International Journal of Forecasting 9 (1993) 527–529

Notes

The Notes section of the *International Journal of Forecasting* contains commentary on the theory and practice of forecasting in the form of communications to the journal such as research notes, teaching tips, practitioners' and consultants' views, and other contributions, especially those that attempt to bridge the gap between new developments in the methodology of forecasting and their practical application. Contributions to this section of the journal can be submitted to any of the four editors.

Accuracy measures: theoretical and practical concerns

Spyros Makridakis
INSEAD, Fontainebleau, France

There has been renewed concern recently [Fildes (1992); Armstrong and Collopy (1992)] about the most appropriate accuracy measure to be used for evaluating forecasting methods and for reporting error statistics. The purpose of this note is to examine accuracy measures from a theoretical and practical point of view and to suggest a modified form of MAPE as the most appropriate measure satisfying both theoretical *and* practical concerns while allowing meaningful relative comparisons.

Accuracy measures, error statistics or measures, and loss functions are alternative ways of conveying information about the ability of a certain forecasting method to predict actual data, either when a model is fitted to such data, or for future periods (post-sample) whose values have not been used to develop the forecasting model. Research has shown that post-sample accuracies are not always related to those of the model that best fits available historical data [Makridakis (1986)]. The correlations between the two are small to start with [0.22 for the first forecasting horizon; Makridakis (1986)] and become equal to zero for horizons longer than four periods ahead. This means that we must judge the appropriateness of whichever measure we use by how effectively it provides information about post-sample accuracies.

For post-sample comparisons, research findings indicate that the performance (accuracy) of different methods depends upon the accuracy measure used. This means that some methods are better when, for example, MAPEs are used while others are better when rankings are utilized, although the various accuracy measures are clearly correlated [see Armstrong and Collopy (1992)]. From a theoretical point of view there is a problem as no single method can be designated as the 'best', although those methods that have been found to perform badly in all accuracy measures can be excluded. This means that the 'best' method will have to be related to the purpose of forecasting—its value for improving decision making and the needs and concerns of the person or situation using the forecasts. Thus, in a one-time auction the method that comes up the best most of the time is to be preferred (e.g.

using rankings or percentage better measures). In empirical comparisons several measures may have to be used to calibrate the results, while in budgeting, MAPEs may be most appropriate as they convey information about average percentage errors which are used to a great extent in reporting accounting results and profits.

Is there a best overall measure that can be used in the great majority of situations and which satisfies both theoretical and practical concerns? From both a theoretical and practical point of view such an accuracy measure must be relative; otherwise, we compare apples and oranges in ways that make little sense [see Fildes and Makridakis (1988); Chatfield, (1988)]. From a theoretical point of view this accuracy measure must be robust from one situation (or data set) to another and not to be unduly influenced by outliers. From a practical point of view it must make sense, be easily understood, and convey as much information about accuracy (or errors) as possible. Finally, we must distinguish between academic research focused mainly on evaluating forecasting competitions and other large scale accuracy studies and reporting the performance of methods or forecasters in business, government or miltary applications. While the former are concerned about averages, the latter want to know what happens in specific cases for particular periods of time.

Computational considerations dictate that, whatever measure is being used, the possibility of division by zero (or a very small number) must *never* exist. Moreover, it it important that we do not divide by a very large number. Otherwise we must set arbitrary upper, or lower, limits that make comparisons doubtful (for instance, the author found the Theil's *U*-Statistic [Makridakis and Hibon (1979)] highly problematic because in several instances the divisor was zero). For this reason I believe that measures such as Geometric Mean or Median using RAEs (Relative Absolute Errors) are not appropriate since their divisor is the difference between the actual values and those of a random walk model which can either be zero (or close to zero), or alternatively a very large value. Winsorizing the RAEs by setting an upper limit of 10 and a lower limit of 0.01 recommended by Armstrong and Collopy (1992) creates the serious problem of non-continuous scales and begs the question of why not 5, 20, or 40 instead of 10; or 0.005, 0.02, or 0.04 instead of 0.01. This problem of setting arbitrary upper and lower limits becomes extreme when the number of series involved is small. Equally important, RAE-based measures mean *absolutely* nothing to decision makers who cannot understand either their meaning or grasp the non-linear scales being reported. What is meant by a Geometric Mean RAE of 0.25, or a Median RAE of 0.75? In my view not much, even if one is familiar with statistical measures. Moreover, Geometric Means cannot be easily computed when a large number of series is involved. For the above reasons, I do not believe that either GMRAE or the MdRAE are appropriate accuracy measures except for specific cases when the number of series involved is neither very small nor very large, when *no* winsorizing is required *and* when the results are reported exclusively for statistical audiences. This is more so as GMRAE and MdRAE do not perform better than MAPE [see exhibit 9 in Armstrong and Collopy, (1992)] over the criteria of reliability, validity, outlier protection, sensitivity, and value to decision making.

As rankings and medians are not relative measures, they are not recommended for general use. The percentage better measure is not appropriate for reporting results (it can only be used in large scale empirical studies) while mean square errors are neither relative nor convey much meaning to decision makers. This leaves the MAPE as the only remaining choice. MAPE is a relative measure that incorporates the best characteristics among the various accuracy criteria. Moreover, it is the only measure (in addition to Percent Better) that means something to decision makers who have trouble even understanding medians, not to mention geometric means. In addition, it can be used for both evaluating large-scale empirical studies and for presenting specific results. Thus, we must look for ways of correcting the disadvantages of MAPE rather than searching for alternative measures which are less desirable than MAPE, even with its current disadvantages, and whose meaning is difficult to understand.

MAPE as an accuracy measure can be influenced by four problems.

(1) Equal errors above the actual value result in a greater APE (Absolute Percentage Error)

than those below the actual value. For instance, when the actual value is 150 and the forecast is 100 (an error of 50) the APE is:

$$\text{APE}_t = \left|\frac{A_t - F_t}{A_t}\right| = \frac{150 - 100}{150} = \frac{50}{150} = 33.33\%$$

where A_t is the actual and F_t the forecast value at period t. However, when the actual is 100 and the forecast 150 the APE is:

$$\text{APE} = \left|\frac{100 - 150}{100}\right| = \frac{50}{100} = 50\%$$

This problem can be easily corrected by dividing the error $(A_t - F_t)$ by the average of both A_t and F_t, i.e.

$$\text{APE}_t = \left|\frac{A_t - F_t}{(A_t + F_t)/2}\right|$$

The above formula will provide the APE of 40% in both cases.

(2) When the value of A_t is small (usually less than 1), it can provide large percentage errors. This problem, although not common, can be solved by excluding from the averaging of absolute percentage errors the series which have values of less than 1 (or some similar cut-off point) as otherwise some extremely large percentage errors can result which can render the MAPE computed meaningless.

(3) In case of unusually large errors, in particular when the value of A_t is small, some Absolute Percentage Error (APE) can become extremely large (outliers) and distort the comparisons in forecasting competitions or empirical studies. This problem can be solved by reporting the MAPEs with *and* without outliers (where an outlier can be defined as any value greater than the MAPE plus, say, three standard deviations). This problem, however, does not exist in practice when reporting forecasting errors, as large errors and outliers are costly and undesirable and must be known to decision makers.

(4) MAPEs cannot be compared directly with Naive 1 (random walk) or Naive 2 (deseasonalized random walk) models as is the case with RAE and therefore Geometric Means and Relative Medians which summarize them. This problem can be elevated by computing the MAPE of Naive 1 and Naive 2 and then showing the difference of the MAPE of a certain method minus that of the MAPE of Naive 1 and Naive 2. This difference tells decision makers the improvement in forecasting accuracy of the method used over an above that of using the last available data as the forecast (with or without seasonality). Such a difference provides similar information as the Geometric Mean or Relative Median while excluding the possibility of dividing by zero. Finally, by also providing the difference between the MAPE of Naive 2 minus that of Naive 1, we can know how much removing just the seasonality improves the forecasts of a certain method.

We (Michèle Hibon and myself) are, at present, re-computing the results of the M-Competition [Makridakis et al. (1982)] using the modified MAPE described above. The preliminary results seem promising. When nine subsamples are used the modified MAPEs are more stable and include fewer outliers. Also, their correlations from one subsample/method to another increase. Although the full results are not yet available, we wish to make the point now that a modification of MAPE may be the most appropriate way to meet theoretical and practical concerns, and do so in a simple and meaningful way.

References

Armstrong, J.S. and F. Collopy, 1992, "Error measures for generalizing about forecasting methods: Empirical comparisons with discussion", International Journal of Forecasting, 8, 69–80, 99–111.

Chatfield, C., 1988, "Apples, oranges and mean square error", International Journal of Forecasting, 4, 515–518.

Fildes, R., 1992, "The evaluation of extrapolative forecasting methods", International Journal of Forecasting, 8, 88–98.

Fildes, R. and S. Makridakis, 1988, "Loss functions and forecasting", International Journal of Forecasting, 4, 545–550.

Makridakis, S., 1986, "The art and science of forecasting: An assessment and future directions", International Journal of Forecasting, 2, 15–39.

Makridakis, S. and M. Hibon, 1979, "Accuracy of forecasting: An empirical investigation with discussion", Journal of the Royal Statistical Society (A), 142, 97–145.

Makridakis, S., A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen and R. Winkler, 1982, "The accuracy of extrapolation (time series) methods: Results of a forecasting competition", Journal of Forecasting, 1, 111–153.