinary to time-aware cross-validation changed almost nothing (902.72 against 901.95), which locates the effect in the evaluation step and not the tuning step. We explain the mechanism, which is evaluation variance rather than leakage bias, and argue that this makes protocol disclosure more important, not less, because the direction of the effect cannot be inferred from the reported number.

Keywords: Validation Protocol; Walk-Forward Evaluation; Forecast Accuracy; Evaluation Variance.

JEL Classification: C52; C53; C22.

1. INTRODUCTION

An accuracy figure is a property of a model evaluated under a protocol, not of the model alone. The time-series forecasting literature has long warned that random cross-validation can overstate accuracy for dependent data (Bergmeir & Benítez, 2012; Tashman, 2000). What is done less often is to hold everything else fixed and measure how much the number moves when only the protocol changes.

1.1 What this paper contributes

- It isolates protocol choice as the only varying factor, holding model class, features, search grid and data constant.
- It separates the tuning protocol from the evaluation protocol and shows the effect sits almost entirely in the latter.
- It documents a case where the effect runs opposite to the standard expectation, and identifies why.
- It reports per-fold validation scores instead of only their mean.

2. LITERATURE REVIEW

2.1 Cross-validation with dependent data

Standard k-fold cross-validation assumes exchangeable observations, and applied to time series it lets a model train on the future of its own test fold (Bergmeir & Benítez, 2012). Bergmeir, Hyndman and Koo (2018) refine this, showing ordinary cross-validation can be valid for purely autoregressive models with uncorrelated errors, a qualification that does not extend to the mixed-predictor setting used here. Tashman (2000) sets out rolling-origin evaluation and argues for multiple origins over single splits, because a single split confounds model quality with the peculiarities of one boundary.

2.2 Two different problems

Tashman's argument is central to what this paper finds, and it concerns something different from leakage. Leakage is a bias: it pushes estimates optimistically in a predictable direction. Evaluation variance is not a bias at all. A single holdout estimate is one draw from a sampling distribution, and in short samples that distribution is wide, so the estimate can land far from the expected out-of-sample error in either direction depending on what the particular window contains. The

common framing that time-aware validation always gives more honest and more pessimistic numbers runs these two together.

3. METHOD

The model is a support vector regression predicting FDI one quarter ahead from six core macroeconomic predictors. Model class, feature set, hyperparameter grid and data are identical across all three conditions. Only the protocol varies. The dataset was originally assembled as 144 quarterly observations from 1990Q1. Everything before 2008 proved to be linearly interpolated from annual figures: FDI, GDP, the exchange rate, inflation and the trade series each change by a constant amount within the year and jump at the boundary. That period has been dropped. Trade openness was also excluded, because the available series divides nominal trade by real GDP and tests as I(2). GDP is rebased in 2025 with an unadjusted splice, so 2025 is dropped too. The working sample is 2008Q2 to 2024Q4, 67 quarters.

- Protocol A: hyperparameters tuned by ordinary three-fold cross-validation on the training 80%, then one model fitted once and evaluated once on the final 20%.
- Protocol B: hyperparameters tuned by four-fold expanding-window cross-validation, then the same single fit and single evaluation. This isolates the tuning step.
- Protocol C: the model is refitted at each step on all data available to that point and predicts only the next quarter, repeated across the holdout.

All three use the same grid (C in {1, 10, 100, 1000}, epsilon in {0.01, 0.1, 1}, gamma in {scale, auto}) with scaling fitted only on training data inside each protocol.

3.1 A note on comparing R-squared

MAE and RMSE are directly comparable here, because all three protocols predict the same 13 quarters. Out-of-sample R-squared is less comparable: under A and B it references the variance of one fixed block evaluated in one shot, while under C it is computed across sequentially updated one-step predictions. Part of any R-squared gap reflects a change in what the benchmark refers to. The comparison below leads with RMSE for that reason.

4. RESULTS

Table 1. The same model under three protocols, test period 2021Q3–2024Q3 (n = 13)

Protocol	Tuning	Evaluation	MAE	RMSE	R ²
A	Ordinary 3-fold CV	Single fit, single prediction	774.93	902.72	-1.8986
B	4-fold TimeSeriesSplit	Single fit, single prediction	774.12	901.95	-1.8937
C	4-fold TimeSeriesSplit	Expanding-window walk-forward	514.16	653.34	-0.5183

A and B differ by 0.77 in RMSE. C improves on both by about 28%. The whole effect sits in the evaluation step.

Table 2. Per-fold validation RMSE within the tuning procedure

Fold	Validation RMSE
Fold 1 (earliest)	721.32
Fold 2	783.31
Fold 3	803.41
Fold 4 (latest)	457.13
Mean	691.29

Spread across folds is 1.8x. On the earlier interpolated sample the same procedure produced a spread of more than tenfold, which was itself an artefact of the interpolation.

5. DISCUSSION

Protocols A and B differ only in how hyperparameters were tuned, and they return RMSE of 902.72 and 901.95. For this model and dataset the choice between ordinary and time-aware cross-validation for tuning made no practical difference once the final evaluation was a single holdout prediction. That is a useful negative result, because it locates the source of the larger gap precisely.

The larger gap is between A and B on one side and C on the other, and its direction is the opposite of what the standard warning predicts. The literature's concern is that naive validation inflates apparent accuracy through leakage. Here the naive protocol reported worse accuracy, with RMSE falling from about 902 to 653 when the evaluation moved to walk-forward. Nothing about leakage explains that.

The mechanism is evaluation variance and it is specific to this sample. Walk-forward refits on an expanding information set and evaluates across 13 separate one-step predictions. The single-shot protocol fits once on the first 49 observations and tests that one model against all 13 remaining quarters at once. The holdout here spans 2021Q3 to 2024Q3, covering the COVID recovery and the 2023 exchange-rate unification, and FDI turns negative within it. A model frozen before that period has no chance to adapt; walk-forward gets to learn from each quarter as it arrives.

This strengthens the case for disclosure rather than weakening it. If protocol choice always biased results one way, a reader could adjust mentally. Because the direction depends on where the difficult observations fall relative to the split, it cannot be inferred from the reported number. A reader given only $RMSE = 653.34$ has no way to know whether the protocol flattered the model or penalised it.

The per-fold spread of 1.8x is modest and worth contrasting with the earlier version of this analysis, which reported a spread above tenfold on the interpolated sample. That instability was a property of the data construction, not of the tuning procedure. Reporting per-fold scores would have surfaced it.

6. CONCLUSION

Holding model, features, grid and data fixed, reported RMSE for one-quarter-ahead Nigerian FDI ranged from 902.72 to 653.34 depending only on the validation protocol, and the direction ran opposite to the usual expectation. The tuning protocol made no practical difference; the evaluation protocol made all of it. Accuracy figures for short macroeconomic series are not interpretable without an explicit statement of protocol, and walk-forward evaluation remains preferable because it matches deployment conditions, not because it is systematically more conservative.

7. LIMITATIONS

One model class, one dataset, one split boundary. Whether the direction of the effect generalises requires replication. No formal test of whether the differences exceed sampling variability. A Diebold-Mariano test on the paired errors would establish this.

Only three protocols. Blocked cross-validation, rolling-window evaluation and purged cross-validation with embargo periods are further options.

The proposed mechanism, that the difficult observations sit inside the holdout, is inferred and not tested. Repeating the comparison across several split boundaries would test it directly.

REFERENCES

- [1]. Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, 191, 192–213.
- [2]. Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120, 70–83.
- [3]. Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263.
- [4]. Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.
- [5]. Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
- [6]. Tashman, L. J. (2000). Out-of-sample tests of forecasting accuracy. *International Journal of Forecasting*, 16(4), 437–450.