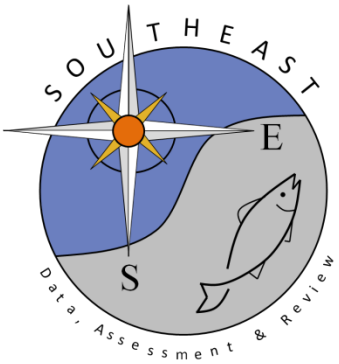
Rolling scaled forecast accuracy

Rob J Hyndman

SEDAR90-RD-39



Rolling scaled forecast accuracy

DATE
20 January 2026

TOPICS
FORECASTING

When we compute a MASE or RMSSE using a rolling origin, should the scaling factor be recalculated every time?

I've been asked this a couple of times, so perhaps it is worth a blog post.

For a simple training/test split, the Mean Absolute Scaled Error (MASE) (Hyndman & Koehler, 2006) is defined as

$$\text{MASE} = \frac{\frac{1}{H}\sum_{t=T+1}^{T+H} |y_t - \hat{y}_{t|T}|}{\frac{1}{T-m}\sum_{t=m+1}^{T} |y_t - y_{t-m}|}$$

and the Root Mean Squared Scaled Error (RMSSE) is defined as

$$\text{RMSSE} = \sqrt{\frac{\frac{1}{H}\sum_{t=T+1}^{T+H} (y_t - \hat{y}_{t|T})^2}{\frac{1}{T-m}\sum_{t=m+1}^{T} (y_t - y_{t-m})^2}}.$$

In both cases, $m = 1$ for non-seasonal data, where m is the seasonal period for seasonal data, and the sum in the numerator is over the *test* set, while the sum in the denominator is over the *training* set. The notation $\hat{y}_{t|T}$ means the forecast of y_t given data $y_1, \dots, y_T$.

These measures are discussed in my forecasting textbook with George Athanasopoulos. The denominator is a scaling factor, introduced so that you can compare MASE or RMSSE values across series of different units. For example, are the forecasts of widget sales more accurate than the forecasts of electricity demand? If all your forecasts are in the same units, then you don't need to remove the scale, and it is simpler to just use MAE or RMSE (i.e., only the numerators of the above equations).

Now, the question is, when we are doing time series cross-validation, and computing forecast accuracy over a series of training/test sets with a rolling origin, does it make sense to compute a different scaling factor each time? For example, a cross-validated MASE for h-step forecasts is

$$\text{MASE}_h = \frac{1}{T-I-h+1}\sum_{t=I}^{T-h} \frac{|y_{t+h} - \hat{y}_{t+h|t}|}{\frac{1}{t-m}\sum_{s=m+1}^{t} |y_s - y_{s-m}|},$$

where the first I observations form the smallest training set, and subsequent training sets increase one observation at a time.

An alternative approach would be to compute the scaling factor across all available data, rather than calculate it separately for each training set. Then MASE would become

$$\text{MASE}_h = \frac{\frac{1}{T-I-h+1}\sum_{t=I}^{T-h}|y_{t+h} - \hat{y}_{t+h|t}|}{\frac{1}{T-m}\sum_{t=m+1}^{T}|y_t - y_{t-m}|}$$

Let's think about the advantages of this alternative:

1. It is (slightly) faster. But the denominators are very fast to compute, so this really doesn't make much difference.
2. It removes a source of variation from the calculation. The scaled errors will be more variable when the denominator changes with each test set, and that makes it harder to see the difference between forecasting methods.
3. It uses more data in computing the scaling factor, which reduces the uncertainty in the estimate of the accuracy measure. This is potentially important if $I \ll T$, so that some training sets are relatively small compared to the available data.

As for disadvantages:

1. The measures are no longer true measures of forecast accuracy because the calculation potentially involves future observations. On the face of it, this seems important, but in reality it isn't. The future observations aren't affecting the forecasts, only the scaling factor, so there is no leakage involved.
2. One of my correspondents suggested that it changed the interpretation. I think the interpretability of these scaled measures is over-rated. They can only be interpreted as a ratio of out-of-sample accuracy to in-sample accuracy, and the two are not necessarily even over the same horizons. How the scaling factor is calculated doesn't really affect the interpretation, because there is not much value in interpretation either way.

So I suggest that computing the scaling factor across all available training data when calculating cross-validated MASE and RMSSE values is a good idea.

That raises the question as to why shouldn't we use all available data in the denominator when doing a simple training/test split? In that context, the only advantage that is relevant is #3 above, and unless the test set is particularly large, it shouldn't make much difference. Nevertheless, I can't see any real problem in using all available data in computing the scaling factor.

References

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International J Forecasting*, *22*(4), 679–688. https://robjhyndman.com/publications/another-look-at-measures-of-forecast-accuracy/