Another look at measures of forecast accuracy

Rob J Hyndman

Department of Econometrics and Business Statistics,

Monash University, VIC 3800, Australia.

Telephone: +61–3–9905–2358

Email: Rob.Hyndman@buseco.monash.edu

Anne B Koehler

Department of Decision Sciences & Management Information Systems

Miami University

Oxford, Ohio 45056, USA.

Telephone: +1–513–529–4826

E-Mail: koehleab@muohio.edu

2 November 2005

Another look at measures of forecast accuracy

Abstract: We discuss and compare measures of accuracy of univariate time series forecasts. The methods used in the M-competition and the M3-competition, and many of the measures recommended by previous authors on this topic, are found to be degenerate in commonly occurring situations. Instead, we propose that the mean absolute scaled error become the standard measure for comparing forecast accuracy across multiple time series.

Keywords: forecast accuracy, forecast evaluation, forecast error measures, M-competition, mean absolute scaled error.

1 Introduction

Many measures of forecast accuracy have been proposed in the past, and several authors have made recommendations about what should be used when comparing the accuracy of forecast methods applied to univariate time series data. It is our contention that many of these proposed measures of forecast accuracy are not generally applicable, can be infinite or undefined, and can produce misleading results. We provide our own recommendations of what should be used in empirical comparisons. In particular, we do not recommend the use of any of the measures that were used in the M-competition and the M3-competition.

To demonstrate the inadequacy of many measures of forecast accuracy, we provide three examples of real data in Figure 1. These show series N0472 from the M3-competition[1], monthly log stock returns for the Walt Disney Corporation, and monthly sales of a lubricant product sold in large containers. Note that the Disney return series and the lubricant sales series both include exact zero observations, and the Disney series contains negative values. 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" method or random-walk method based on the most recent observation; (3) simple exponential smoothing and (4) Holt's method. We do not suggest these are the best methods for these data, but they are all simple methods that are widely applied. We compare the in-sample performance of the methods (based on one-step-ahead forecasts) and the out-of-sample performance (based on forecasting the data in the hold-out period using only information from the fitting period).

Tables 1–3 show some forecast error measures for these methods applied to the example data. The acronyms are defined below and we explicitly define the measures in Sections 2 and 3. The relative measures are all computed relative to a naïve (random walk) method.

MAPE	Mean Absolute Percentage Error
MdAPE	Median Absolute Percentage Error
sMAPE	Symmetric Mean Absolute Percentage Error
sMdAPE	Symmetric Median Absolute Percentage Error
MdRAE	Median Relative Absolute Error
GMRAE	Geometric Mean Relative Absolute Error
MASE	Mean Absolute Scaled Error

[1]Data downloaded from http://www.forecasters.org/data/m3comp/m3comp.htm

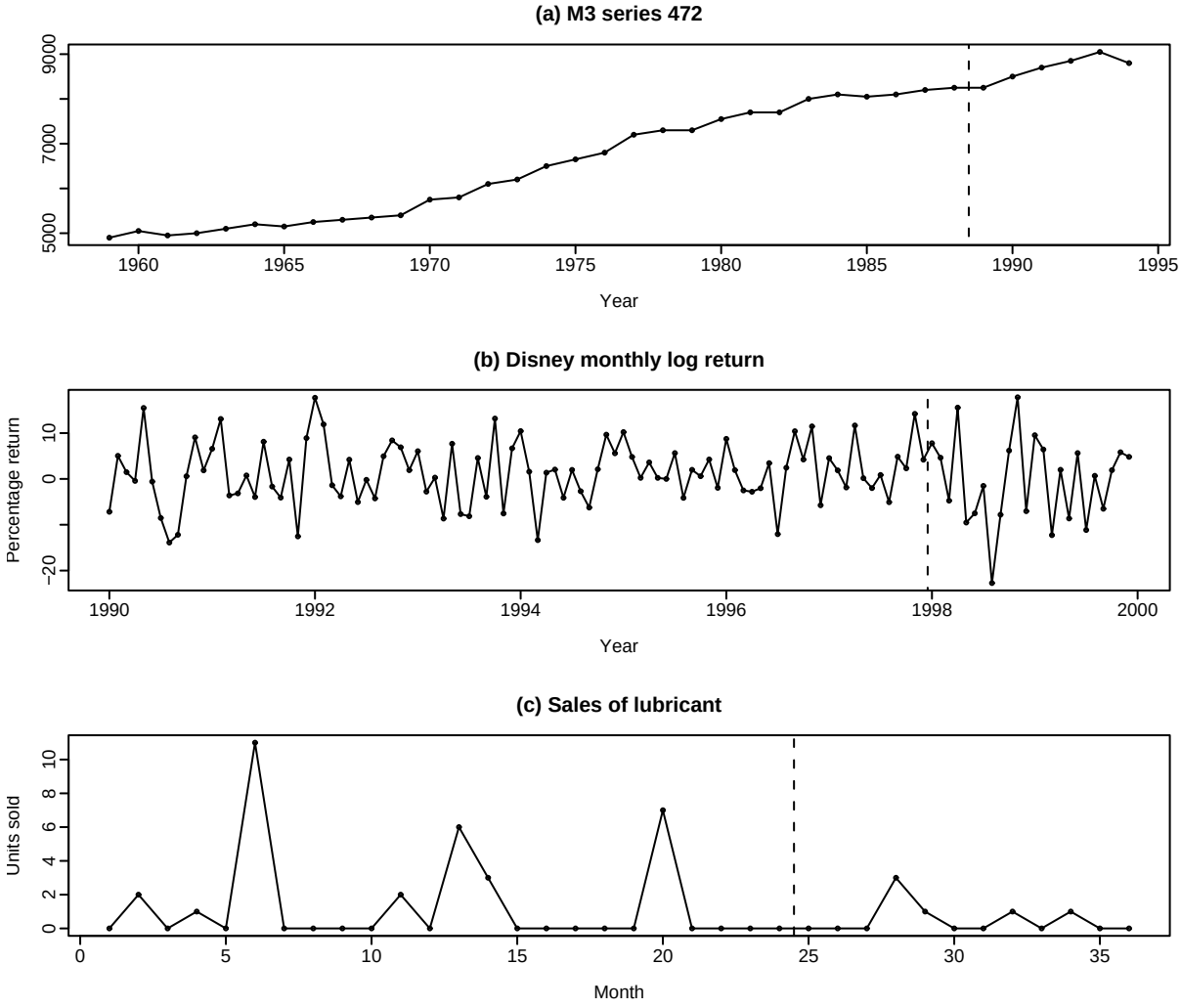

Figure 1: *Example (a): Series 472 from the M3-competition. Example (b): ten years of monthly log stock returns for the Walt Disney Corporation, 1990–1999. Data source: Tsay (2002), chapter 1. Example (c): three years of monthly sales of a lubricant product sold in large containers. Data source: 'Product C' in Makridakis, Wheelwright and Hyndman (1998), chapter 1. The vertical dashed lines indicate the end of the data used for fitting and the start of the "hold-out" set used for out-of-sample forecast.*

Example A	Mean		Random walk		SES		Holt	
	In	Out	In	Out	In	Out	In	Out
MAPE	14.09	25.57	2.01	5.00	2.08	5.03	1.31	3.94
MdAPE	17.44	26.13	1.61	5.71	1.86	5.74	1.04	4.20
sMAPE	0.16	0.29	0.02	0.05	0.02	0.05	0.01	0.04
sMdAPE	0.19	0.30	0.02	0.06	0.02	0.06	0.01	0.04
MdRAE	6.50	4.61	Undefined	Undefined	1.02	1.01	0.50	0.80
GMRAE	∞	∞	Undefined	Undefined	∞	∞	∞	∞
MASE	7.88	17.23	1.00	3.42	1.04	3.44	0.66	2.69

Table 1: *Forecast error measures for M3 series N0472.*

Example B	Mean		Random walk		SES		Holt	
	In	Out	In	Out	In	Out	In	Out
MAPE	∞	96.56	∞	125.90	∞	102.32	∞	113.90
MdAPE	101.61	96.92	159.25	119.19	97.06	98.63	98.70	97.91
sMAPE	-0.80	-2.04	∞	-3.03	-1.10	-0.57	-0.80	-1.12
sMdAPE	0.91	0.35	0.66	0.22	1.01	0.43	0.91	0.48
MdRAE	0.71	0.82	1.00	1.00	0.63	0.88	0.64	0.94
GMRAE	0.66	1.06	1.00	1.00	0.62	1.01	0.60	1.10
MASE	0.72	1.01	1.00	1.06	0.70	1.02	0.70	1.04

Table 2: *Forecast error measures for Disney stocks.*

Example C	Mean		Random walk		SES		Holt	
	In	Out	In	Out	In	Out	In	Out
MAPE	∞	∞	Undefined	Undefined	∞	∞	∞	∞
MdAPE	∞	∞	Undefined	Undefined	∞	∞	∞	∞
sMAPE	1.68	1.39	Undefined	Undefined	1.62	1.37	1.65	1.62
sMdAPE	2.00	2.00	Undefined	Undefined	2.00	2.00	2.00	2.00
MdRAE	1.00	∞	Undefined	Undefined	1.17	∞	0.91	∞
GMRAE	∞	∞	Undefined	Undefined	∞	∞	∞	∞
MASE	0.89	0.39	1.00	0.15	0.83	0.35	0.79	0.20

Table 3: *Forecast error measures for lubricant sales.*

In these tables, we have included measures that have been previously recommended for use in comparing forecast accuracy across many series. Most textbooks recommend the use of the MAPE (e.g., Hanke and Reitsch, 1995, p.120, and Bowerman, O'Connell & Koehler, 2004, p.18) and it was the primary measure in the M-competition (Makridakis, et al., 1982). In contrast, Makridakis, Wheelwright & Hyndman (1998, p45) warn against the use of the MAPE in some circumstances, including those encountered in these examples. Armstrong and Collopy (1992) recommended the use of GMRAE, MdRAE and MdAPE. Fildes (1992) also recommended the use of MdAPE and GMRAE (although he described the latter as the relative geometric root mean square error or GRMSE). The MdRAE, sMAPE and sMdAPE were used in the M3-competition (Makridakis & Hibon, 2000).

The M-competition and M3-competition also used rankings amongst competing methods. We do not include those here as they are dependent on the number of methods being considered. They also give no indication of the size of the forecast errors. Similarly, both competitions included measures based on the percentage of times one method was better than a benchmark method. Again, such measures are not included here as they do not indicate the size of the errors.

To our knowledge, the MASE has not been proposed before. We consider it the best available measure of forecast accuracy and we argue for it in Section 3.

Note that there are many infinite values occurring in Tables 1–3 due to division by zero. Division by numbers close to zero also results in very large numbers. The undefined values arise due to the division of zero by zero. Some of these are due to computations of the form $Y_t/(Y_t - Y_{t-1})$ where $Y_{t-1} = Y_t = 0$ and others are due to computations of the form $(Y_t - Y_{t-1})/(Y_t - Y_{t-1})$ where $Y_t = Y_{t-1}$. In the latter case, it is algebraically possible to cancel the numerator and denominator, although numerical results will be undefined. Also note that the sMAPE can take negative values although it is meant to be an "absolute percentage error".

Note that with random walk forecasts, the in-sample results for MASE and all results for MdRAE and GMRAE are 1 by definition, as they involve comparison with naïve forecasts. However, some of the values for MdRAE and GMRAE are undefined as explained above.

Of the measures in Tables 1–3, only the MASE can be used for these series due to the occurrence of infinite and undefined values. These three series are not degenerate or unusual—intermittent demand data often contain zeros and many time series of interest to forecasters

take negative observations. The cause of the problems with M3 series N0472 is the occurrence of consecutive observations taking the same value, something that occurs very often with real data.

2 A critical survey of accuracy measures

Let Y_t denote the observation at time t and F_t denote the forecast of Y_t. Then define the forecast error $e_t = Y_t - F_t$. The forecasts may be computed from a common base time, and be of varying forecast horizons. Thus, we may compute out-of-sample forecasts $F_{n+1}, \dots, F_{n+m}$ based on data from times $t = 1, \dots, n$. Alternatively, the forecasts may be from varying base times, and be of a consistent forecast horizon. That is, we may compute forecasts $F_{1+h}, \dots, F_{m+h}$ where each F_{j+h} is based on data from times $t = 1, \dots, j$. The in-sample forecasts in the examples above were based on the second scenario with $h = 1$. A third scenario arises when we wish to compare the accuracy of methods across many series at a single forecast horizon. Then we compute a single F_{n+h} based on data from times $t = 1, \dots, n$ for each of m different series.

We do not distinguish these scenarios in this paper. Rather, we simply look at ways of summarizing forecast accuracy assuming that we have m forecasts and that we observe the data at each forecast period.

We use the notation $\text{mean}(x_t)$ to denote the sample mean of $\{x_t\}$ over the period of interest (or over the series of interest). Analogously, we use $\text{median}(x_t)$ for the sample median and $\text{gmean}(x_t)$ for the geometric mean.

2.1 Scale-dependent measures

There are some commonly used accuracy measures whose scale depends on the scale of the data. These are useful when comparing different methods on the same set of data, but should not be used, for example, when comparing across data sets that have different scales. Nevertheless, the MSE was used by Makridakis et al., 1985, in the M-competition. This inappropriate use of the MSE was widely criticized (e.g., Chatfield, 1988; Armstrong and Collopy, 1992).

The most commonly used scale-dependent measures are based on the absolute error or squared

errors:

$$\begin{aligned}
\text{Mean Square Error (MSE)} &= \text{mean}(e_t^2) \\
\text{Root Mean Square Error (RMSE)} &= \sqrt{\text{MSE}} \\
\text{Mean Absolute Error (MAE)} &= \text{mean}(|e_t|) \\
\text{Median Absolute Error (MdAE)} &= \text{median}(|e_t|)
\end{aligned}$$

Often, the RMSE is preferred to the MSE as it is on the same scale as the data. Historically, the RMSE and MSE have been popular, largely because of their theoretical relevance in statistical modelling. However, they are more sensitive to outliers than MAE or MdAE which has led some authors (e.g., Armstrong, 2001) to recommend against their use in forecast accuracy evaluation.

2.2 Measures based on percentage errors

The percentage error is given by $p_t = 100e_t/Y_t$. Percentage errors have the advantage of being scale-independent, and so are frequently used to compare forecast performance across different data sets. The most commonly used measures are:

$$\begin{aligned}
\text{Mean Absolute Percentage Error (MAPE)} &= \text{mean}(|p_t|) \\
\text{Median Absolute Percentage Error (MdAPE)} &= \text{median}(|p_t|) \\
\text{Root Mean Square Percentage Error (RMSPE)} &= \sqrt{\text{mean}(p_t^2)} \\
\text{Root Median Square Percentage Error (RMdSPE)} &= \sqrt{\text{median}(p_t^2)}\,.
\end{aligned}$$

These measures have the disadvantage of being infinite or undefined if $Y_t = 0$ for any t in the period of interest, and having an extremely skewed distribution when any Y_t is close to zero. This means, for example, that the MAPE is often substantially larger than the MdAPE. Where the data involves small counts (which is common with intermittent demand data; see Gardner, 1990) it is impossible to use these measures as occurrences of zero values of Y_t occur frequently. Excessively large (or infinite) MAPEs were avoided in the M3-competition by only including data that were positive (Makridakis and Hibon, 2000, p.462). However, this is an artificial solution that is impossible to apply in practical situations.

A further disadvantage of methods based on percentage errors is that they assume a meaning-

ful zero. For example, they make no sense in measuring forecast error for temperatures on the Fahrenheit or Celsius scales.

The MAPE and MdAPE also have the disadvantage that they put a heavier penalty on positive errors than on negative errors. This observation led to the use of the so-called "symmetric" measures (Makridakis, 1993) defined by

$$\text{Symmetric Mean Absolute Percentage Error (sMAPE)} = \text{mean}(200|Y_t - F_t|/(Y_t + F_t))$$
$$\text{Symmetric Median Absolute Percentage Error (sMdAPE)} = \text{median}(200|Y_t - F_t|/(Y_t + F_t))$$

The problems arising from small values of Y_t may be less severe for sMAPE and sMdAPE. However, ever there if Y_t is close to zero, F_t is also likely to be close to zero. Thus, the measure still involves division by a number close to zero.

As was seen in the examples in Section 1, sMAPE and sMdAPE can take negative values. It would seem more natural to define them with absolute values in the denominator, and so avoid this problem, but this is not what is usually done. Further, these measures are not as "symmetric" as their name suggests. For the same value of Y_t, the value of $2|Y_t - F_t|/(Y_t + F_t)$ has a heavier penalty when forecasts are low compared to when forecasts are high. See Goodwin and Lawton (1999) and Koehler (2001) for further discussion on this point.

Some authors (e.g., Swanson, et al., 2000) have noted that measures based on percentage errors are often highly skewed, and therefore transformations (such as logarithms) can make them more stable. See Coleman and Swanson (2004) for further discussion.

2.3 Measures based on relative errors

An alternative way of scaling is to divide each error by the error obtained using another standard 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 random walk where F_t is equal to the last observation; this is what was used in the examples in Section 1.

Then we can define:

$$\begin{aligned}\text{Mean Relative Absolute Error (MRAE)} &= \text{mean}(|r_t|)\\ \text{Median Relative Absolute Error (MdRAE)} &= \text{median}(|r_t|)\\ \text{Geometric Mean Relative Absoluate Error (GMRAE)} &= \text{gmean}(|r_t|)\end{aligned}$$

and so on. Armstrong and Collopy (1992) recommended the use of relative absolute errors, especially the GMRAE and MdRAE. Fildes (1992) also prefers the GMRAE although he expresses it in an equivalent (but more complex) form as the square root of the geometric mean of squared relative errors. This equivalence does not seem to have been noticed by any of the discussants in the commentary by Ahlburg et al. (1992).

A serious deficiency in relative error measures is that e_t^* can be small. In fact, r_t has infinite variance because e_t^* has positive probability density at 0. One common special case is when e_t and e_t^* are normally distributed, in which case r_t has a Cauchy distribution.

Armstrong and Collopy (1992) recommend the use of "winsorizing" to trim extreme values. This will avoid the difficulties associated with small values of e_t^*, but adds some complexity to the calculation and a level of arbitrariness as the amount of trimming must be specified.

2.4 Relative measures

Rather than use relative errors, one can use relative measures. For example, let MAE_b denote the MAE from the benchmark method. Then, a relative MAE is given by

$$\text{RelMAE} = \text{MAE}/\text{MAE}_b\,.$$

Similar measures can be defined using RMSEs, MdAEs, MAPEs, etc. Note that Armstrong and Collopy refer to the relative MAE as CumRAE.

When the benchmark method is a random walk, and the forecasts are all one-step forecasts, the relative RMSE is Theil's U statistic (Thiel, 1966, chapter 2), sometimes called $U2$. In fact, Theil's definition is ambiguous and the relative RMSPE with the random walk as a benchmark method is also sometimes called Theil's U statistic (e.g., in Makridakis, Wheelwright and Hyndman, 1998).

Thompson's (1990) LMR measure is simply $\log(\text{RelMSE})$. While this has some nice statistical properties, it is not so easily interpreted which is possibly why it has not been widely used.

The random walk or "naïve" method (where F_t is equal to the last observation) is the most common benchmark method for such calculations, although another frequently used possibility is the mean method (where F_t is equal to the mean of all observations). For seasonal data, the "naïve2" method is sometimes used for comparison; this gives forecasts based on the last observation adjusted for seasonality using classical decomposition (Makridakis, Wheelwright and Hyndman, 1998).

An advantage of these methods is their interpretability. For example relative MAE measures the improvement possible from the proposed forecast method relative to the benchmark forecast method. When RelMAE < 1, the proposed method is better than the benchmark method and when RelMAE > 1, the proposed method is worse than the benchmark method.

However, they require several forecasts on the same series to enable a MAE (or MSE) to be computed. One common situation where it is not possible to use such measures is where one is measuring the out-of-sample forecast accuracy at a single forecast horizon across multiple series. It makes no sense to compute the MAE across series (due to their different scales).

A related approach is to use the percentage of forecasts for which a given method is more accurate than the random walk. This is often known as "Percent Better" and can be expressed as

$$\text{PB(MAE)} = 100\,\text{mean}(I\{\text{MAE} < \text{MAE}_b\})$$
$$\text{PB(MSE)} = 100\,\text{mean}(I\{\text{MSE} < \text{MSE}_b\})$$

However, these give no indication about the amount of improvement possible. Thus, it is possible to have one method that performs very slightly better than the benchmark method for 99 series but much worse on 1 series, thus giving it a PB score of 99 even though the benchmark method is preferable.

3 Scaled errors

Relative measures and measures based on relative errors both try to remove the scale of the data by comparing the forecasts with those obtained from some benchmark forecast method, usually the naïve method. However, they both have problems. Relative errors have a statistical distribution with undefined mean and infinite variance. Relative measures can only be computed when there are several forecasts on the same series, and so cannot be used to measure out-of-sample forecast accuracy at a single forecast horizon.

We propose a new but related idea that is suitable for all situations, by scaling the error based on the *in-sample* MAE from the naïve (random walk) forecast method. Thus, a scaled error is defined as

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

which is clearly 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

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

Related measures such as Root Mean Squared Scaled Error (RMSSE) and Median Absolute Scaled Error (MdASE) can be defined analogously. Billah et al. (2005) used a similar error measure when they computed the absolute value of the forecast error as a percentage of the in-sample standard deviation. However, this approach has the disadvantage that the denominator grows with the sample size for non-stationary series containing a unit root. Scaling by the in-sample naïve MAE only assumes that series has no more than one unit root, which is almost always true for real data.

When MASE < 1, the proposed method gives, on average, smaller errors than the one-step errors from the naïve method. If multi-step forecasts are being computed, it is possible to scale by the in-sample MAE computed from multi-step naïve forecasts.

We propose that measures based on scaled errors should become the standard approach in comparing forecast accuracy across series on different scales. They have a meaningful scale, are widely applicable, and are not subject to the degeneracy problems seen in the examples in Section 1. The only circumstance under which these measures would be infinite or undefined is when *all* historical observations are equal.

Of these measures, we prefer MASE as it is less sensitive to outliers and more easily interpreted than RMSSE, and less variable on small samples than MdASE. If the RMSSE is used, it may be preferable to use the in-sample RMSE from the naïve method in the denominator of q_t.

In some circumstances an asymmmetric loss function may be preferred (see, e.g., Lawrence & O'Connor, 2005), in which case some other (asymmetric) function of the scaled errors may be appropriate. Diebold, 2001, gives some examples of suitable asymmetric functions.

4 Application to M3-competition data

We demonstrate the use of MASE using the M3-competition data (Makridakis & Hibon, 2000). Figure 2 shows the MASE at each forecast horizon for five forecasting methods applied to the M3-competition data. The errors have been scaled by the one-step in-sample forecast errors from the naïve method, and then averaged across all series. So a value of 2 indicates that the out-of-sample forecast errors are, on average, about twice as large as the in-sample one-step forecast errors from the naïve method. Because the scaling is based on one-step forecasts, the scaled errors for multi-step forecasts are typically larger than one. The methods Theta, RobustTrend and ForecastPro were part of the M3-competition, and are described in Makridakis & Hibon (2000). The HKSG method uses the state space modelling approach of Hyndman, Koehler, Snyder and Grose (2002), but only including the additive models.

Table 4 gives the MASE for each of the M3 methods along with the HKSG method. Here, the absolute scaled errors have been averaged across all out-of-sample forecast horizons used in the M3 competition, and then averaged across all series. The best performing method in each category is highlighted with its MASE in bold.

Comparing Table 4 with the results of the original M3 analysis (Makridakis & Hibon, 2000) shows that MASE does not substantially affect the main conclusions about the best-performing methods. In particular, as with the original M3 analysis, the Theta method (Assimakopoulos &

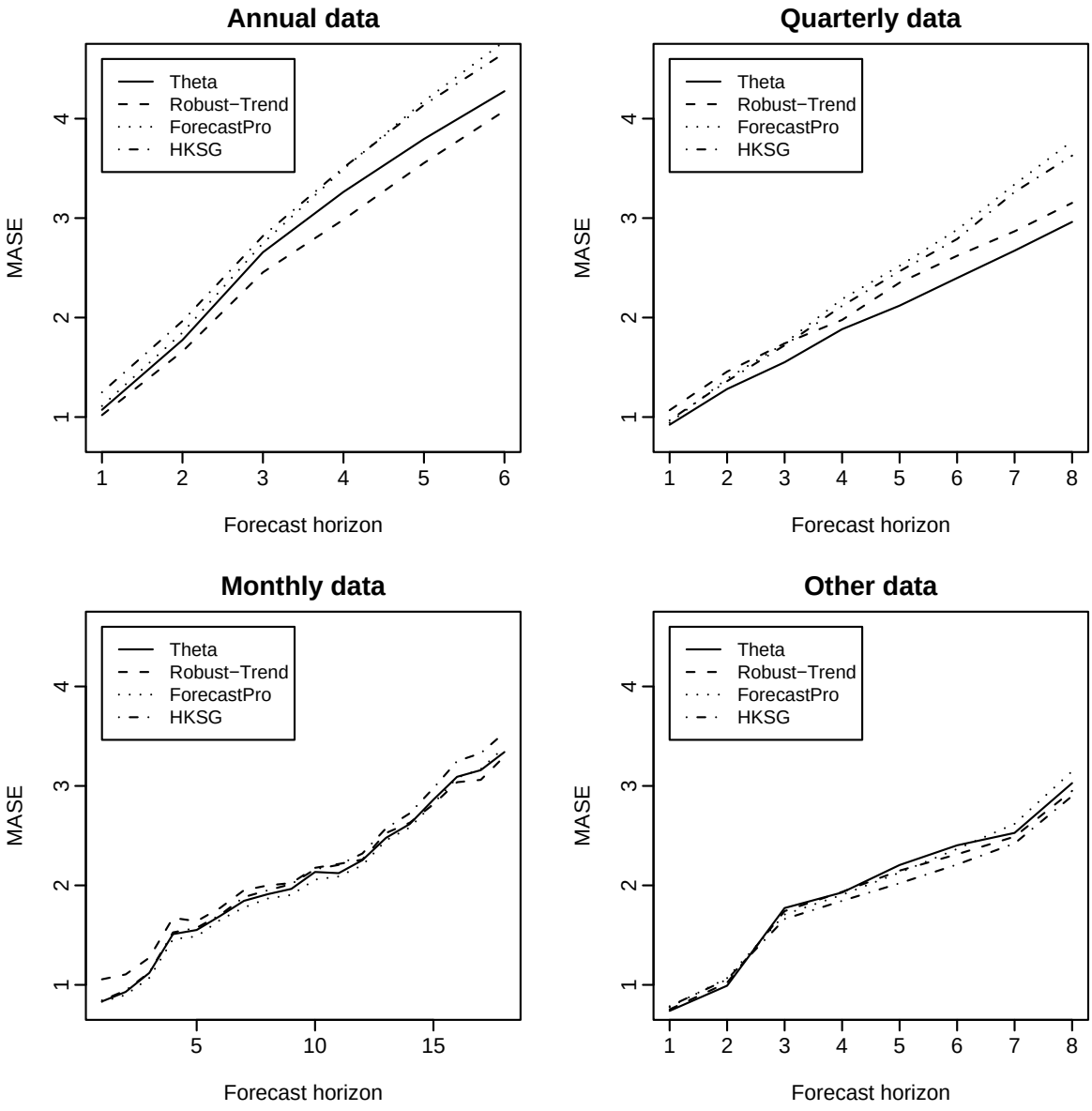

Figure 2: *Mean Absolute Scaled Errors at different forecast horizons for five forecasting methods applied to the M3-competition data.*

	Yearly	Quarterly	Monthly	Other	All
Theta	2.81	**1.97**	2.08	1.95	**2.20**
Theta-sm	2.81	2.00	2.09	1.95	2.22
Robust-Trend	**2.63**	2.15	2.14	1.92	2.23
Comb SHD	2.88	2.05	2.12	2.09	2.26
ForecastX	2.77	2.22	2.20	1.97	2.31
ForecastPro	3.03	2.35	**2.04**	1.96	2.33
Dampen	3.03	2.10	2.18	2.08	2.34
HKSG	3.06	2.29	2.15	**1.86**	2.36
RBF	2.72	2.19	2.27	2.70	2.37
BJ automatic	3.16	2.21	2.21	2.30	2.42
Flores/Pearce1	2.94	2.23	2.31	2.27	2.42
Holt	3.18	2.40	2.15	2.04	2.43
ARARMA	3.48	2.29	2.07	2.05	2.43
SmartFcs	3.00	2.39	2.23	2.08	2.43
PP-autocast	3.02	2.12	2.44	2.09	2.46
Pegels	3.58	2.30	2.13	1.85	2.47
Flores/Pearce2	3.02	2.41	2.27	2.34	2.47
Autobox3	3.18	2.45	2.23	2.01	2.47
Automatic ANN	3.06	2.35	2.34	2.13	2.49
Winter	3.18	2.37	2.43	2.04	2.55
SES	3.17	2.27	2.44	3.14	2.59
Autobox1	3.68	2.61	2.20	2.12	2.62
Naive2	3.17	2.28	2.50	3.13	2.62
Autobox2	2.75	2.20	3.39	1.90	2.87

Table 4: *Mean Absolute Scaled Error for the M3-forecasting competition. All methods were participants in the M3-competition except for HKSG which is based on the method of Hyndman, Koehler, Snyder and Grose (2002).*

Nikolopoulos, 2000) does very well. Hyndman & Billah (2003) demonstrated that this method is equivalent to simple exponential smoothing (SES) with drift where the drift is half the value of the slope of a linear regression fitted to the data. Thus, it provides a form of shrinkage which limits the ability of the model to produce anything wildly inaccurate.

5 Conclusion

Despite two decades of papers on measures of forecast error, we believe that some fundamental problems have been overlooked. In particular, the measures used in the M-competition and the M3-competition, and the measures recommended by other authors, all have problems—they can give infinite or undefined values in commonly occurring situations.

We propose that scaled errors become the standard measure for forecast accuracy, where the forecast error is scaled by the in-sample mean absolute error obtained using the naïve forecast-

ing method. This is widely applicable, and is always defined and finite except in the irrelevant case where all historical data are equal. This new measure is also easily interpretable: values of MASE greater than one indicate the forecasts are worse, on average, than in-sample one-step forecasts from the naive method.

Of course, there will be situations where some of the existing measures may still be preferred. For example, if all series are on the same scale, then the MAE may be preferred because it is simpler to explain. If all data are positive and much greater than zero, the MAPE may still be preferred for reasons of simplicity. However, in situations where there are very different scales including data which are close to zero or negative, we suggest the MASE is the best available measure of forecast accuracy.

6 Acknowledgments

We thank Michelle Hibon for kindly providing the forecasts submitted to the M3-competition, and two anonymous referees for providing some thoughtful comments.

References

AHLBURG, D.A., CHATFIELD, C., TAYLOR, S.J., THOMPSON, P.A., WINKLER, R.L., MURPHY, A.H., COLLOPY, F., and FILDES, R. (1992) A commentary on error measures. *International J. Forecasting*, **8**, 99–111.

ARMSTRONG, J.S. (2001) "Evaluating forecasting methods", Chapter 14 in *Principles of forecasting: a handbook for researchers and practitioners*, ed., J.S. Armstrong. Kluwer Academic Publishers: Norwell, MA.

ARMSTRONG, J.S., and COLLOPY, F. (1992) Error measures for generalizing about forecasting methods: empirical comparisons. *International J Forecasting*, **8**, 69–80.

ASSIMAKOPOULOS, V. and NIKOLOPOULOS, K. (2000) The theta model: a decomposition approach to forecasting. *International Journal of Forecasting* **16**, 521–530.

BILLAH, B., KING, M.L., SNYDER, R.D., and KOEHLER, A.B. (2005) Exponential smoothing model selection for forecasting. Working Paper 6/05, Department of Econometrics and Business Statistics, Monash University, Australia.

BOWERMAN, B.L., O'CONNELL, R.T., and KOEHLER, A.B. (2004) *Forecasting, time series and regression: an applied approach*, Thomson Brooks/Cole: Belmont, CA.

CHATFIELD, C. (1988) Apples, oranges and mean square error. *International J. Forecasting*, **4**, 515–518.

CLEMENTS, M.P., AND HENDRY, D.F. (1993) On the limitations of comparing mean square forecast errors. *J. Forecasting*, **12**, 617–637.

COLEMAN, C.D., and SWANSON, D.A. (2004) On MAPE-R as a measure of estimation and forecast accuracy. Working paper. Center for Population Studies, University of Mississippi. Accessed 18 May 2005.
`http://www.olemiss.edu/depts/population_studies/WorkingPapers.html`

COLLOPY, F., and ARMSTRONG, J.S. (2000) Another error measure for selection of the best forecasting method: the unbiased absolute percentage error. Updated October 2000. Accessed 18 May 2005.
`http://hops.wharton.upenn.edu/forecast/paperpdf/armstrong-unbiasedAPE.pdf`

DIEBOLD, F.X. (2001) *Elements of forecasting*, 2nd ed., South-Western: Cincinnati, Ohio.

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

GARDNER, E. (1990) Evaluating forecast performance in an inventory control system. *Management Science*, **36**, 490–499.

GOODWIN, P., and LAWTON, R. (1999) On the asymmetry of the symmetric MAPE. *International Journal of Forecasting*, **4**, 405–408.

GRANGER, C.W.J., and PESARAN, M.H. (2000) Economic and statistical measures of forecast accuracy. *J. Forecasting*, **19**, 537–560.

HANKE, J.E., and REITSCH, A.G. (1995) *Business forecasting*, 5th ed., Prentice-Hall: Englewood Cliffs, NJ.

HYNDMAN, R.J., KOEHLER, A.B., SNYDER, R.D., and GROSE, S. (2002) A state space framework for automatic forecasting using exponential smoothing methods. *International J. Forecasting*, 18(3), 439–454.

KOEHLER, A.B. (2001) The asymmetry of the sAPE measure and other comments on the M3-competition. *International J. Forecasting*, **17**, 537–584.

LAWRENCE, M., and O'CONNOR, M. (2005) Judgmental forecasting in the presence of loss functions. *International J. Forecasting*, **21**, 3–14.

MAKRIDAKIS, S. (1993) Accuracy measures: theoretical and practical concerns. *International J. Forecasting*, **9**, 527–529.

MAKRIDAKIS, S., ANDERSON, A., CARBONE, R., FILDES, R., HIBON, M., LEWANDOWSKI, R., NEWTON, J., PARZEN, P., and WINKLER, R. (1982) The accuracy of extrapolation (time series) methods: results of a forecasting competition. *J. Forecasting*, **1**, 111–153.

MAKRIDAKIS, S., and HIBON, M. (2000) The M3-competition: results, conclusions and implications. *International J. Forecasting*, **16**, 451–476.

MAKRIDAKIS, S., WHEELWRIGHT, S., and HYNDMAN, R.J. (1998) *Forecasting: methods and applications*, 3rd ed., John Wiley & Sons: New York.

THEIL, H. (1966) *Applied economic forecasting*. Rand McNally: Chicago, IL.

THOMPSON, P.A. (1990) An MSE statistic for comparing forecast accuracy across series. *International J. Forecasting*, **6** 219–227.

TSAY, R.S. (2002) *Analysis of financial time series*, John Wiley & Sons: New York.

SWANSON, D.A., TAYMAN, J., and BARR, C.F. (2000) A note on the measurement of accuracy for subnational demographic estimates. *Demography*, **2**, 193–201.