Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons

By J. Scott Armstrong and Fred Collopy

Reprinted with permission form *International Journal of Forecasting*, 8 (1992), 69-80.

Abstract: This study evaluated measures for making comparisons of errors across time series. We analyzed 90 annual and 101 quarterly economic time series. We judged error measures on reliability, construct validity, sensitivity to small changes, protection against outliers, and their relationship to decision making. The results lead us to recommend the Geometric Mean of the Relative Absolute Error (GMRAE) when the task involves calibrating a model for a set of time series. The GMRAE compares the absolute error of a given method to that from the random walk forecast. For selecting the most accurate methods, we recommend the Median RAE (MdRAE) when few series are available and the Median Absolute Percentage Error (MdAPE) otherwise. The Root Mean Square Error (RMSE) is not reliable, and is therefore inappropriate for comparing accuracy across series.

Keywords: Forecast accuracy, M-Competition, Relative absolute error, Theil's U.

1. Introduction

Over the past-two decades, many studies have been conducted to identify which method will provide the most accurate forecasts for a given class of time series. Such generalizations are important because organizations often rely upon a single method for a given type of data. For example, a company might find that the Holt-Winters' exponential smoothing method is accurate for most of its series, and thus decide to base its cash flow forecasts on this method. This paper examines error measures for drawing conclusions about the relative accuracy of extrapolation methods.

Conclusions about the accuracy of various forecasting methods typically require comparisons across many time series. However, it is often difficult to obtain a large number of series. This is particularly a problem when trying to specify the best method for a well-defined set of conditions; the more specific the conditions, the greater the difficulty in obtaining many series. Thus, it is important to identify which error measures are useful given few series, and which are appropriate for a larger number of series.

Error measures also play an important role in calibrating or refining a model so that it will forecast accurately for a set of time series. That is, given a set of time series, the analyst may wish to examine the effects of using different parameters in an effort to improve a model.

We first discuss issues involved in comparing the accuracy of different forecasting methods across time series. Next, we provide empirical comparisons of six error measures with respect to reliability, construct validity, and other criteria. Then we examine the selection of

errors when one has a small number of series. Finally, we provide guidelines for the selection of an error measure.

2. Making comparisons across series

Comparisons of errors across series typically involve many methods and many series. Because the use of multiple measures can be cumbersome, a single error measure is desirable. We expected the choice of an error measure to vary, however, according to the situation. In particular, we expected the choice to depend on the number of time series available and on whether the task is to select the most accurate method or to calibrate a given model.

This section examines technical issues that arise in the choice of an error measure. These include differences in scale across the series, the amount of change that occurs over the forecast horizon, and the presence of extreme forecast errors (outliers). We also discuss ways of summarizing error measures across series.

2.1. Scaling

The scale of the data often varies considerably among series. Series with large numbers might dominate comparisons. Despite this problem, the need for unit-free measures was not widely appreciated in the early 1980s as shown by Carbone and Armstrong (1982). They asked 145 forecasting experts what error measures they preferred when generalizing about the accuracy of different forecasting methods. Practitioners selected the Root Mean Square Error (RMSE) more frequently than any other measure, although it is not unit-free. Academicians had an even stronger preference for the RMSE.

The RMSE has been used frequently to draw conclusions about forecasting methods. For example, Zellner (1986) claimed that the Bayesian method was the most accurate method in the M-competition because its RMSE was lowest. However, Chatfield (1988), in a re-examination of the M-Competition data, showed that five of the 1001 series dominated the RMSE rankings. The remaining 996 series had little impact on the RMSE rankings of the forecasting methods.

Researchers now seem to prefer unit-free measures for comparing methods. One such measure is the percentage of forecasts for which a given method is more accurate than the random walk (Percent Better). Another way to control for scale is to use percentage errors; that is, to calculate the error as a percentage of the actual value. Perhaps the most widely used unit-free measure is the Mean Absolute Percentage Error (MAPE). A disadvantage of the MAPE is that it is relevant only for ratio-scaled data (i.e., data with a meaningful zero). (Our study, however, examined economic and demographic data, which typically involve ratio-scaled data.) Another disadvantage of the MAPE is that it puts a heavier penalty on forecasts that exceed the actual than on those that are less than the actual. For example, the MAPE is bounded on the low side by an error of 100%, but there is no bound on the high side.

2.2. Amount of change

Predictions are more difficult for series where large changes occur over the forecast horizon. The Percent Better avoids this problem by discarding information about the amount of change. Another approach is to employ relative errors and compare the forecast errors from a given model against those from another model. One candidate for an alternative model is the random walk. The random walk is attractive because it is simple and easily interpreted. Theil's U2 [Theil (1966, ch. 2)] compares the RMSE for a proposed model with the RMSE for the random walk [see Bliemel (1973) for a discussion of Theil's measure]. Theil's U2 has not been widely used by forecasters outside of economics. Only two percent of the 145 academicians and practitioners in the survey by Carbone and Armstrong (1982) preferred Theil's measure for the evaluation of forecasting methods.

We propose the Relative Absolute Error, or RAE, as a simple and easily communicated alternative to Theil's U2. It is calculated for a given time series by dividing the absolute forecast error (at a given horizon, *h*) for a proposed model ($| F_{m\,h} - A_{\,h} |$) by the corresponding error for the random walk ($| F_{rw\,h} - A_{\,h} |$) Thus, if the proposed model had a forecast of 120, while the random walk had a forecast of 105 and the actual was 115, the RAE would be the absolute value of 120 minus 115 divided by the absolute value of 105 minus 115 (i.e., 5/10 or 0.50).

For a single horizon, the RAE is equivalent to the relative absolute percentage error (the ratio of the absolute percentage error for a proposed model divided by the absolute percentage error for the random walk). Thus, the relative error measure adds another level of control (for change) to the MAPE, which, in turn, adds control (for scale) to the mean absolute error. For a single horizon (e.g., an annual forecast for year h in the future), the RAE and Theil's U2 are equivalent. They are not equivalent when cumulated over a forecast horizon, as we describe later.

We believe that the primary advantage of the RAE over Theil's U2 is the ease of interpretation and communication. To see its advantage, try to explain Theil's U2 and the RAE to a practitioner. Or ask forecasters to describe Theil's U2 to you.

Relative error measures do not relate closely to the economic benefits associated with the use of a particular forecasting method. For example, for the RAE, progressive reductions of errors from 40 to 20, then from 20 to 10, then from 10 to 5, and so forth would all be judged to be of equal importance. Thus, relative error measures are typically inappropriate for managerial decision making.

2.3. Outliers

A single series may dominate the analysis because it has a much larger or smaller error than that found for other time series in the summary. This situation might arise because of a mistake in recording the data. Outliers are especially troublesome when the goal is to select from among a set of forecasting methods. They are less of a problem for calibrating a model.

The Percent Better (than a comparison model) is immune to outliers. The RAE, by controlling for differences in scale and the amount of change that occurs in series over the forecast horizon, is less likely to be affected by outliers. Nevertheless, the RAE can explode for series where the error from the random walk is very small. For example, if the random walk forecast had no error and the proposed model had some error, their ratio would be infinite. Outliers also can occur with small errors for the proposed model.

2.4. Summarizing across series

Various measures of central tendency are available to summarize the errors of each forecasting method across a set of time series. Arithmetic means can be used to summarize the Percent Better and the APE (thus providing the MAPE). Geometric means can be used to summarize relative error measures. For the RAE, then, we use the Geometric Mean of the RAE (GMRAE). We also use a geometric mean to summarize Theil's U2.

Outliers create problems when one tries to summarize across series. For example, if the RAE for one series in a set were infinite, the GMRAE across this set of series would be infinite. Similarly, if the RAE for a single series were zero, the GMRAE would be zero. The effect of outliers can be reduced by trimming, which discards high and low errors. Using medians is an extreme way to trim as it removes all values higher and lower than the middle value. The Median APE (MdAPE) reduces the bias in favor of low forecasts, thus offering an advantage over the MAPE. The MdAPE also provides a standard trimming rule, and this aids in comparisons across studies that are reported in the literature. Winsorizing, which replaces extreme values with certain limits, can also be used to temper the impact of outliers. To its advantage, Winsorizing retains some information about the high and low errors. However, it does raise an issue as to the proper limits. We Winsorized all summaries based on the RAE. Values less than 0.01 were replaced by 0.01, and those greater than 10 were replaced by 10. The Appendix details how error measures were summarized in this study.

3. Comparing the error measures

This section describes the data, reliability, construct validity, and other criteria. The statistical analysis was done using the Macintosh version of Statview. An associate independently confirmed the results using SPSS-X (version 4) on a Sun computer. This direct replication identified one data entry error and one procedural error.

3.1. Data

The data and forecasts were drawn from annual and quarterly time series from the M-competition [Makridakis et al. (1982)].[1] They were stratified by demographic, macroeconomic, industry, and company categories. Ten equal-sized stratified subsamples were created; they contained 18 annual and 20 quarterly series (one subsample of quarterly data contained 21 rather

[1] These data and forecasts were provided by Everette Gardner, Michele Hibon, and Spyros Makridakis. For these analyses, we used all annual and quarterly series whose identification numbers ended in a 2, 3, 4, 6, or 7.

than 20 series). Then we randomly selected five annual and quarterly subsamples. This procedure yielded a total of 90 annual series and 101 quarterly series.

The data, all ratio scaled, come from several countries and from different time spans. We used only one starting point to forecast each series and the starting points differed among series.

3.2. Reliability

Reliability addresses the question of whether repeated application of a procedure will produce similar results. To assess reliability, we examined the extent to which an error measure produced the same accuracy rankings for extrapolation methods when it was applied to different samples from a set of time series. Specifically, we calculated one-year-ahead forecast errors for each forecasting method for all five subsamples of 18 annual series each. We ranked 11 forecasting methods: linear trend, moving average, single exponential smoothing, ARR exponential smoothing, Holt's exponential smoothing, Brown's linear exponential smoothing, Brown's quadratic exponential smoothing, automatic AEP, Bayesian, and two methods using combined forecasts from these methods. For details on these methods see Makridakis et al. (1982). For each horizon, we ranked the forecasting methods for accuracy according to each of six error measures. We repeated this procedure for six-year-ahead forecast errors. Then we did similar analyses for one quarter-ahead and eight-quarter-ahead forecasts for the five subsamples of quarterly series using 13 methods (the original 11 methods plus the seasonally adjusted random walk and HoltWinters).

Exhibit 1
Reliability of the error measures
(Average Spearman correlations for pairwise comparisons among five subsamples.)

Error	Quarterly		Annual		Average
Measure	1-ahead	8-ahead	1-ahead	6-ahead	
RMSE	0.14	–0.13	0.26	0.54	0.20
MdAPE	0.14	0.22	0.46	0.79	0.40
MAPE	0.59	0.61	0.49	0.30	0.50
GMRAE	0.38	0.17	0.81	0.74	0.53
MdRAE	0.39	0.43	0.79	0.72	0.58
Percent Better	0.60	0.59	0.82	0.78	0.70

To what extent did rankings based on a given error measure hold up across the five subsamples? We analyzed this question by calculating Spearman rank-order correlation coefficients, r_s, for the accuracy rankings between each pairing of the subsamples. We then averaged these ten pairwise correlations. Exhibit 1 summarizes results for each of the error measures by period and horizon.

Rankings of methods based on the RMSE were highly unreliable. For example, the average r_s for 8-ahead quarterly forecasts was –0.13. Across the four tests, the average pairwise correlation for the RMSE was only 0.20. Given such a low level of reliability, the use of the

RMSE would require many series. Prior research using this criterion would be suspect except where the comparisons involved many series.

The average r_s for the MAPE was 0.50 over the four tests. Surprising to us, the MdAPE was not more reliable than the MAPE; its average r_s was 0.40.

As expected, the relative error measures were the most reliable. The GMRAE's reliability was 0.53, while those for the MdRAE and Percent Better were 0.58 and 0.7O, respectively. To illustrate the differences in reliability, we calculated what sample sizes would yield the same statistical significance as provided by the Percent Better for 18 annual one-ahead forecasts.[2] The necessary sample sizes are: 18 series using GMRAE, 19 using MdRAE, 49 using MAPE, 55 using MdAPE, and 170 using RMSE.

Exhibit 2
Agreement among accuracy rankings for a set of 18 annual series
(Spearman correlations for one-year horizon.)

Error measure	MAPE	MdAPE	Percent Better*	GMRAE	MdRAE
RMSE	0.44	0.42	0.11	0.03	–0.31
MAPE		0.83	0.17	0.68	0.28
MdAPE			0.09	0.40	0.06
Percent Better*				0.46	0.65
GMRAE					0.79

*To keep the sign consistent, we used 'Percent Worse' rather than 'Percent Better' for the correlation.

3.3. Construct validity

Reliability examines whether the same results are produced, but it does not demonstrate that the proper thing is being measured. Construct validity asks whether a measure does, in fact, measure what it purports to measure. We were interested in the extent to which the various measures assess the 'accuracy'of extrapolation methods. To examine this, we compared rankings of the forecasting methods by each of the error measures. If they are measuring the same thing, and doing so reliably, the intercorrelations among the various measures should be high.

We first examined rankings of one-ahead forecasts for a set of 18 annual series. These results, given in Exhibit 2, show substantial agreement among some rankings of the methods, such as the 0.68 correlation between MAPE and GMRAE. But the rankings by some error measures were not highly correlated with the rankings by other measures, such as the 0.03 between the RMSE and the GMRAE. Thus, the choice of an error measure can affect conclusions about the relative accuracy of forecasting methods. As an extreme example, the

[2] To compute the equivalent sample sizes, we first obtained the significance of the reliability for the percent better measure for the 18 annual 1-ahead forecasts, where r_s was 0.82. Z = r_s times the square root of (N- 1); thus, Z = 3.38. We then asked what sample size would be required to achieve this Z-score given the estimated r_s for each of the error measures in the annual 1-ahead column of Exhibit 1.

rankings based on RMSE were negatively correlated with those from the MdRAE. Overall, the average Spearman correlation between pairs of measures was 0.34, indicating only modest construct validity. The three relative error measures, however, produced rankings that were similar to one another. The average Spearman rho among these rankings was 0.63 (see the three correlations in the bottom right-hand corner of Exhibit 2).

If the different measures assess the same thing, then their intercorrelations should increase as the reliability of the measures is increased. Conversely, if they were measuring independent constructs, an increase in reliability should not increase their intercorrelations. We examined whether the six error measures converged with one another as the number of series was increased from 18 to 90. They did. Compare Exhibit 2 with Exhibit 3. Twelve of the correlations increased, one did not change, and only two decreased. The average Spearman rho among the six measures for the 18 series in Exhibit 2 was 0.34, while that for all 90 series was 0.68. This increase supports the belief that the different error measures are measuring the same construct.

As a further test of construct validity, we examined whether the accuracy rankings of methods obtained by using one error measure were similar to the rankings obtained by other error measures. First, we constructed a consensus ranking by averaging the rankings from each of the six error scores for the full sample of 90 annual series and for the 101 quarterly series. Then, each error measure's ranking of forecasting methods was correlated with the consensus ranking. Exhibit 4 summarizes the results. Most of the measures were strongly correlated with the consensus. These results suggest that the measures each provide reasonable measures of accuracy when the number of series is fairly large. In particular, the relative errors and absolute percentage errors provided similar accuracy rankings

Exhibit 3
Agreement among accuracy rankings for 90 annual series
(Spearman correlations for one-year horizon.)

Error measure	MAPE	MdAPE	Percent Better*	GMRAE	MdRAE
RMSE	0.51	0.82	0.29	0.79	0.79
MAPE		0.73	0.35	0.79	0.70
MdAPE			0.47	0.97	0.96
Percent Better*				0.46	0.58
GMRAE					0.95

*To keep the sign consistent, we used ‘Percent Worse’ rather than ‘Percent Better’ for the correlation

Exhibit 4
Correlation with the consensus
(Spearman correlations for the full samples.)

Error Measure	Quarterly		Annual		Average
	1-ahead	8-ahead	1-ahead	6-ahead	
RMSE	0.65	0.17	0.80	0.86	0.62
Percent Better	0.81	0.79	0.64	0.51	0.69
MdRAE	0.69	0.92	0.96	0.96	0.88
MdAPE	0.78	0.88	0.96	0.93	0.89
GMRAE	0.80	0.85	0.93	0.97	0.89
MAPE	0.83	0.87	0.80	0.94	0.90

The RMSE had the lowest correlation to the consensus (r = 0.62), probably due to the RMSE's poor reliability. The low correlation also may be due to the RMSE's emphasis on assessing large errors; it is the only measure here that uses a squared error term.

Although the Percent Better is a reliable measure, it was not highly correlated with the consensus (r_s = 0.69). This low correlation is probably because the Percent Better is the only measure that ignores the magnitude of the errors.

3.4 Other criteria

Other criteria should also be considered when selecting an error measure. These criteria include expense, understandability, sensitivity, and relationship to decision making. The first two are easily dealt with; while medians are somewhat more expensive to calculate than means, none of the measures is expensive given current computer capabilities. All of the methods considered here are relatively easy to understand; however, squared error terms may be more difficult for some people to understand.

Sensitivity: For calibration, it is desirable to have a sensitive error measure so as to reveal the effects of changes. The measure should indicate the effect on accuracy when a change is made in a parameter for a given model.

Median error measures are not sensitive. In developing rules for rule-based forecasting [Collopy and Armstrong (1992)], we found medians to be of little value because of their low sensitivity. Percent Better is not sensitive because once a method is more accurate than the random walk for a given series, further improvements in forecasting that series produce no change when summarizing across series. Similarly, the Percent Better gives no credit for reducing the error from an extremely poor forecast to the point where it is almost as accurate as that for the random walk.

Relationship to Decision Making: None of the error measures is ideal for aiding decision making. However, the RMSE describes the magnitude of the error in terms that would be relatively more useful to decision makers. Consistent with this, Carbone and Armstrong (1982) found that

practitioners preferred the RMSE to all other error measures. The MAPE and MdAPE are less appealing because percentages do not have obvious implications for decision making. For example, in inventory control, losses would be related to dollars of additional inventory or to opportunity costs from lost sales, not to percentage errors. Relative measures have the least relationship to decision making. The Percent Better does not recognize the amount of improvement at all. The GMRAE, as noted above, gives as much credit for an improvement in a relatively constant series as for a corresponding percentage improvement in a series that changes substantially.

4. Selection of error measures given few series

We examined two procedures for comparing methods when the number of series is small. One is to use a consensus based on a variety of error measures. The other procedure is to use errors over the forecasting horizon, which we refer to as the cumulative-horizon error.

4.1. Consensus error measures

Forecasts are used in different ways and for different decisions. As a result, multiple error measures might be relevant. However, the use of multiple measures makes the comparisons more difficult. A consensus would simplify comparisons. Also because each error measure has defects, a consensus based on multiple measures might compensate for these defects.

To assess the value of a consensus error measure, we first prepared a consensus ranking for the forecasting methods in each subsample by averaging the ranks. We posed this question: If the analyst used only the GMRAE for a sample of series, would the conclusions differ from those based on a consensus from the same subsample? Therefore, we compared the consensus rankings of forecasting methods with those provided by the GMRAE. This test required some criterion; for this, we asked which measure for this sample would provide the most appropriate conclusion about the relative accuracy of the methods for the full sample of 90 series. We used the MdAPE to rank methods for the full sample of series; that is, this ranking was treated as 'truth' for purposes of this test. Exhibit 5 presents the results in the first two columns. For annual one-ahead forecasts, GMRAE rankings for subsamples of 18 series were highly correlated on average (r_s = 0.88) with the rankings when all 90 series ('Full Sample') were ranked using the MdAPE. This is similar to the correlation obtained when we used the consensus to rank the accuracy of the methods in the subsample (r_s = 0.81). The consensus offers modest improvements for the quarterly data (0.56 versus 0.41), but none for the annual (0.81 versus 0.85). This result suggests that the consensus would be more useful where the reliability is lower. (Exhibit 1 shows that reliability was lower for quarterly than for annual data for 10 of the 12 comparisons.)

Exhibit 5
Comparison of consensus and GMRAE for calibration using a small number of series
(Average Spearman correlations.)

Situation	Full sample MdAPE vs. Subsample		Full sample consensus vs. Subsample	
	GMRAE	Consensus	GMRAE	Consensus
Quarterly 1-ahead	0.46	0.57	0.59	0.69
Quarterly 8-ahead	0.35	0.56	0.41	0.65
Quarterly average	0.41	0.56	0.50	0.67
Annual 1-ahead	0.88	0.81	0.89	0.83
Annual 6-ahead	0.81	0.81	0.85	0.88
Annual average	0.85	0.81	0.87	0.56

We repeated the analysis with 'truth' based on the consensus rankings for the 90 series as the criterion. As shown in the last two columns of Exhibit 5, the results were similar to those using the MdAPE as the criterion. The consensus again offered modest gains for the quarterly data (0.67 versus 0.50) and no gains for the annual data (0.86 versus 0.87).

4.2. Cumulative-horizon error

Instead of basing comparisons of methods on a single forecast horizon, such as a one-ahead or a six-ahead annual forecast, one might summarize across forecast horizons (e.g., the error over the next six annual forecasts). One advantage of the cumulative-horizon error is simplicity. Using this single measure for calibration would be preferable to examining the error measure for each forecast horizon.

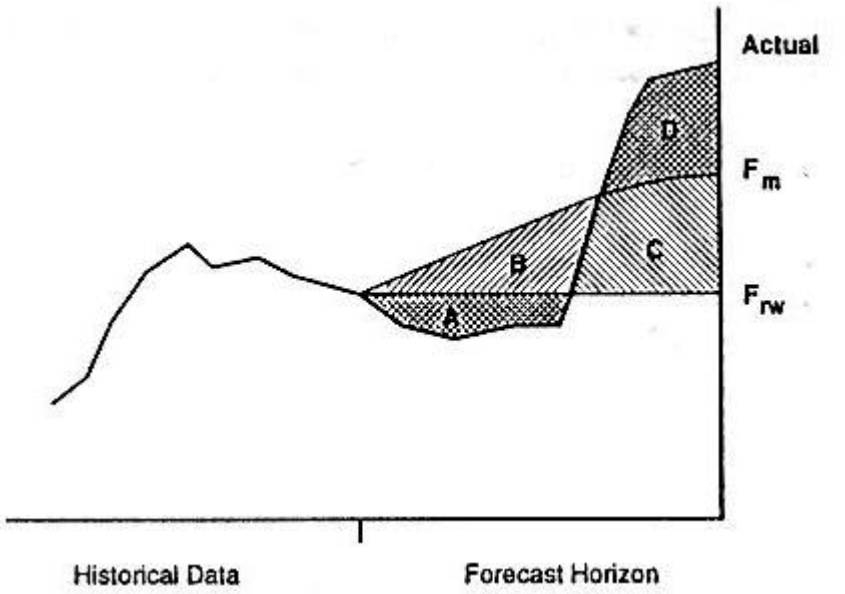
To examine the impact of cumulating errors across horizon, we focused on the RAE. As we have said, the RAE is like Theil's U2 for single horizon forecasts. However, the RAE differs from Theil's U2 when errors are cumulated over the forecast horizon (e.g., for annual forecasts covering years one through six.) To obtain Theil's U2 over the forecast horizon, one calculates a geometric mean of the errors for each horizon for a given series. In contrast, the Cumulative RAE takes the arithmetic sum of the absolute error for the proposed method over the forecast horizon and divides it by the corresponding error for the random walk. It is calculated for a series as follows:

$$CumRAE_m = \frac{\sum_{h=1}^{H}(|F_{m,h} - A_h|)}{\sum_{h=1}^{H}(|F_{rw,h} - A_h|)}$$

where F is the forecast value, *m* designates the forecasting method being evaluated, *rw* designates the random walk, *h* represents the forecast horizon, *A* is the actual value, and *H* is the

number of periods in the forecast horizon. To summarize across series, the geometric mean (designated as GMCumRAE) or the median (MdCumRAE) can be used.

Exhibit 6



$$\text{CumRAE}_m = \frac{A+D+B}{A+D+C} = \frac{\text{Absolute Deviation between Actual and Forecast from Model M } (F_m)}{\text{Absolute Deviation between Actual and Forecast from Random Walk F } (F_{rw})}$$

Exhibit 6 illustrates the Cumulative RAE. The error for the model to be evaluated (Model *m*) is the area between the forecast for that model (F_m) and the actual outcome; these are the areas A + D + B. This is divided by the deviations between the forecast from the random walk (F_{rw}) and the actual outcome; these are the areas A + D + C.

The CumRAE should be less influenced by outliers than the RAE. Given a larger number of forecasts, it is less likely that the random walk error will be near zero (making the denominator in the RAE very small). It is also less likely that the proposed model will be almost perfect (making the numerator very small). To test this, we calculated RAEs from all of the methods for all of the series, quarterly and annual. We defined outliers as values lower than 0.01 or greater than 10. Exhibit 7 summarizes the percentages of outliers for the short-range, long-range, and cumulative-horizon RAEs. The single horizon forecasts had some outliers, but outliers were less common for the cumulative horizons than for single horizons. For the annual data, only 0.3% of the CumRAEs were outliers, while for the quarterly data only 0.7% were outliers.

Exhibit 7
Frequency of RAE outliers (percentages)

	Less than 0.01	Greater than 10	Totals
Annual			
1-ahead	0.4	1.7	2.1
6-ahead	0.4	2.0	2.4
Cumulative	0.0	0.3	0.3
Quarterly			
1-ahead	0.8	3.3	4.1
8-ahead	0.4	3.1	3.5
Cumulative	0.1	0.6	0.7

To assess whether the protection against outliers and the gain in reliability could be achieved without a substantial loss in validity, we analyzed the effect of generalizing from cumulative-horizon forecast errors, rather than from single-horizon errors. We did this for quarterly and annual data and for one-ahead and in-ahead forecasts. Exhibit 8 summarizes the results. To our surprise, the GMCumRAE did not perform as well as did the single horizon one-ahead GMRAE; the GM-CumRAE rankings for the subsamples were not as highly correlated to the full sample one-ahead rankings as were the one-ahead rankings for the subsamples (0.51 versus 0.77). Thus, there appears to be a loss in validity for the one-ahead forecasts. The GMCumRAE might, however, be more useful for smaller samples of series than those we examined.

Exhibit 8
Stability of cumulative horizon RAEs vs. horizon RAEs
(Average Spearman correlations.)

Time interval	Full sample 1-ahead vs. Subsample		Full sample *h*-ahead vs. Subsample	
	1-ahead	Cumulative	*h*-ahead	Cumulative
Quarterly	0.64	0.33	0.44	0.47
Annual	0.89	0.69	0.87	0.91
Average	0.77	0.51	0.65	0.69

We expected the CumRAE and Theil's U2 to give comparable results for the cumulative-horizon error. To examine this, we used these two error measures to rank the accuracy of the 11 methods used for the 90 annual series. The results were indeed similar, as the Spearman rank order correlation was 0.996. We used the same procedure to compare the accuracy rankings for the 13 methods used for the 101 quarterly series; the Spearman correlation was also very high, at 0.991.

5. Guidelines for selecting error measures

To aid in the selection of an error measure, we rated each error measure as good, fair, or poor for each criterion. We based the ratings for reliability and construct validity on the empirical results. The other ratings represent our subjective judgments. None of the error measures that we examined was superior on all criteria. Exhibit 9 shows the ratings.

Exhibit 9
Ratings of the error measures

Error measure	Reliability	Construct validity	Outlier protection	Sensitivity	Relationship to decisions
RMSE	Poor	Fair	Poor	Good	Good
Percent Better	Good	Fair	Good	Poor	Poor
MAPE	Fair	Good	Poor	Good	Fair
MdAPE	Fair	Good	Good	Poor	Fair
GMRAE	Fair	Good	Fair	Good	Poor
MdRAE	Fair	Good	Good	Poor	Poor

Calibration requires a sensitive error measure. When a change is made in a model, it should be easy to see how this affects its performance. Good sensitivity is provided by only three of these measures: RMSE, MAPE, and GMRAE. We recommend the GMRAE because the RMSE has poor reliability, and because the MAPE is biased in favor of low forecasts.

For selection among forecasting methods, the primary criteria are reliability, construct validity, protection against outliers, and the relationship to decision making. Sensitivity is not so important for selecting methods. When only very small sets of series are available, the MdRAE is appropriate; it is as reliable and as valid as the GMRAE, and it offers better protection against outliers. Given a moderate number of series, reliability becomes a less important issue. The MdAPE would be appropriate because of its closer relationship to decision making.

The RMSE is unreliable. Related to this is its poor protection against outliers. We do not recommend the RMSE for assessing the level of accuracy. As noted, it was not useful for the 1001 series in the M-competition [Chatfield (1988)].

6. Limitations

Our analysis applies only to the choice of errors for generalizing from comparisons across multiple time series. The conclusions do not necessarily apply to the examination of a single time series. In that case, the selection of an error measure should relate more closely to any decisions that may be based on the forecast.

Because we examined only ratio-scaled economic and demographic data, we do not know whether our conclusions apply to other types of data. Also, our study used only extrapolation methods. Other measures may be desirable for judgmental and econometric methods.

While our study examined many commonly used errors measures, it ignored others. The adjusted MAPE (or MAPE-A) divides the absolute error by the average of the actual and predicted; this compensates for the MAPE's favorable treatment of forecasts that are too low. The R2, widely used because of its availability in statistical packages, provides information about covariation. The mean absolute deviation (MAD) is extensively used in inventory control as it is closely related to decision making. Turning point errors are of interest to economists. These measures are discussed in Armstrong (1985, pp. 346-356). They do not exhaust the possibilities. For example, the Root Median Square Error (RMdSE) would avoid the outlier problem and might provide a reliable measure. Another candidate is the root mean square percentage error (RMSPE).

Our study ignores large errors, which are sometimes the primary concern. For example large errors have disproportionate impacts for forecasts involving weather (droughts, floods, or hurricanes), electrical power (shortages), and cash flow (if low for a financially weak company). Other error measures might be appropriate in these situations. For example, the RMSE might be appropriate here if scaling is not a problem.

The selection of an error measure is dependent upon the situation. None of the error measures was superior on all criteria. To calibrate the parameters of a given model, we recommend the GMRAE. To select among forecasting methods, we recommend the MdRAE when using a small number of time series, and the MdAPE when many series are available.

The RMSE has been widely used for comparing forecasting methods. Our study suggests that the RMSE should not be used for generalizing about the level of accuracy of alternative forecasting methods because of its low reliability. The MAPE should not be used if large errors are expected because it is biased in favor of low forecasts. Consensus errors offered slight advantages where reliability was suspect. Cumulative horizon errors were examined, but they produced no benefits.

References

Armstrong, J. Scott, 1985, *Long-Range Forecasting* (Wiley, New York).

Bliemel, Friedhelm W., 1973, 'Their's forecast accuracy coefficient: A clarification,' *Journal of Marketing Research*, 10, 444-446.

Carbone, Robert and J. S. Armstrong, 1982, 'Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners,' *Journal of Forecasting*, 1, 215-217.

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

Collopy, Fred and J. S. Armstrong, 1992, 'Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations,' *Management Science* (forthcoming).

Makridakis, Spyros et al., 1982, 'The accuracy of extrapolation (time series) methods: Results of a forecasting competition,' *Journal of Forecasting*, 1, 111-153.

Thei, Henri, 1966, *Applied Economic Forecasting* (Rand-McNally, Chicago).

Zellner, Arnold, 1986, "A tale of forecasting 1001 series. The Bayesian knight strikes again," *International Journal of Forecasting*, 2, 491-494.

Appendix

The following notation is used for the definitions of error measures that follow:

m	is the forecasting method,
rw	is the random walk method,
h	is the horizon being forecast,
s	is the series being forecast,
$F_{m,h,s}$	is the forecast from method m for horizon h of series s,
$A_{h,s}$	is the actual value at horizon h of series s,
H	is the number of horizons to be forecast, and
S	is the number of series being summarized.

The absolute percentage error (APE) for a particular forecasting method for a given horizon of a particular series is defined as

$$APE_{m,h,s} = \left| \frac{F_{m,h,s} - A_{h,s}}{A_{h,s}} \right|$$

The APEs for a particular forecasting method are summarized across series by

$$? \ ?P?_{m,h} = \frac{\sum_{s-1}^{S} ?PE_{m,h,s}}{S} x\,100 \quad \text{or by}$$

$MdAPE_{m,h} =$ Observation $\frac{S+1}{2}$ if S is odd, or

the mean of observations $\frac{S}{2}$ and

$\frac{S}{2}$ + 1 if S is even, where the observations are rank-ordered by $APE_{m,h,s.}$

The relative absolute error (RAE) for a particular forecasting method for a given horizon of a particular series is defined as

$$RAE_{m,h,s} = \frac{\left|F_{m,h,s} - A_{h,s}\right|}{\left|F_{rw,h,s} - A_{h,s}\right|}.$$

The Winsorized RAEs are defined by

$$WRAE_{m,h,s} = \begin{cases} 0.01 if RAE_{m,h,s} < 0.01 \\ RAE_{m,h,s} \; if \; 0.01 \le RAE_{m,h,s} \le 10 \\ 10 \; if RAE_{m,h,s} > 10 \end{cases}$$

Because we always recommend Winsorizing of the RAE, we drop the *W* below and in the text. The Winsorized RAEs for a particular forecasting method are summarized across series by

$$GMRAE_{m,h} = \left[\prod_{s-1}^{S} RAE_{m,h,s}\right]^{1/S} \text{ or by}$$

$MdRAE_{m,h}$ = Observation $\frac{S+1}{2}$ if S is odd, or

the mean of observations $\frac{S}{2}$ and

$\frac{S}{2}$ + 1 if S is even, where the observations are rank-ordered by $RAE_{m,h,s}$

The RAEs for a particular forecasting method are summarized across all of the *H* horizons on a particular series by

$$CumRAE_{m,s} = \frac{\sum_{h=1}^{H}\left|F_{m,h,s} - A_{h,s}\right|}{\sum_{h=1}^{H}\left|F_{rw,h,s} - A_{h,s}\right|}$$

The CumRAE is Winsorized in the same way as the GMRAE is Winsorized. The CumRAEs for a particular forecasting method are summarized across series by

$$GMCumRAE_m = \left[\prod_{s=1}^{S} CumRAE_{m,s}\right]^{1/S} \text{ or by}$$

$MdCumRAE_m$ = Observation $\frac{S+1}{2}$ if S is odd, or

the mean of observations $\frac{S}{2}$ and

$\frac{S}{2}$ + 1 if S is even, where the observations are rank-ordered by Winsorized

$CumRAE_{m,s}$

The root mean squared errors (RMSEs) for a particular forecasting method are summarized across series by

$$RMSE_{m,h} = \left(\frac{\sum_{s=1}^{S}\left(F_{m,h,s} - A_{h,s}\right)^2}{S}\right)^{1/2}.$$

Theil's US for a particular forecasting method on a particular series is

$$U2_{m,h,s} = \frac{\left[\frac{1}{h}\sum_{h=1}^{H}(F_{m,h,s} - A_{h,s})^2\right]^{1/2}}{\left[\frac{1}{h}\sum_{h=1}^{H}(F_{rw,h,s} - A_{h,s})^2\right]^{1/2}}.$$

To summarize across series, a geometric mean is calculated by

$$GMU2_{m,h} = \left[\prod_{s=1}^{S} U2_{m,h,s}\right]^{1/S}.$$

Percent Better is calculated as

$$\text{Percent Better}_{m,h} = \frac{\sum_{s=1}^{S} j_s}{S} x100$$

where $j_s = \begin{matrix} 1 \text{ if } \left|F_{m,h,s} - A_{h,s}\right| < \left|F_{rw,h,s} - A_{h,s}\right| \\ 0 \text{ otherwise.} \end{matrix}$

Consensus Rank for the six measures we examined was calculated by

Consensus Rank = $\left[\sum_{i=1}^{6} R_{i,m}\right] / 6,$

Where $R_{i,m}$ is the ranking given by measure i to method m.

For simplicity, we generally eliminate the subscripts in the text.