Forecast Evaluation

BSAD 8310: Business Forecasting — Lecture 6

Department of Economics
University of Nebraska at Omaha

Spring 2026

1 The Evaluation Problem

2 Error Metrics

3 Walk-Forward Validation

4 The Diebold-Mariano Test

5 Forecast Combination

6 Key Takeaways and Roadmap

The Evaluation Problem

Five model families, one retail series — how do we know which one to use?

Every lab in Lectures 1–5 held out a single final window of H periods.

Three failure modes:

1. **Lucky split:** one origin may favor one model by chance
2. **Test-set leakage:** tuning on the test set inflates performance
3. **Horizon conflation:** $h = 1$ and $h = 12$ accuracy are very different — averaging hides the pattern

A single split is a photograph. Walk-forward evaluation is a movie.

What rigorous evaluation requires:

- Multiple forecast origins (not one T)
- Horizon-specific accuracy ($h = 1, \ldots, H$)
- A loss function matched to business costs
- Statistical testing to separate signal from noise

This lecture delivers all four. (See L01 footnote: "walk-forward evaluation: Lecture 6.")

Before evaluating, a map of what we now have:

Family	**Lecture**	**Strength**	**Weakness**
Benchmarks	L1	Simple, fast	No dynamics
Regression / AR	L2	Interpretable	Assumes linearity
ETS / Exp. smooth.	L3	Trend & seasonality	Univariate only
ARIMA / SARIMA	L4	Box-Jenkins; flexible	Complex identification
VAR / ARIMAX / ECM	L5	Multivariate dynamics	Parameter-heavy

Evaluation is the final step in the modeling cycle: build $\rightarrow$ fit $\rightarrow$ **evaluate out-of-sample** $\rightarrow$ decide. No model is universally best across all series, horizons, and loss functions.

Error Metrics

RMSE, MAE, MAPE, and MASE each answer a different business question.

Let $e_{T+h} = y_{T+h} - \hat{y}_{T+h|T}$.

$\text{RMSE}_h = \sqrt{\frac{1}{n}\sum e_{T+h}^2}$	*scale-dep.*
$\text{MAE}_h = \frac{1}{n}\sum \lvert e_{T+h} \rvert$	*scale-dep.*
$\text{MAPE}_h = \frac{100}{n}\sum \lvert e_{T+h}/y_{T+h} \rvert$	*scale-free*
$\text{MASE}_h = \text{MAE}_h \,/\, \bar{q}$	*scale-free*

$\bar{q}$ = in-sample MAE of seasonal naive.

Decision guide:

RMSE: large errors costly

MAE: equal cost per unit

MASE < 1: beats naïve

MAPE: only when $y_t \gg 0$

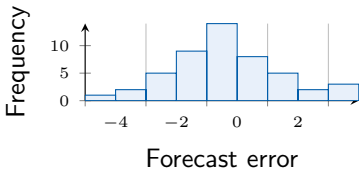
Never report RMSE alone — show MAE alongside.

Socratic: two series, RMSE(A) < RMSE(B) but MASE(A) > MASE(B). Which model is "better"?

RMSE penalizes large errors quadratically; MAE treats all errors equally.

When RMSE $\gg$ MAE:

- Error distribution has heavy tails
- Occasional very large misses
- If those misses are costly $\Rightarrow$ RMSE is the right metric



Stockout cost $= 5\times$ overstock cost. The loss function is *asymmetric* and favors penalizing large underestimates.

$\Rightarrow$ RMSE aligns better than MAE.

$\Rightarrow$ Ideally: use asymmetric loss $L(e) = c_1 e^+ + c_2 e^-$, where $e^+ = \max(e, 0)$, $e^- = \max(-e, 0)$.

Never report only RMSE — always show MAE to reveal whether accuracy is driven by a few outlier periods.

MAPE failure cases:

- $y_{T+h} \approx 0 \Rightarrow$ MAPE $\to \infty$ (e.g., new product launch, zero-demand months)
- **Asymmetric:** when $y > 0$, under-forecasts contribute at most 100% to MAPE (bounded: $\hat{y} = 0 \Rightarrow |e/y| = 1$); over-forecasts are unbounded above ($\hat{y} \to \infty \Rightarrow$ MAPE $\to \infty$)
- Percentage errors favor low-variance series

MASE (Hyndman and Koehler 2006):

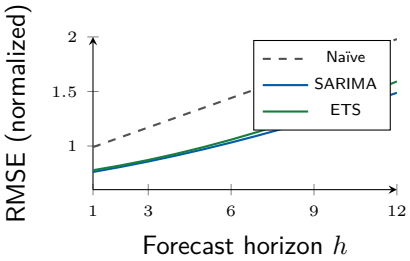
$\text{MASE} < 1 \Rightarrow$ beats seasonal naïve

$\text{MASE} > 1 \Rightarrow$ worse than naïve

Works for intermittent demand, zero values, and cross-series comparison.

Practical recommendation: report RMSE + MAE for absolute accuracy; MASE for benchmarking; MAPE only when $y_t \gg 0$. *Socratic: a new product launches in month 1 with $y_1 = 0$. Which metric can you not compute, and what do you use instead?*

Key insight: the winner at $h = 1$ is not always the winner at $h = 12$.



Always report accuracy **by horizon**, not just as a single average.

Why profiles diverge:

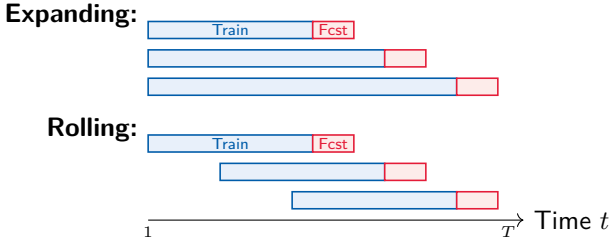
- At $h = 1$: model captures short-run dynamics
- At $h = 12$: model must extrapolate seasonal patterns
- Naïve accumulates error with horizon; ARIMA/ETS maintain seasonal anchor

Lab 6 plots this profile on RSXFS with 36 walk-forward origins.

Walk-Forward Validation

A single test set gives one data point. Walk-forward gives a distribution.

At each **forecast origin** t: fit model on $\{y_1, \ldots, y_t\}$, forecast $\{y_{t+1}, \ldots, y_{t+H}\}$.



Expanding: *train grows at each origin.* ***Rolling:*** *fixed-size window shifts forward. Both produce a time series of forecast errors per horizon* h.

Use expanding window when:

- Data-generating process is *stable* over time
- More training data $\rightarrow$ better parameters
- Long series with no structural breaks

Use rolling window when:

- Structural breaks (e.g., COVID-19 shock)
- Time-varying parameters / regime change
- Fixed window forces model to "forget" old patterns

Minimum training size: choose T_0 large enough for reliable estimation. For SARIMA$(p,d,q)(P,D,Q)_m$: at least $3m + p + P$ observations.

For monthly data: $T_0 \geq 36$ months.

Python: `TimeSeriesSplit` in `sklearn.model_selection` implements expanding-window CV; rolling requires a manual loop.

1. Choose: first origin T_0, window type, horizon H
2. **For** each origin $t = T_0, T_0 + 1, \ldots, T - H$:
   (a) Fit model on $\{y_1, \ldots, y_t\}$
   (b) Generate forecasts $\hat{y}_{t+h|t}$, $h = 1, \ldots, H$
   (c) Store errors $e_{t,h} = y_{t+h} - \hat{y}_{t+h|t}$
3. Average over origins per horizon:

$$\mathrm{RMSE}_h = \sqrt{\frac{1}{n_h} \sum_t e_{t,h}^2}$$

4. Plot horizon profile; run DM test for significance

Walk-forward errors are the closest thing to a **live forecast track record**: each $e_{t,h}$ represents the error you would have made if you had used this model in real time.

$e_{t,h}$ generalizes e_{T+h} from the Error Metrics section to multiple origins t; the metric formulas are the same, averaged over t.

Number of origins $n_h = T - T_0 - H + 1$ for $h = H$; with monthly data, 24–36 origins is typical.

Illustrative results (36 expanding-window origins, $H = 12$):

Model	RMSE$_{h=1}$	RMSE$_{h=3}$	RMSE$_{h=12}$	MASE
Naïve (seasonal)	2,810	3,145	4,210	1.00
SARIMA(1,1,1)(0,1,1)	1,720	2,040	3,480	0.76
ETS (auto-AIC)	1,690	2,020	3,510	0.75
ARIMAX (+ sentiment)	1,640	1,980	3,530	0.73
VAR (BIC order)	1,660	2,100	3,620	0.77
Equal-weight combo	**1,600**	**1,920**	**3,390**	**0.71**

Combination beats all individual models across every horizon. Is the ETS vs. SARIMA gap statistically real?
→ Section 4.

The Diebold-Mariano Test

Visual inspection says Model A looks better. The DM test asks: is the gap statistically real?

1995

Let $d_t = g(e_{1t}) - g(e_{2t})$ be the **loss differential** (e.g., $d_t = e_{1t}^2 - e_{2t}^2$ for MSE loss). Then:

$$DM = \frac{\bar{d}}{\widehat{\mathrm{se}}(\bar{d})} \xrightarrow{d} \mathcal{N}(0,1), \quad \bar{d} = \tfrac{1}{n}\textstyle\sum_{t=1}^{n} d_t$$

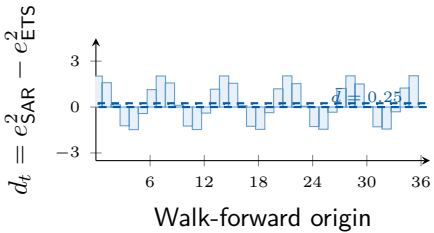
H_0: equal predictive accuracy $\big(\mathbb{E}[d_t] = 0\big)$.

$\widehat{\mathrm{se}}(\bar{d})$ uses a **HAC variance estimate** to account for serial correlation in d_t.

Why HAC? For an h-step forecast, d_t is serially correlated up to lag $h-1$ (overlapping forecast errors). Ignoring this understates the standard error.

Use Newey-West bandwidth $= h-1$ for the HAC variance; with $h = 1$ (one-step forecasts), d_t is white noise under H_0.

SARIMA vs. ETS on RSXFS ($h = 1$, 36 origins, MSE loss):



Decision rule:

- $DM > 1.645$ (one-sided): Model 2 significantly better
- $|DM| > 1.96$ (two-sided): models differ significantly
- Fail to reject: insufficient evidence — do not conclude models are equal

$\bar{d} > 0$: SARIMA has slightly higher MSE. If $DM > 1.645$, use ETS.

When the DM test applies:

- Non-nested models (e.g., SARIMA vs. ETS)
- Stationary loss differentials d_t
- Sufficient number of origins ($n \geq 20$)

When it does not apply:

- **Nested models** (e.g., AR(1) vs. AR(2)): under H_0 the DM statistic is non-standard. Use the Clark-West correction (Clark and West 2007)
- Very small samples ($n < 15$): use finite-sample correction (Harvey et al. 1997)

DM tests **equal expected loss**, not model fit.

Always use *out-of-sample* walk-forward errors — never in-sample residuals.

Always specify the loss function $g(\cdot)$: MSE, MAE, or asymmetric loss.

Socratic: if $\bar{d} > 0$ but $DM = 0.8$, what conclusion do you draw? What would increase power?

Forecast Combination

No single model wins every period. Combining forecasts diversifies model risk.

The portfolio analogy: diversification reduces portfolio variance when assets are imperfectly correlated.

$$\mathrm{Var}\left(\tfrac{1}{2}e_1 + \tfrac{1}{2}e_2\right) = \tfrac{1}{4}\left(\sigma_1^2 + 2\rho\sigma_1\sigma_2 + \sigma_2^2\right)$$

If $\rho < 1$: the combined error variance is smaller than the average of individual variances.
Empirical finding: Bates and Granger (1969) showed that weighted combinations *generally* outperform the best individual model in empirical comparisons.

If two models have equal accuracy but their errors are **not perfectly correlated**, the simple average outperforms either model alone.

SARIMA–ETS error correlation ≈ 0.72 *on RSXFS* $\Rightarrow$ *significant gains from combining.*

Strategy	Formula	Key property	Practical note
Equal weights	$\hat{y}^C = \frac{1}{K}\sum_k \hat{y}^{(k)}$	No estimation needed	Surprisingly robust
RMSE weights	$w_k \propto \mathrm{RMSE}_k^{-1}$	Rewards accuracy	Simple; no overfitting
OLS weights	$\hat{w}$ from regressing y on $\hat{y}^{(k)}$	Optimal in theory	Prone to overfitting

The combination puzzle (Timmermann 2006; Stock and Watson 2004): estimated OLS weights *often underperform* equal weights out-of-sample. Three reasons:

1. Estimation error in $\hat{w}$
2. Model instability / structural breaks
3. High collinearity among $\hat{y}^{(k)}$

OLS weights can be negative or > 1. Constrain to $w_k \geq 0$, $\sum w_k = 1$ in practice. Alternatively: shrink toward equal weights via ridge.

Extending our walk-forward results (RSXFS, 36 origins):

Method	RMSE ($h = 1$)	RMSE ($h = 12$)	MASE (avg)
Best individual (ARIMAX)	1,640	3,530	0.73
RMSE-weighted combo	1,620	3,410	0.72
Equal-weight combo	**1,600**	**3,390**	**0.71**
OLS combo (unconstrained)	1,680	3,580	0.76

Decision rule: when uncertain about which model to trust, **combine** (equal weights as default). When you have strong theoretical or empirical reasons to prefer one model, use it alone — but verify with the DM test.

The M4 competition (Makridakis et al. 2020): top hybrid submissions combined diverse methods. Equal-weight ensembles ranked in the top 25% of all 61 participating methods.

Metric choice is a business decision: align RMSE/MAE/MASE with the cost structure; MAPE fails near zero.

Walk-forward validation gives a distribution of errors across origins — one split is not sufficient.

Diebold-Mariano test determines whether accuracy differences are statistically real (use out-of-sample errors, HAC standard errors).

Nested model comparisons require the Clark-West correction; standard DM is invalid for nested models.

Equal-weight combination typically matches or beats the best individual model — the combination puzzle.

We now have five model families + rigorous evaluation. The next step: machine learning for forecasting.

Classical forecasting is now complete: benchmarks → regression → ETS → ARIMA → multivariate → **evaluation**.

Lecture 7: Machine Learning Introduction — bias-variance tradeoff, train/validation/test discipline, and how cross-validation extends walk-forward ideas to ML models.

Aspect	**Classical forecasting**	**ML forecasting**
Model structure	Specified (ARIMA, VAR)	Learned from data
Assumptions	Linearity, normality	Minimal
Overfitting risk	Lower (few parameters)	High (regularize!)
Evaluation	Walk-forward; DM test	Same + CV grid search

Lab 6: walk-forward on RSXFS, DM test (SARIMA vs. ETS), forecast combination.

- Bates, John M. and C. W. J. Granger (1969). "The Combination of Forecasts". In: *Operational Research Quarterly* 20.4, pp. 451–468.
- Clark, Todd E. and Kenneth D. West (2007). "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models". In: *Journal of Econometrics* 138.1, pp. 291–311.
- Diebold, Francis X. and Roberto S. Mariano (1995). "Comparing Predictive Accuracy". In: *Journal of Business & Economic Statistics* 13.3, pp. 253–263.
- Harvey, David, Stephen Leybourne, and Paul Newbold (1997). "Testing the Equality of Prediction Mean Squared Errors". In: *International Journal of Forecasting* 13.2, pp. 281–291.
- Hyndman, Rob J. and Anne B. Koehler (2006). "Another Look at Measures of Forecast Accuracy". In: *International Journal of Forecasting* 22.4, pp. 679–688.
- Makridakis, Spyros, Evangelos Spiliotis, and Vassilios Assimakopoulos (2020). "The M4 Competition: 100,000 Time Series and 61 Forecasting Methods". In: *International Journal of Forecasting* 36.1, pp. 54–74.

Stock, James H. and Mark W. Watson (2004). "Combination Forecasts of Output Growth in a Seven-Country Data Set". In: *Journal of Forecasting* 23.6, pp. 405–430.

Timmermann, Allan (2006). "Forecast Combinations". In: *Handbook of Economic Forecasting* 1, pp. 135–196.