ANOTHER LOOK AT FORECAST-ACCURACY METRICS FOR INTERMITTENT DEMAND

by Rob J. Hyndman

Preview: Some traditional measurements of forecast accuracy are unsuitable for intermittent-demand data because they can give infinite or undefined values. Rob Hyndman summarizes these forecast accuracy metrics and explains their potential failings. He also introduces a new metric—the mean absolute scaled error (MASE)—which is more appropriate for intermittent-demand data. More generally, he believes that the MASE should become the standard metric for comparing forecast accuracy across multiple time series.

Rob Hyndman is Professor of Statistics at Monash University, Australia, and Editor in Chief of the *International Journal of Forecasting*. He is an experienced consultant who has worked with over 200 clients during the last 20 years, on projects covering all areas of applied statistics, from forecasting to the ecology of lemmings. He is coauthor of the well-known textbook, *Forecasting: Methods and Applications* (Wiley, 1998), and he has published more than 40 journal articles. Rob is Director of the Business and Economic Forecasting Unit, Monash University, one of the leading forecasting research groups in the world.

- There are four types of forecast-error metrics: scale-dependent metrics such as the mean absolute error (MAE or MAD); percentage-error metrics such as the mean absolute percent error (MAPE); relative-error metrics, which average the ratios of the errors from a designated method to the errors of a naïve method; and scale-free error metrics, which express each error as a ratio to an average error from a baseline method.
- For assessing accuracy on a single series, I prefer the MAE because it is easiest to understand and compute. However, it cannot be compared across series because it is scale dependent; it makes no sense to compare accuracy on different scales.
- Percentage errors have the advantage of being scale independent, so they are frequently used to compare forecast performance between different data series. But measurements based on percentage errors have the disadvantage of being infinite or undefined if there are zero values in a series, as is frequent for intermittent data.
- Relative-error metrics are also scale independent. However, when the errors are small, as they can be with intermittent series, use of the *naïve method* as a benchmark is no longer possible because it would involve division by zero.
- The scale-free error metric I call the mean absolute scaled error (MASE) can be used to compare forecast methods on a single series and also to compare forecast accuracy between series. This metric is well suited to intermittent-demand series because it never gives infinite or undefined values.

Introduction: Three Ways to Generate Forecasts

There are three ways we may generate forecasts (F) of a quantity (Y) from a particular forecasting method:

1. We can compute forecasts from a common origin t (for example, the most recent month) for a sequence of forecast horizons $F_{n+1},...,F_{n+m}$ based on data from times $t = 1,...,n$. This is the standard procedure implemented by forecasters in real time.

2. We can vary the origin from which forecasts are made but maintain a consistent forecast horizon. For example, we can generate a series of one-period-ahead forecasts $F_{1+h},...,F_{m+h}$ where each F_{j+h} is based on data from times $t = 1,..., j$. This procedure is done not only to give attention to the forecast errors at a particular horizon but also to show how the forecast error changes as the horizon lengthens.

3. We may generate forecasts for a single future period using multiple data series, such as a collection of products or items. This procedure can be useful to demand planners as they assess aggregate accuracy over items or products at a location. This is also the procedure that underlies forecasting competitions, which compare the accuracy of different methods across multiple series.

While these are very different situations, measuring forecast accuracy is similar in each case. It is useful to have a forecast accuracy metric that can be used for all three cases.

An Example of What Can Go Wrong

Consider the classic intermittent-demand series shown in Figure 1. These data were part of a consulting project I did for a major Australian lubricant manufacturer.

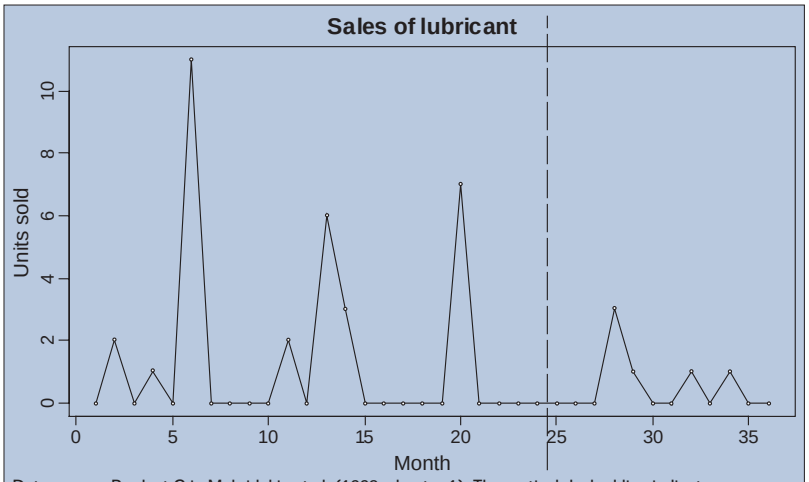
Suppose we are interested in comparing the forecast accuracy of four simple methods: (1) the historical mean, using data up to the most recent observation; (2) the *naïve* or random-walk method, in which the forecast for each future period is the actual value for this period; (3) simple exponential smoothing; and (4) Croston's method for intermittent demands (Boylan, 2005). For methods (3) and (4) I have used a smoothing parameter of 0.1.

I compared the *in-sample* performance of these methods by varying the origin and generating a sequence of one-period-ahead forecasts – the second forecasting procedure described in the introduction. I also calculated the *out-of-sample* performance based on forecasting the data in the hold-out period, using information from the fitting period alone. These out-of-sample forecasts are from one to twelve steps ahead and are not updated in the hold-out period.

Table 1 shows some commonly used forecast-accuracy metrics applied to these data. The metrics are all defined in the next section. There are many infinite values occurring in Table 1. These are caused by division by zero. The undefined values for the naïve method arise from the division of zero by zero. The only measurement that always gives sensible results for all four of the forecasting methods is the MASE, or the mean absolute scaled error. Infinite, undefined, or zero values plague the other accuracy measurements.

In this particular series, the out-of-sample period has smaller errors (is more predictable) than the in-sample period because the in-sample period includes some relatively large observations. In general, we would expect out-of-sample errors to be larger.

Figure 1. Three Years of Monthly Sales of a Lubricant Product Sold in Large Containers



Data source: Product C in Makridakis et al. (1998, chapter 1). The vertical dashed line indicates the end of the data used for fitting and the start of the holdout set used for out-of-sample forecasting.

Table 1. Forecast-Accuracy Metrics for Lubricant Sales

		Mean		Naïve		SES		Croston	
		In	Out	In	Out	In	Out	In	Out
GMAE	Geometric Mean Absolute Error	1.65	0.96	0.00	0.00	1.33	0.09	0.00	0.99
MAPE	Mean Absolute Percentage Error	∞	∞	–	–	∞	∞	∞	∞
sMAPE	Symmetric Mean Absolute Percentage Error	1.73	1.47	–	–	1.82	1.42	1.70	1.47
MdRAE	Median Relative Absolute Error	0.95	∞			0.98	∞	0.93	∞
GMRAE	Geometric Mean Relative Absolute Error	∞	∞	–	–	∞	∞	∞	∞
MASE	Mean Absolute Scaled Error	0.86	0.44	1.00	0.20	0.78	0.33	0.79	0.45

Measurement of Forecast Errors

We can measure and average forecast errors in several ways:

Scale-dependent errors

The forecast error is simply, $e_t = Y_t - F_t$, regardless of how the forecast was produced. This is on the same scale as the data, applying to anything from ships to screws. Accuracy measurements based on e_t are therefore scale-dependent.

The most commonly used scale-dependent metrics are based on absolute errors or on squared errors:

$$\text{Mean Absolute Error (MAE)} = \text{mean}(|e_t|)$$
$$\text{Geometric Mean Absolute Error (GMAE)} = \text{gmean}(|e_t|)$$
$$\text{Mean Square Error (MSE)} = \text{mean}(e_t^2)$$

where "gmean" is a geometric mean.

The MAE is often abbreviated as the MAD ("D" for "deviation"). The use of absolute values or squared values prevents negative and positive errors from offsetting each other.

Since all of these metrics are on the same scale as the data, none of them are meaningful for assessing a method's accuracy across multiple series.

For assessing accuracy on a single series, I prefer the MAE because it is easiest to understand and compute. However, it cannot be compared between series because it is scale dependent.

For intermittent-demand data, Syntetos and Boylan (2005) recommend the use of GMAE, although they call it the GRMSE. (The GMAE and GRMSE are identical because the square root and the square cancel each other in a geometric mean.) Boylan and Syntetos (this issue) point out that the GMAE has the flaw of being equal to zero when any error is zero, a problem which will occur when both the actual and forecasted demands are zero. This is the result seen in Table 1 for the naïve method.

Boylan and Syntetos claim that such a situation would occur only if an inappropriate forecasting method is used. However, it is not clear that the naïve method is always inappropriate. Further, Hoover indicates that division-by-zero errors in intermittent series are expected occurrences for repair parts. I suggest that the GMAE is problematic for assessing accuracy on intermittent-demand data.

Percentage errors

The percentage error is given by $p_t = 100e_t/Y_t$. Percentage errors have the advantage of being scale independent, so they are frequently used to compare forecast performance between different data series. The most commonly used metric is

$$\text{Mean Absolute Percentage Error (MAPE)} = \text{mean}(|p_t|)$$

Measurements based on percentage errors have the disadvantage of being infinite or undefined if there are zero values in a series, as is frequent for intermittent data. Moreover, percentage errors can have an extremely skewed distribution when actual values are close to zero. With intermittent-demand data, it is impossible to use the MAPE because of the occurrences of zero periods of demand.

The MAPE has another disadvantage: it puts a heavier penalty on positive errors than on negative errors. This observation has led to the use of the "symmetric" MAPE (sMAPE) in the M3-competition (Makridakis & Hibon, 2000). It is defined by

$$\text{sMAPE} = \text{mean}(200\,|Y_t - F_t| / (Y_t + F_t))$$

However, if the actual value Y_t is zero, the forecast F_t is likely to be close to zero. Thus the measurement will still involve division by a number close to zero. Also, the value of sMAPE can be negative, giving it an ambiguous interpretation.

Relative errors

An alternative to percentages for the calculation of scale-independent measurements involves dividing each error by the error obtained using some benchmark method of forecasting. Let $r_t = e_t/e_t^*$ denote the relative error where e_t^* is the forecast error obtained from the benchmark method. Usually the benchmark method is the naïve method where F_t is equal to the last observation. Then we can define

$$\text{Median Relative Absolute Error (MdRAE)} = \text{median}(|r_t|)$$
$$\text{Geometric Mean Relative Absolute Error (GMRAE)} = \text{gmean}(|r_t|)$$

Because they are not scale dependent, these relative-error metrics were recommended in studies by Armstrong and Collopy (1992) and by Fildes (1992) for assessing forecast accuracy across multiple series. However, when the errors are small, as they can be with intermittent series, use of the naïve method as a benchmark is no longer possible because it would involve division by zero.

Scale-free errors

The MASE was proposed by Hyndman and Koehler (2006) as a generally applicable measurement of forecast accuracy without the problems seen in the other measurements. They proposed scaling the errors based on the *in-sample* MAE from the naïve forecast method. Using the naïve method, we generate one-period-ahead forecasts from each data point in the sample. Accordingly, a scaled error is defined as

$$q_t = \frac{e_t}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|}$$

The result is independent of the scale of the data. A scaled error is less than one if it arises from a better forecast than the average one-step, naïve forecast computed in-sample. Conversely, it is greater than one if the forecast is worse than the average one-step, naïve forecast computed in-sample.

The mean absolute scaled error is simply

$$\text{MASE} = \text{mean}(|q_t|)$$

The first row of Table 2 shows the intermittent series plotted in Figure 1. The second row gives the naïve forecasts, which are equal to the previous actual values. The final row shows the naïve-forecast errors. The denominator of q_t is the mean of the shaded values in this row; that is the MAE of the naïve method.

Table 2. Monthly Lubricant Sales, Naïve Forecast

	In-sample	Out-of-sample
Actual Y_t	0 2 0 1 0 1 0 0 0 0 2 0 6 3 0 0 0 0 0 7 0 0 0 0	0 0 0 3 1 0 0 1 0 1 0 0
Naïve forecast $\hat{Y}_t$	0 2 0 1 0 1 0 0 0 0 2 0 6 3 0 0 0 0 0 7 0 0 0	0 0 0 0 0 0 0 0 0 0 0 0
Error $\lvert Y_t - \hat{Y}_t \rvert$	2 2 1 1 1 1 0 0 0 2 2 6 3 3 0 0 0 0 7 7 0 0 0	0 0 0 3 1 0 0 1 0 1 0 0

The only circumstance under which the MASE would be infinite or undefined is when all historical observations are equal.

The in-sample MAE is used in the denominator because it is always available and it effectively scales the errors. In contrast, the out-of-sample MAE for the naïve method may be zero because it is usually based on fewer observations. For example, if we were forecasting only two steps ahead, then the out-of-sample MAE would be zero. If we wanted to compare forecast accuracy at one step ahead for ten different series, then we would have one error for each series. The out-of-sample MAE in this case is also zero. These types of problems are avoided by using in-sample, one-step MAE.

A closely related idea is the MAD/Mean ratio proposed by Hoover (this issue) which scales the errors by the in-sample mean of the series instead of the in-sample mean absolute error. This ratio also renders the errors scale free and is always finite unless all historical data happen to be zero. Hoover explains the use of the MAD/Mean ratio only in the case of in-sample, one-step forecasts (situation 2 of the three situations described in the introduction). However, it would also be straightforward to use the MAD/Mean ratio in the other two forecasting situations.

The main advantage of the MASE over the MAD/Mean ratio is that the MASE is more widely applicable. The MAD/Mean ratio assumes that the mean is stable over time (technically, that the series is "stationary"). This is not true for data which show trend, seasonality, or other patterns. While intermittent data is often quite stable, sometimes seasonality does occur, and this might make the MAD/Mean ratio unreliable. In contrast, the MASE is suitable even when the data exhibit a trend or a seasonal pattern.

The MASE can be used to compare forecast methods on a single series, and, because it is scale-free, to compare forecast accuracy across series. For example, you can average the MASE values of several series to obtain a measurement of forecast accuracy for the group of series. This measurement can then be compared with the MASE values of other groups of series to identify which series are the most difficult to forecast. Typical values for one-step MASE values are less than one, as it is usually possible to obtain forecasts more accurate than the naïve method. Multistep MASE values are often larger than one, as it becomes more difficult to forecast as the horizon increases.

The MASE is the only available accuracy measurement that can be used in all three forecasting situations described in the introduction, and for all forecast methods and all types of series. I suggest that it is the best accuracy metric for intermittent demand studies and beyond.

References

Armstrong, J. S. & Collopy F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons, *International Journal of Forecasting*, 8, 69–80.

Boylan, J. (2005). Intermittent and lumpy demand: A forecasting challenge, *Foresight: The International Journal of Applied Forecasting*, Issue 1, 36-42.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods, *International Journal of Forecasting*, 8, 81–98.

Hyndman, R. J. & Koehler, A. B. (2006). Another look at measures of forecast accuracy, *International Journal of Forecasting*. To appear.

Makridakis, S. & Hibon, M. (2000). The M3-competition: Results, conclusions and implications, *International Journal of Forecasting*, 16, 451–476.

Makridakis, S. G., Wheelwright, S. C. & Hyndman, R. J. (1998). *Forecasting: Methods and Applications* (3rd ed.), New York: John Wiley & Sons.

Syntetos, A. A. & Boylan, J. E, (2005). The accuracy of intermittent demand estimates, *International Journal of Forecasting*, 21, 303-314.

Contact Info:
Rob J. Hyndman
Monash University, Australia
Rob.Hyndman@buseco.monash.edu