Another Error Measure for Selection of the Best Forecasting Method: The Unbiased Absolute Percentage Error

Fred Collopy
The Weatherhead School of Management
Case Western Reserve University
Cleveland, Ohio 44106

J. Scott Armstrong
The Wharton School
University of Pennsylvania
Philadelphia, PA 19104

Updated October 2000
[Published only on the Internet: cite as
http://hops.wharton.upenn.edu/forecast/paperpdf/armstrong-unbiasedAPE.pdf]

Research has suggested that the selection of an error measure has an important effect on the conclusions about which of a set of forecasting methods is most accurate (Armstrong and Collopy 1992; Fildes 1992). This research had concluded that the Relative Absolute Error (RAE) is a useful measure, especially when making comparisons across a small set of time series, where the time series differ substantially (Armstrong and Collopy 1992). Makridakis (1993) discussed some problems with the RAE and proposed an alternative measure. We refer to his proposed measure as the Unbiased Absolute Percentage Error (or UAPE). We re-examine problems with the RAE and then provide empirical evidence on the performance of the UAPE. First, however, we provide some historical perspectives that led to the development of the RAE.

History of the RAE

Originally, error measures for assessing the performance of forecasting models were developed by statisticians. Because of its computational convenience and theoretical relevance to statistics, they proposed the Mean Square Error (MSE). The MSE was the primary measure used for comparing forecasting methods for a long time. Empirical research (e.g., Chatfield 1988; Armstrong and Collopy 1992; Fildes 1992) has shown that the MSE is inappropriate for comparing methods, primarily because it is unreliable. A variety of other error measures are better candidates for such comparisons, though the research to date has not produced definitive conclusions about which to use.

A consensus does seem to be emerging, though, that relative error measures are most relevant. The roots of these measures date back at least to Ohlin and Duncan (1949). They created an Index of Predictive Efficiency by taking the difference of the errors from two forecasting methods and dividing by the error of one of the methods. Theil (1966) proposed his U2, computed by dividing the mean square error for a proposed model by the mean square error for the random walk. Armstrong and Collopy (1992) built upon these foundations to propose the Relative Absolute Error or RAE, where the absolute error of the proposed model is divided by the absolute error of the random walk (Theil's U2 and the RAE are equivalent for a single forecast for a single horizon). In short, the RAE rests on a long tradition of relative error measures.

An Alternative to the RAE

Makridakis (1993) questioned the value of the RAE and proposed instead a modification of the MAPE (Mean Absolute Percentage Error). Specifically he suggested an adjustment to the denominator of the APE so that the average of the actual and the forecast values is used instead of the actual value alone (as is done in standard computations of the APE). Armstrong and Collopy (1992) had referred to this as the adjusted APE. It would seem more descriptive to name it for its primary benefit. As noted by Makridakis, it is unbiased with respect to the scale of the error, so we refer to it here as the unbiased absolute percentage error (UAPE). Armstrong (1985, pp. 348 and 355) discusses the characteristics of this error measure and compares it with other measures.

Makridakis suggests that the RAE is not very meaningful for decision-making (other than making the decision about which method to use). We agree. But we believe that the UAPE may also lack meaning. With the APE, the manager is told that the error was, say, fifteen percent of the actual. For the UAPE, the manager is told that the error is fifteen percent of the average of the actual and the forecast that was produced for that period. *We* find it a bit more confusing, but whether managers will also be confused is an empirical issue that has not been investigated.

Makridakis suggested that one way in which the UAPE is superior to the RAE is that it is easier to summarize across series. A simple average can be used. We agree that this is conceptually simpler and that it is more understandable to managers than are geometric means.[1]

The UAPE is constrained to be between 0 and 200 and it avoids the APE's bias in favor of low forecasts. In effect, the error is symmetric with respect to the scale of the errors. The UAPE also avoids problems associated with dividing by small numbers when the actual is close to zero. Furthermore, it avoids the need for trimming.[2] Given these characteristics, it seemed reasonable to expect that it might offer improved reliability over the APE.

To assess the reliability of the UAPE, we followed a procedure used in Armstrong and Collopy (1992), which compared the average correlations among rankings for subsets of M-competition data. For annual data, we ranked 11 forecasting methods for accuracy of one-year ahead forecasts according to each of the proposed error measures. We did this for five subsets of 18 series each, using the arithmetic mean. We ranked the methods according to their accuracy and calculated Spearman correlations for the ten pairwise comparisons, which were then averaged. We did the same thing for the six-year ahead forecasts. For quarterly data, we followed the same procedure using 12 forecasting methods and 20 series in each subsample (one had 21 series). The results are presented in the Exhibit. The column labeled "average" provides the average of 40 pairwise correlations.

---

[1] Makridakis also says that "Geometric means cannot easily be computed when a large number of series is involved." This is true only if you directly calculate the geometric mean (using multiplications and then taking the nth root). However, one can transform to logs, add the logs, take an average, then take the anti-log of the average. This can be done with widely-used spreadsheet programs.

[2] The issue of trimming is a complex one. For example, while Makridakis criticized our use of Winsorizing, a trimming procedure that substitutes a boundary value for any observation beyond that boundary, he also proposed a form of trimming for certain situations. In point #2 on page 529, he said that when the value of A_t is small "usually less than 1," it should be "excluded from the averaging." This is a more severe trimming procedure than Winsorizing because it eliminates all information about that value. In point #3 (page 529)M where he deals with large errors, Makridakis proposed a severe trimming procedure, as he suggested that one should present one summary with outliers excluded and another with them included.

Exhibit
Reliability of the error measures
(average Spearman correlations for pairwise comparison among five subsamples).

	Quarterly		Annual		Average
	1-ahead	**8-ahead**	**1-ahead**	**6-ahead**	
MAPE	0.55	0.61	0.49	0.30	0.49
GMRAE	0.34	0.21	0.81	0.74	0.52
MdRAE	0.29	0.43	0.79	0.72	0.56
Mean UAPE	0.63	-0.04	0.53	0.56	0.42
Median UAPE	0.38	0.31	0.75	0.69	0.53

The average correlation of the rankings using the Mean UAPE was 0.42, in contrast to the MAPE's 0.48. This result was surprising to us. However, it should be noted that the sample sizes were small and that the estimates varied widely across the conditions that we examined.

The Median UAPE's average correlation of 0.53, on the other hand, suggests that it had about the same reliability as the Geometric Mean of the RAE (GMRAE), with a correlation of 0.52 and as the Median RAE (MdRAE), with a correlation of 0.56. Here also, tests with many more subsets would be desirable.

Makridakis suggests the creation of a relative measure by taking the difference between the UAPE for the proposed method and that for the seasonally-adjusted random walk for each forecasted value. While this may aid interpretation, the rank order of the forecasts would not change under this transformation as the same constant is being subtracted from the forecasts from each method.

Summary

The search for the most effective forecast error measures for making comparisons across series is still in process and Makridakis proposed a measure that we have referred to the UAPE. The UAPE has desirable features. As noted it is a relative measure and it is unbiased. It needs no trimming and it can be summarized using the arithmetic mean. The reliability of the mean UAPE was not as good as the GMRAE's, but that of the Median UAPE was comparable to that for the GMRAE and the MdRAE. In effect, then, trimming was useful in this test. Given that it is unbiased and that it has a simpler summary measure, the UAPE is worthy of further research. Pending the results of further research, however, we recommend the use of relative measures such as the MdRAE. The RAE is based on a long tradition of relative error measures.

Acknowledgements: Partial support for this research has been provided by the U.S. Navy Personnel R & D Center and by the Office of Naval Research (under grant number N00014-92-J-1544).

References

Armstrong, J. Scott , 1985, *Long-Range Forecasting: From Crystal Ball to Compute*r. New York: John Wiley.

Armstrong, J. Scott and F. Collopy, 1992, "Error measures for generalizing about forecasting methods: Empirical comparisons," *International Journal of Forecasting*, 8, 69-80.

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

Fildes, Robert, 1992, "The evaluation of extrapolative forecasting methods," *International Journal of Forecasting*, 8, 81-98.

Makridakis, Spyros, 1993, "Accuracy measures: Theoretical and practical concerns," *International Journal of Forecasting*, 9, 527-529.

Ohlin, Lloyd E. and O. D. Duncan, 1949, "The efficiency of prediction in criminology," *American Journal of Sociology*, 54, 442-452.

Theil, Henri, 1966, *Applied Economic Forecasting*. Chicago: Rand McNally,