Published in *International Journal of Forecasting* (1992), 8, 99-111

A Commentary on Error Measures

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

Error measures and the choice o f a forecast method, Dennis A. Ahlburg,* Industrial Relations Center, Carlson School of Management, University of Minnesota.

Is the choice of an error measure to identify the most accurate forecasting method a question of personal taste? It appears that this may be the case although the papers by Armstrong and Collopy and by Fildes argue that it should not be. Carbone and Armstrong (1982) found that Root Mean Square Error (RMSE) was the most preferred measure of forecast accuracy. This is despite the fact that it is widely accepted that unit-free measures are necessary for comparisons among forecast methods and RMSE is not unit-free. Mean Absolute Percentage Error (MAPS) was the most widely used unit-free accuracy measure but is relevant only for ratio-scaled data (as Armstrong and Collopy point out, this is not a problem with most economic and demographic data) and is consistent with a loss function linear in percentage, not absolute errors. This may not be appropriate in some forecasting applications. As a rough check on the relevance of these findings to my own area, population forecasting, I surveyed seventeen papers dealing with population forecasts (national, state, county, and city). See the papers referenced in Ahlburg (1987) and Land (1986) and the papers by Ahlburg (1982), Smith (1987), and Smith and Sincich (1991). I found ten used MAPE, four used RMSE, three used Root Mean Square Percentage Error (RMSPE), and three used Theil's U (four used several measures). None justified the choice of error measure. Given this state of affairs, Armstrong and Collopy and Fildes set out to assist the forecaster in the task of choosing the most appropriate measure of forecast accuracy to select the best forecasting method.

Where the forecaster knows the loss function applicable to the problem at hand, the choice of error measure is somewhat less difficult. For instance, if the loss function is linear then the search for an accuracy measure can be confined to linear measures. However where this is not the case and where the forecasting problem involves many series the forecaster needs guidance. Armstrong and Collopy attempt to do so empirically (based on reliability, validity, outliers, sensitivity, and interpretability). While they are able to provide guidelines in this search they show there is no single answer. Selection of an error measure, and thus an extrapolation method, is dependent on use and upon the situation. Their results, and those of Fildes on reliability, outliers, and validity, are disturbing, especially where the number of series is small. For the forecaster of a small number of series, as is often the case in demography where one may be forecasting population by forecasting births, deaths, and migration or be forecasting fertility by 5 year age groups, the choice of an error measure and forecasting method is something of a craps shoot. In the small series case Armstrong and Collopy suggest using the Median Relative Absolute Error (MdRAE) but, as they point out, this measure has significant problems in interpretability for managerial decision making. Perhaps here the Relative Geometric Mean Square Error (RGRMSE) that performed well in Fildes' tests and has a clear interpretation might be useful. The small number of series case is the most problematic and I think needs more attention.

Armstrong and Collopy place much weight on validity or, as measured by them, consensus. That is, do the error measures measure the same thing. This criterion weights the tables against measures that have a different focus, such as percentage error measures that penalize large errors more than smaller ones, and percentage better that does not take account of error size. I would place less weight on consensus.

A particularly sobering finding is the poor performance of the RMSE, the most commonly used error measure. Its use is related to its popularity with statisticians and its interpretability in relation to business decisions and not to its efficiency in choosing accurate forecasting methods. The reliability of RMSE is poor and it is scale dependent. Another favorite, the MAPS, does reasonably well in the Armstrong-Collopy comparisons (better, I think, than they acknowledge) but is strongly rejected by Fildes on statistical grounds.

* The author is grateful to Scott Armstrong, Fred Collopy, and Robert Fildes for comments.

While it is difficult to find a single measure of accuracy, all is not lost because both papers find support for the usefulness of modifications of these well known methods. The RGRMSE overcomes the problems of the RMSE and the Relative Median Absolute Percentage Error (RMdAPE) those of the MADE. Both papers find RMAPE useful although Fildes shows a preference for the RGRMSE based on interpretability. These measures are related to Theil's U2 but use levels of variables not changes as in Theil's original measure. I think a move to relative accuracy measures is overdue and the random walk is a good choice for the comparator since it is often in the minds of forecasters. Other comparators may be relevant in other applications, for example, Keilman (1991) used a modified U2 for evaluating population forecasts with the population in the year before the jump-off year as the alternative forecast.

Interpretability seems to be an important issue in the use of accuracy measures. Armstrong and Collopy like the RAE because of its ease of interpretation relative to Theil's U2 and Fildes likes the interpretability of RGRMSE. I find Theil's U2 more appealing on interpretability than Armstrong and Collopy (see Ahlburg 1982) but agree that our ability to communicate the meaning of accuracy statistics is a key to spreading the use of better statistics among practitioners.

The conclusion reached by Armstrong and Collopy that no single accuracy measure is appropriate in most situations is true but perhaps too pessimistic. These papers clearly reject some frequently used measures and provide others that have desirable statistical properties and perform well in applications. The next step is to get forecasters to use '[appropriate] accuracy results to guide our choices among methods' [Beaumont and Isserman (1987: 1005)].

References

Ahlburg, D.A., (1982), "How accurate are the US Bureau of the Census projections of total live births?" *Journal of Forecasting*, 1, 365-374.

Ahlburg, D.A., (1987), "Population forecasting," in: S. Makridakis and S. Wheelwright, eds., *The Handbook of Forecasting: A Manager's Guide*, second ed. (Wiley, New York) 135-149.

Beaumont, P. and A. Isserman, (1987), "Comment," *Journal of the American Statistical Association*, 82, 1004-1009.

Keilman, N.W., (1991), *Uncertainty in National Population Forecasting: Issues, Backgrounds, Analyses, Recommendations*. Swets and Zeitlinger, Amsterdam.

Land, K.,(1986), "Methods for national population forecasts: A review," *Journal of the American Statistical Association*, 81, 888-901.

Smith, S. (1987), "Test of forecast accuracy and bias for county population projections," *Journal of the American Statistical Association*, 82, 991-1003.

Smith, S. and T. Sincich, (1991), "An empirical analysis of the effect of length of forecast horizon on population forecast errors," *Demography*, 28, 261-274.

A commentary on error measures, Chris Chatfield, School of Mathematical Sciences, University of Bath, Bath, Avon, BA2 7AY, UK.

These two papers address important issues in the evaluation of forecasting methods. I welcome the main thrust of both papers and would like to briefly summarize the results for the benefit of the general reader. I am just a little worried that the "message" may get rather lost in these longish papers with so many tables to digest.

I suppose the average forecaster (is there such an animal?) wants to know (a) which forecasting method to use, and (b) which error measure should be used to assess forecasting accuracy. The answers to these questions are of course situation-dependent. For example the choice of method depends on the type of data and the expertise available. Here I concentrate on the following dichotomy of situations. At one extreme there is a single series to forecast, an appropriate probability model is fitted and optimal forecasts are then made. At the other extreme, the practitioner with a large number of series to forecast, may decide to use the same all-purpose procedure whatever the individual series look like. The former situation is perhaps more familiar to the statistician, and the latter to the operational researcher.

Error measures

For a single series it is perfectly reasonable to fit a model by *least squares* (as statisticians customarily do) and evaluate forecasts from different models and/or methods by the mean square error (MSE) of the forecasts. However, once you apply the same method to a group of series, it can be disastrous to average raw MSE across series as MSE is scale-dependent. Unfortunately this was done with the M-competition results so that the MSE results therefrom should be disregarded. Various alternatives are explored by Armstrong and Collopy and detailed recommendations are made. They seem reasonable enough but involve three different measures and so are perhaps somewhat overly complicated. Fildes on the other hand plumps for the geometric root MSE (GRMSE). This may be unfamiliar to many readers, in which case I recommend reading Section 1.2 closely. The key point is that the measure should be *scale-independent so* that multiplying all the numbers in one series by the same constant (e.g. expressing sales in pounds rather than dollars) should have no effect on the overall comparison of methods for a group of series. It is also valuable if the error measure has the property of not being unduly affected by outliers, as demonstrated by Fildes for GRMSE.

Another important point demonstrated by Fildes is that forecasting comparisons should not rely on a single time origin for each series but that the error measure should be averaged *across time* in some way, as well as across different series.

Choice of forecasting method

As regards choice of forecasting method, the results of the M-competition and similar studies have long been controversial. Is it really true that simple smoothing methods can give forecasts which are as good as Box-Jenkins forecasts for example? Do the results generalize to other groups of series and/or single series?

The value, or otherwise, of forecasting competitions has been discussed by various authors [e.g. Chatfield (1988a, section 4.1)]. They do tell us something, but only part of the story. They have been mainly helpful in comparing automatic forecasting methods for large disparate groups of series. The series in the M-competition were about as varied as one can get! In contrast Fildes examines a large group of series from the *same* company for the *same* variable. Naturally the data are much more *homogeneous* than the series in the M-competition. In particular they all show a negative trend for example which renders simple exponential smoothing inappropriate at, a stroke. Fildes uses background knowledge, an exploratory examination of the data, plus a comparative evaluation of different methods to select a forecasting method which is designed to cope with the particular situation, including the presence of trend and of outliers. This case study makes it is clear that the results of one forecasting competition need not necessarily generalize to another. Furthermore the results from forecasting competitions certainly do not generalize to the single-series situation, where the analyst will still have the difficult task of identifying the model appropriate to the given data. Perhaps the greatest benefit of the M-competition has been, not the results as such, but the 'by-products' in making us think more clearly about such issues as error measures and replicability.

Tables and graphs

For my final comments, I switch to a completely different topic. One of my 'hobby-horses' is lamenting the generally poor standard of tables and graphs which the computer age seems to have done little to improve.

The general rules for presenting clear tables and graphs are 'well-known' [e.g. Chatfield (1988b, section 6.5)] yet are often disregarded. For example there should be a clear self-explanatory title, the units of measurement should be stated, axes should be labeled, the number of significant figures in tables should be carefully chosen, and so on. The overriding rule is that the exhibit should be clear, easy-to-understand and preferably self-explanatory without looking at the text.

While I have seen many worse examples than those given in these two papers (which have improved during the refereeing process); I have also seen better and I suspect that many readers will be bemused by some of them. For example in Fildes' Exhibit 2, the horizontal axis, time, is measured in months (I think) and the vertical axis is the "number of circuits" not "circuits." In Armstrong- Collopy's Exhibit 6, I can guess what RW is, I can look up RAE and Fm, yet still end up vaguely confused. Likewise Fildes' Exhibit 5 is likely to bring on a headache rather than a dawning of light. If the results at lead 12 are like those at lead 6, is it necessary to give them? If the geometric mean is better than the arithmetic mean, why give the latter? Ah well, I could give far worse examples from other recent published papers, including figures with no title, unlabelled axes and so on. I urge all readers, whether acting as author, referee or editor, to give this topic the attention it deserves. It really can "make" a paper if the tables and graphs are clear, but this can take much more effort than many people realize. It is not obvious for example how to present Fildes' Exhibit 5 in the "best" way. In my experience it can take several iterations to get tables and graphs to a publishable state, and even then further improvements are often possible which can sometimes be seen more easily by a fresh eye.

References

Chatfield, C., (1988a), "What is the best method of forecasting?" *Journal of Applied Statistics*, 15, 19-38.

Chatfield, C., (1988b), *Problem-Solving: A Statistician's Guide*. Chapman and Hall, London.

Comparing forecasts in finance, Stephen J. Taylor, Department of Accounting and Finance, Lancaster University, UK.

The two papers by Armstrong, Collopy and Fildes draw attention to the important problem of deciding how to combine forecasting results for a set of time series. Their contributions will I hope stimulate further work in this area. In these comments I reflect on how the issues raised by the authors could relate to forecasting market prices and I conclude by noting some lessons that can be learnt from the papers.

Armstrong and Collopy give results for economic and demographic data and Fildes studies telecommunications data with references to inventory and production control. One way to evaluate the methodological contributions of these two papers is to consider their implications for forecasting methodology in a different area. I will attempt to do this for the application area I know best, namely Finance. Comparisons of forecasting methods in Finance are frequently made for several series but relatively rarely is an effort made to combine results across series. The following examples are therefore hypothetical but all are based on important contemporary problems in Finance.

Consider three forecasting tasks involving the market prices of financial assets where the researcher could have access to a sample of time series from a large, homogenous population. The researcher might work for a bank or might be an academic; both scenarios are plausible and a distinction between applied and theoretical forecasting should not be sought.

First, we might have an interest in predicting the future values of prices and the data could be the weekly prices of several hundred U.S. stocks for several years. Second, we might want to predict the volatility of prices, defined to be the standard deviation of changes in the logarithms of prices. The data could include daily high, low and close prices and also volatility figures implied by option prices, for a selection of financial assets, e.g. stock indices or exchange rates for several countries. Third, it might be desired to compare forecasts of spot exchange rates from three sources: the spot market, the forward market and surveys of experts; the sample would contain figures for perhaps a dozen currencies. Note that only the first task is a pure extrapolative forecasting exercise. The other two tasks involve information from additional sources (the options market in one case, experts in the other) and this is commonplace in Finance where it is often of interest to identify the most relevant source of information.

Standard methodology for all three tasks would include working with the logarithms of prices. The first differences of log price changes are closer to stationarity than the first differences of prices. Taking logarithms immediately deals with the issue of scale dependence mentioned in both papers. Most researchers would then compare forecasting methods by using the root mean square error (RMSE) criterion. Taking logs and then using RMSE is essentially the same as working with root mean square percentage errors which is one of the measures mentioned but not studied by Armstrong and Collopy. My guess is that presented with many series most researchers would take a simple average across series of the relative RMSE, defined as the RMSE for a method divided by the RMSE for a random walk forecast.

A single cross-section would certainly not be used to compare methods. Instead time series would be split for the first two tasks, parameters estimated for the first period and then a set of forecasts evaluated for the second period. The entire sample would be the second period for the third task. Furthermore, forecast errors would not be combined across forecast horizons. Results would usually be given by forecast horizon for a selection of horizons. Outliers would arise in all three forecasting tasks, because there are occasional dramatic changes in market prices. Trimming or excluding outliers would be considered dubious because they are a real phenomenon having important practical consequences.

The choice of an error measure for a Finance problem does not usually receive much attention. Thus choosing between a squared error measure and an absolute error measure could be regarded as a neglected choice. Any serious claim to have found important conclusions from a forecasting study would usually be accompanied by an evaluation of the consequences for market agents. Investigating the economic consequences of forecasts in Finance requires consideration of subjects such as systematic risk (for the first task especially), hedging (for the second task) and optimal portfolios (inter alia for the third task), rather than a view on error measures and loss functions.

The joint evaluation of forecast results for several series in Finance often reduces at a practical level to the performance of a portfolio formed partially by combining the decisions from a set of forecasts.

So what might a researcher into a Finance problem learn from these papers? First, the robustness of results to the choice between an absolute error measure and a squared error measure is a relevant issue, especially considering the magnitude of some outliers. Second, it is unwise to apply the results from one forecasting competition to a different context. Third, there must always be a preference for studying as many series as possible and this should be accompanied by serious consideration of how results are to be combined across series.

A statistician in search of a population, Patrick A. Thompson, Department of Decision and Information Sciences, College of Business Administration, University of Florida, Gainesville, FL 32605, USA.

Newbold (1983) was wrong when he called it the "competition to end all competitions." Although it is true that no new major forecasting contests have appeared in the decade since the M-Competition was performed, a recent series of papers suggests another competition is going on right now. Rather than being concerned with which technique wins the game, however, this new contest asks "what is the best way to keep score?"

As one of the contestants in this new game, it is interesting to note that we all seem to be coming up with more or less the same answer: relative error. Armstrong and Collopy advocate using a ratio of mean absolute errors, Fildes suggests a ratio of geometric root mean squared errors, while I [Thompson (1990, 1992)] opt for the logarithm of a mean squared error ratio. Of course, the use of relative error statistics is not new, with Theil (1966), Newbold and Granger (1974), McLaughlin (1975) and Fildes and Makridakis (1988) suggesting earlier variants.

On an individual series, all of these relative error statistics should give similar rankings of the accuracy of a set of forecasting techniques. Armstrong and Collopy show that their (CumRAE) statistic strongly agrees with the ranking given by Theil's (U2) statistic, and since mine (LMR) is a 1-1 transformation of U2, CumRAE and LMR would also agree.

Over many series, we differ on how to compute the `usual accuracy' of a technique. Fildes takes logarithms of his geometric mean ratios, indicates that these appear to be normally distributed across series, then averages them. LMR usually displays a symmetric distribution across series, and is also averaged. Because their statistic has a skewed distribution over many series, Armstrong and Collopy use either the geometric mean or the median of the CumRAE.

Our largest disagreement, however, concerns the type of inferences we attempt to make. Newbold (1983) strongly criticized the authors of the M-Competition report for performing formal statistical inference, pointing out that it was difficult to `draw a random sample of all time series'. In Thompson (1990), I indicated that a forecasting competition is similar to a repeated measures experiment in which the experimental units (the series) are treated as blocking factors. If the series are not selected randomly, however, it would be a *fixed* block design, under which inference extends only to the series included in the contest. Because of this limitation, I stopped short of performing formal inference.

Both the Armstrong- Collopy and Fildes papers contain formal statistical inference. Fildes uses a sample selected from a large and homogeneous set of telephone company series, thus has a well-defined population to work with. His inferences wouldn't necessarily extend beyond this population; in particular, personal experience suggests his lognormal findings do not extend to the M-Competition data. Armstrong and Collopy make inferences based on samples of the M-Competition data. In effect, they have *defined* the M-Competition data as their population! While I have often thought the cause of (at least academic) forecasting would be served if we did adopt a large collection of series as our "target population," I have too many reservations about how the M-Competition series were selected to accept it as mine.

References

Fildes, R. and S. Makridakis, (1988), "Forecasting and loss functions," *International Journal of Forecasting,* 4, 545-550.

McLaughlin, R.L., (1975), "The real record of the economic forecasters," *Business Economics,* 10, 28-36.

Newbold, P., (1983), "The competition to end all competitions," *Journal of Forecasting,* 2, 276-279.

Newbold, P. and C.W.T. Granger, (1974), "Experience with forecasting univariate time series and the combination of forecasts," *Journal of the Royal Statistical Society,* Series A, 137, 131-165.

Theil, H., (1966), "Measuring the accuracy of point predictions," ch. 2 in *Applied Economic Forecasting* Rand-McNally, Chicago.

Thompson, P., (1990), "An MSE statistic for comparing forecast accuracy across series," *International Journal of Forecasting*, 6, 219-227.

Thompson, P., (1991), "Evaluation of the M-Competition forecasts via log mean squared error ratio," *International Journal of Forecasting*, 331-334.

On seeking a best performance measure or a best forecasting method,[*] Robert L. Winkler, Fuqua School of Business, Duke University, Durham, North Carolina 27706, USA and Allan H. Murphy, Departments of Atmospheric Sciences and Statistics, Oregon State University, Corvallis, Oregon 97331, USA.

A great deal of emphasis in forecasting has been placed on the search for a `best' forecasting method or a `true' model. The search involves the evaluation of different methods or models, and in this endeavor also the focus often seems to be on the question of which performance measure is `best' in specific types of situations. To the extent that seeking out "best" methods and measures has helped us to learn about characteristics of altemative methods and measures, it has been very beneficial. In forecasting applications, however, this mindset is, in our opinion, restrictive and counterproductive.

Our discussion will concentrate primarily on the evaluation of forecasting methods or models. If we are going to choose a method and use it to generate a forecast for a given event or variable, what we ultimately care about is the relationship between the forecast and the observation (the actual event or value of the variable). We obviously cannot say in advance exactly what the forecast or observation will be for the situation of interest: If we could, we could forecast perfectly. Instead, we typically have data in the form of forecasts and the corresponding observations for previous situations. Setting aside the thorny question of whether the previous situations are `similar' to the current situation (which includes, but is not limited to, the distinction between fitted errors and actual forecasting errors), how can the information from past situations be used to evaluate forecasting methods?

The information consists of pairs of forecasts and observations. If we can ignore time-related issues (e.g., more recent observations being of greater interest), the information can be represented fully in terms of the joint distribution of forecasts and observations [Murphy and Winkler (1987, 1992)]. We view this distribution as the basic unit of analysis and then proceed to look at various distributions and summary measures based on the joint distribution. For starters, the joint distribution can be factored into a marginal distribution and a set of conditional distributions in two different ways, and the resulting distributions relate to different characteristics of forecasting performance. For example, the conditional distributions of the observations given the forecasts reflect the calibration of the forecasts, while the conditional distributions of the forecasts given the observations reflect the ability of the forecasts to discriminate among different values for the observations.

Comparison of entire distributions for different forecasting methods is a complex endeavor. Looking at entire distributions is valuable, but it is just the starting point of our distributions-oriented approach. Summary measures of the distributions are more readily interpretable. Means, standard deviations, fractiles, and other summary measures of the marginal and conditional distributions can provide useful information about the characteristics of a forecasting method. The correlation between forecasts and observations is also of interest. Moreover, all of this information can be helpful in refining the method or suggesting alternate methods that might not have been considered.

Note that the distributions and summary measures we have been discussing relate to the joint, marginal, and conditional distributions involving the forecasts and observations. In contrast, most performance measures focus on the errors; the differences between forecasts and observations. Restricting consideration to errors implicitly assumes that the specific values of the forecast and observation are not important as long as their difference is known. Under this assumption, we can consider the distribution of errors and then move on to summary measures of that distribution. Common performance measures such as the mean square error fall into this category. Moreover, some measures (the mean square error in particular) sometimes can be decomposed into components that tell us something about individual characteristics of the forecasting method (e.g., bias, calibration, discrimination).

In evaluating an individual forecasting method or model, then, there are many measures of performance that tell us different things - about the method or model. Moreover, in a comparative evaluation of different methods, new classes of distributions and measures are of interest. If we want to compare two methods, we can use the above-discussed procedures to evaluate each method individually. That is, we can look at the joint distribution of forecasts and observations for the first method and the many distributions and summary measures that can be found from that joint distribution, and then do the same thing for the second method. In a full comparative evaluation,

[*] This work was supported in part by the National Science Foundation under Grant SES-9106440.

however, the fundamental distribution of interest is the trivariate distribution of the forecasts from the first method, the forecasts from the second method, and the observations. As in individual evaluation, we can generate a variety of distributions and summary measures from the trivariate distribution. This approach enables us to investigate relationships between the two methods in a more complete fashion. Of course, if the past data on the two methods pertain to different cases (i.e., the methods were not used for the same situations and therefore do not have a common set of observations), we are faced with an evaluation problem involving a fundamental distribution defined over four variables. Clearly, comparative evaluation under these conditions is even more complex (necessarily involving additional marginal and conditional distributions) than that faced in the "matched" comparative evaluation problem just discussed.

The output from a comparative evaluation includes relative measures that provide information about how a forecasting method performs relative to another method. We like to distinguish between overall measures of the performance of an individual method (such as the mean square error) and relative measures (such as the percentage improvement in mean square error from one method to another) by calling the former measures of accuracy and the latter measures of skill.

Our primary message is that no single measure can capture all of the nuances of a forecasting method or the relative merits of different forecasting methods. As Fildes notes, "A single general loss function can seldom capture the complexities of how forecasting models are used." Thus, the search for a single "best" measure is doomed to failure. There simply is no single, all-purpose measure, and we feel that an emphasis on using multiple measures is much more effective in terms of providing greater insight and guiding efforts to improve forecasting methods.

The emphasis on multiple measures does not simply mean multiple overall measures of accuracy such as mean square error, mean absolute error, and mean absolute percentage error. These traditional measures are trying to get at the same thing (accuracy) in slightly different ways. It is much more effective to look at a variety of measures, including overall measures such as mean square error but also measures from the marginal and conditional distributions, in order to understand better the strengths and weaknesses of a forecasting method in a given class of situations. Moreover, we feel that the focus in forecast evaluation should be shifted from a measure-oriented approach to a distributions-oriented approach in which distributions are primary and other measures are viewed as summarizing different aspects of the distributions.

Now we return to the issue identified in the first sentence of these comments, the emphasis on the search for a "best" forecasting method or a `true' mode. As we move away from the use of a single performance measure to the consideration of multiple measures that provide information about multiple characteristics, we recognize that different methods tend to have somewhat different strong points and weak points. Evidence suggests that no single method is able to dominate on all dimensions of performance, and any choice of a single method involves implicit if not explicit tradeoffs. An alternative is to abandon the search for the `best' method and use multiple methods. Valuable information is provided by differences among the forecasts from multiple methods, and if a single forecast is desired the multiple forecasts can be combined. Empirical studies as well as theoretical results suggest that combining forecasts leads to better performance in many (perhaps most) cases [e.g. Clemen (1989)], but the notion of seeking out a single `best' method or `true' model seems to hold away in the forecasting community. As these comments are being written, it is hard, for example, to imagine that there is a `true' model for forecasting the movements of the economy over the next six months, any more than there is a `true' probability that the French franc will rise against the U.S. dollar in the next six months, yet economic forecasters often seem to be searching for such a model.

In conclusion, we note that the papers by Fildes and Armstrong and Collopy carry on a long tradition in forecasting of focusing on overall measures of performance based on errors and seeking a "best" measure, often in an attempt to find a "best" method. Within that tradition, they have some useful things to say. In contrast, we espouse a view that emphasizes multiple measures based on the joint distribution of forecasts and observations and sees advantages of combining forecasts from multiple methods with different characteristics. This view, which might have been somewhat impractical not too long ago, is now feasible to implement given current computing capabilities. We feel it will lead to an improved understanding of the performance characteristics of forecasting methods and to improved forecasts.

References

Clemen, R.T., (1989), "Combining forecasts: A review and annotated bibliography," *International Journal of Forecasting*, 5, 559-583.

Murphy, A.H. and R.L. Winkler, (1987), "A general framework for forecast verification," *Monthly Weather Review*, 115, 1330-1338.

Murphy, A.H. and R.L. Winkler, (1992), "Diagnostic verification of probability forecasts," *International Journal of Forecasting*, 7, 435-455.

Generalization and communication issues in the use o f error measures: A reply, Fred Collopy, The Weatherhead School, Case-Western Reserve University, Cleveland, Ohio 44118, USA and J. Scott Armstrong, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

We agree with most of what the commentators say about Armstrong and Collopy (1992), hereafter referred to as "AC," and Fildes (1992), hereafter referred to as "F." Here, we address three issues where we do not agree entirely:

(1) Can the results from the M-competition be generalized?
(2) Is Theil's U2 easy to communicate?
(3) Would a richer set of measures lead to improvements in the selection and development of forecasting methods?

Our own answers to these questions are "yes," "no," and "probably not," respectively.

Generalizability

Thompson (1992) views the M-Competition data as a population of economic and demographic series beyond which one cannot generalize. Taylor (1992) echoes this concern; he states that it is unwise to apply the results of one competition to another. In contrast, we see generalizing as a primary function of research on forecasting. Researchers should make empirical comparisons of methods on actual data in an effort to generalize to similar data. If generalization cannot be done, there would be little reason for conducting this kind of research.

It is difficult to define the domain of all possible time series. However, one can select series that are representative of other series. This was the strategy used in the M-competition and in F. Furthermore, the characteristics of these series can be described [as we have done for annual M-competition series in Collopy and Armstrong (1992), using 18 features]. We believe that findings from studies on actual data can be generalized to other economic and. demographic series. In Armstrong and Collopy (1993), we showed that the conclusions about forecasting methods based on analyses of the M-competition data were very similar when we repeated the analyses on four other data sets. We would be happy to cooperate with any attempts to extend the A & C study of error measures to other data.

Comparisons with Theil's U2

Chatfield (1992) and Ahlburg (1992) favor the use of Theil's U2. Ahlberg believes that Theil's U2 is easy to understand; but then he has written a paper on it [Ahlburg (1984)]. We believe that Theil's U2 is a highly desirable measure, so we were surprised that its use is limited primarily to economics. (We examined citations to two of Theil's books that discuss this measure. Of the 185 citations in the *Social Science Citation Index* from 1981 to 1991, at least two-thirds of the citations were by academic economists.) The survey of 145 forecasting researchers and practitioners conducted by Carbone and Armstrong (1982) showed that only two percent of them selected Theil's U2 for comparisons across series. Our guess is that Theil's U2 is underused because it is difficult to communicate to forecasters and decision makers.

The RAE is easier to communicate than Theil's U2. The term "Relative Absolute Error" is descriptive, while the term `U2' is not. Also, the procedure is a bit simpler than that for Theil's U2 as it does not use squared terms. Like researchers and practitioners, we have had difficulty understanding and remembering Theil's U2. When we began our work on rule-based forecasting (Collopy and Armstrong, 1992), we needed a reliable and sensitive measure that would enable us to draw conclusions from small sets of series. To improve reliability we developed a measure, the RAE, to control for scale, outliers, and change over the forecast horizon. In searching the literature to learn whether the RAE has been used previously, we rediscovered Theil's U2, a measure that also provided the reliability that we were seeking. Ironically, we discovered Theil's U2 in one of the authors' previous works (Armstrong 1985)!

As we have shown, Theil's U2 and the RAE have similar benefits. We advocate that one of these measures be used when making comparisons among forecasting methods. The RAE has not been used previously and Theil's U2 has been underused for the comparison of forecasting methods.

Use o f a richer set of performance measures

Winkler and Murphy (1992) argue for a richer set of forecast performance measures when- they suggest examining distributions of forecasts and predictions. Would such additional information improve decisions by researchers and forecasters? This is an empirical question. Prior research suggests that using additional information can be a risky and costly strategy.

In our opinion, the primary purpose of statis tics is to effectively communicate a large body of information. One key to communication is simplification. Complex concepts and complex measures are sometimes ignored, even when relevant. Let us illustrate this with our work on rule-based forecasting (Collopy and Armstrong, 1992). To examine the effects of changes in the rules that we were using to weight forecasts from multiple methods, we made about 500 runs over a three-year period and produced millions of forecasts. We examined six error measures and thus produced millions of forecast errors, yielding several thousand summary statistics. Examining and comparing these statistics was a formidable task. We are not convinced that our thousands of decisions would have been improved had we replaced each of these statistics with a richer set of information. Clearly our decision task would have been more substantial.

We are probably not alone in our inability to make decisions based on many variables. For example, Dudycha and Naylor (1966) showed that adding information about a less important variable in a two-variable model decreased the subjects' ability to make good predictions. Somehow, then, information about thousands of comparisons must be reduced to simple and understandable metrics so that different researchers can agree about statements such as "Method A is superior to Method B for situation X."

Given a richer set of metrics, people may focus on information that confirms their prior beliefs. This occurred in the commentary on the M-competition, where the authors of the original study used different error measures to support their positions (Armstrong and Lusk, 1983).

In any event, the first order of business is to ensure that each of the measures that you do use is appropriate for the task. Consequently, we thought it was unfortunate that Winkler and Murphy used the Mean Square Error as an example of an overall measure. The A & C and F studies concluded that this measure was inappropriate for comparing methods across series.

We hope that these papers will encourage further research on this topic. Replications and extensions would help to better define the conditions under which various measures are most appropriate. In the meantime, to avoid biases and inefficient decision-making by forecasting researchers, we think one should make well-justified a priori choices of error metrics. We were interested to learn from Ahlburg's (1992) examination of 17 population forecasting studies that none of the authors justified their use of error measures. The current papers provide specific recommendations to help researchers choose error measures.

References

Ahlburg, Dennis A., (1984), "Forecast evaluation and improvement using Theil's decomposition," *Journal of Forecasting*, 3, 345-351.

Ahlburg, Dennis, 1992, "Commentary on error measures: Error measures and the choice of a forecast method," *International Journal of Forecasting*, 8, 99-111.

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

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

Armstrong J. Scott and F. Collopy, (1993), "Causal forces: Structuring knowledge for time series extrapolation," *Journal of Forecasting*, 12, 103-115.

Armstrong, J. Scott and E.J. Lusk, (1983), "Research on the accuracy of alternative extrapolation models: Analysis of forecasting competition through open peer review," *Journal of Forecasting*, 2, 259-311.

Carbone, Robert and J. S. Armstrong, (1982), "Evaluation of extrapolative forecasting methods: Results of a survey of academicians and practitioners," *Journal of Forecasting*, 1, 214-217.

Chatfield, Chris, (1992) "Commentary on error measures," *International Journal of Forecasting*, 8, 100-102.

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*, 38, 1394-1414.

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

Taylor, Stephen, J., (1992), "Commentary on error measures: Comparing forecasts in finance," *International Journal of Forecasting*, 8, 102-103.

Thompson, Patrick A., (1992), "Commentary on error measures: A statistician in search of a population," *International Journal of Forecasting*, 8, 103-104.

Winkler, Robert L. and Allan H. Murphy, (1992), "Commentary on error measures: On seeking a best performance measure or a best forecasting method," *International Journal of Forecasting*, 8, 104-107.

On error measures: A response to the commentators - the best error measure?, Robert Fildes, Management School, Lancaster University, Lancaster, LA14YX, UK.

There is much common ground between Armstrong and Collopy's paper (which I will subsequently refer to as AC), the commentators' viewpoints, and my own contribution [F: Fildes (1992)]. Both papers attempt to move beyond a complicated set of motherhood statements to critique and extend the existing literature on error measures by demonstrating some of their statistical properties. As Ahlburg (1992) and Taylor (1992) remark from the differing perspectives of demographics and finance, error measures in their respective fields are typically selected arbitrarily without due thought to the consequences. What both papers have shown, at least to the satisfaction of the commentators, is that in general certain standard error measures such as MAPS and MSE cannot be satisfactorily interpreted when used for comparing forecasting methods across data series.

There will, of course, be occasions when the empirical performance of the MADE or MSE is wholly adequate. However, in deciding on the error measures to use in a forecasting experiment, the researcher must make a choice a priori. Both AC and F suggest certain criteria by which the adequacy of an error measure should be judged. They include: reliability (AC, F), sensitivity to outliers (AC, F), interpretability (AC, F), reliability/ stability (AC, F), validity/ consensus (AC), descriptive of the underlying distribution (F). Both AC and F propose relative measures as being better behaved overall when judged by these criteria than absolute measures.

The AC and F studies then go on to show that a number of different measures such the geometric root mean square error (or Thompson's variant), or Theil's U2 seem to perform adequately based on these criteria, at least for the M-Competition data and the Telecommunications data series. This conclusion on error measure performance places a new requirement on forecasting researchers: to justify the use of their chosen error measure(s) with reference to the criteria established by AC and F.

The essential feature of an error summary statistic is to summarize the underlying error distribution. As Winkler and Murphy remark, this, in itself, is a simplification; a complete summary would include the joint (trivariate) distribution of the forecasts from the two methods under consideration as well as the observations themselves. While I do not argue with their analysis of the problem, I do doubt both the importance and the practicality of their suggestions for many business and economic applications. When analyzing method performance on a single series these distributions give rise to various diagnostic statistics useful in testing the method's adequacy (compared to some underlying statistical model). Even within this simple framework the analyst will need to find summary measures for these distributions based on prior analysis of the forms that are likely to prove satisfactory. This need arises because of the potential instability of a method's parameters (and the corresponding forecast distributions); this problem is not addressed in Winkler and Murphy's suggestions.

The typical use of error measures, however, is to aid in evaluating a forecasting method's performance (compared to alternatives) across a population of time series. In standard software implementations ad hoc measures have been used for this [Fildes and Beard (1992)]. In more thorough research papers, the authors [see for example, Makridakis et al. (1982)] have sought to avoid the problem by including most standard error measures in any evaluation, although often neglecting the relative measures recommended here by AC and F. Winkler and Murphy's recommendation of examining the error distribution relies on there being a wealth of time series data for each data series. In most forecasting comparisons, F being an exception, this is not the case and cross-sectional evidence is needed to help in choosing between methods.

In summary, the use of summary error measures to describe the underlying error distribution is common, and often the basis of major misinterpretations. The use of relative error measures and some consideration of the underlying sampling distribution has been amply justified by AC and F. Other, more complex alternatives have yet to prove their worth.

Let me turn now to some minor points. Like Ahlburg I disagree with AC's stress on consensus as a desirable characteristic. One can easily contrive examples where rankings will be reversed in shifting from one measure to another. Consensus among measures tells us more about the data series being used than the validity of a measure. The failure of two measures that are otherwise well-behaved to agree (eg. an absolute and a percentage error measure) is relevant material to be highlighted in the research - not a reason to exclude either measure.

Naturally I welcome Chatfield's clarifications of the points I wished to make about the choice of a forecasting method. Jenkin's concern that forecasting competitions use only one time origin from which to generate the forecasts was proved all too correct for the telecommunications data. While the analysis has not been carried out for the M-competition data I believe the same problem arises. Without an understanding of the fluctuations over time in the relative error measures (and the corresponding rankings) any recommendations on the best method are bound to be suspect.

While I agree with Chatfield about the quality of tables and graphs, he is as aware as I am that improving these exhibits is easier said than done. My own solution was to ask the referees and commentators. They were apparently no better than me!

Taylor's comment on the exclusion of 'outliers' from various calculations deserves further consideration. While such an exclusion may well be necessary if a simple summary is to be found, relegating the excluded outlier from further analysis should in general be avoided. The `Finance literature's' solution uses stable distributions with infinite variance and/or mean as models for the errors in order to include such extremes. I prefer `robust' approaches to estimating simpler models of the errors.

Thompson's (and Newbold's 1983) point about the need for some specific notion of population and sample is important. Potentially there are inferences to be made from the data series being analyzed to a larger population and from one time interval to another (for the same series). Although F's data series were not randomly selected they were chosen according to a well specified rule from a defined population. They therefore differ from the M-Competition data series, and the results of applying a sub-set of the M-Competition techniques to them also differ. However to abandon any thought of generalization from such competitions is foolish. After all, we have lived with the notion that an ARIMA model is a suitable representation of a time series for twenty years without specifying an explicit population for which the ARIMA representation is supposed to be suitable.

Writing now as editor of the International Journal of Forecasting I expect future authors to accept the consequences of these studies for comparative research. Let us hope we can convince the editors and referees of other journals as well as software designers to take the choice of error measure seriously. *It is not a matter of personal preference.*

References

Ahlburg, Dennis, (1992), "A commentary on error measures: Error measures and choice of a forecast method," *International Journal of Forecasting, 8*, 99-111.

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

Chatfield, Chris, (1992), "A commentary on error measures," *International Journal of Forecasting*, 8, 100-102.

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

Fildes, Robert and Beard, Charles, (1992), "Forecasting systems for production and inventory control," *International Journal of Operations and Production Management*, 12, No. 5, forthcoming.

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

Newbold, Paul, (1983), "The competition to end all competitions," *Journal of Forecasting*, 2, 276-279.

Taylor, Stephen J., (1992), "A commentary on error measures: Comparing forecasts in finance", *International Journal of Forecasting*, 8, 102-103.

Thompson, Patrick A., (1992), "A commentary on error measures: A statistician in search of a population," *International Journal of Forecasting*, 8, 103-104.

Winkler, Robert L. and Murphy, Allan H. (1992), "A commentary on error measures: On seeking a best performance measure or a best forecasting method," *International Journal o f Forecasting*, 8, 104-107.