6-1992

Generalization and Communication Issues in the Use of Error Measures: A Reply

Fred Collopy

J. Scott Armstrong
University of Pennsylvania, armstrong@wharton.upenn.edu

Recommended Citation
Collopy, F., & Armstrong, J. S. (1992). Generalization and Communication Issues in the Use of Error Measures: A Reply. *International Journal of Forecasting*, *8* (1), 107-109. http://dx.doi.org/10.1016/0169-2070(92)90015-2

Generalization and Communication Issues in the Use of Error Measures: A Reply

Abstract
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.

Disciplines
Business | Business Administration, Management, and Operations | Business Analytics | Business Intelligence | Management Sciences and Quantitative Methods | Marketing | Organizational Behavior and Theory

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.