Published in J*ournal of Forecasting*, 20, 2001, 273-283.

Identification of Asymmetric Prediction Intervals through Causal Forces

J. Scott Armstrong
The Wharton School, University of Pennsylvania

Fred Collopy
Weatherhead School of Management, Case Western Reserve University

Abstract

When causal forces are specified, the expected direction of the trend can be compared with the trend based on extrapolation. Series in which the expected trend conflicts with the extrapolated trend are called contrary series. We hypothesized that contrary series would have asymmetric forecast errors, with larger errors in the direction of the expected trend. Using annual series that contained minimal information about causality, we examined 671 contrary forecasts. As expected, most (81%) of the errors were in the direction of the causal forces. Also as expected, the asymmetries were more likely for longer forecast horizons; for six-year-ahead forecasts, 89% of the forecasts were in the expected direction. The asymmetries were often substantial. Contrary series should be flagged and treated separately when prediction intervals are estimated, perhaps by shifting the interval in the direction of the causal forces.

We extend the use of causal forces (Armstrong and Collopy, 1993) by testing whether they can help identify series that would produce asymmetric prediction intervals. Doing so would allow forecasters to adjust the confidence that they have in the forecasts, and perhaps to consider alternate procedures for estimating prediction intervals for these series.

Traditional approaches to constructing prediction intervals, which are based on the fit of a model to the data, are well reasoned and statistically sophisticated, but often of little value to forecasters. Chatfield (2001) reviewed the literature on this topic and concluded that traditional prediction intervals are often poorly calibrated for forecasting. In particular, they tend to be too narrow (i.e., too many actual observations fall outside the specified intervals).

The distributions of *ex ante* forecast errors differ substantially from the distributions of fitting errors (Makridakis and Winkler, 1989). (By *ex ante* forecast errors, we mean those based on forecasts that go beyond the calibration data and use no information from the forecast period.) The *ex ante* distribution provides a better guide to uncertainty than does the distribution of errors based on the fit to historical data. However, a higher than expected percentage of the forecasts still falls outside specified prediction intervals.

Because prediction intervals are typically too small, one obvious response is to enlarge them. Gardner (1988) used such an approach with traditional extrapolation forecasts. He calculated the standard deviation of the empirical *ex ante* errors for each forecast horizon and then multiplied the standard deviation by a factor based on the Chebyshev inequality. The resulting larger prediction intervals improved the calibration in terms of the percentage

of actual values that fell within the limits. However, widening the prediction intervals will not solve the calibration problem if the errors are asymmetric. The limits will be too wide on one side and too narrow on the other.

In the M-Competition (Makridakis *et al.,* 1987), leading academic researchers used additive extrapolation models whose forecast errors proved to be asymmetric. For six-year-ahead extrapolation forecasts using Holt's exponential smoothing, 33.1% of the actual values fell above the upper 95% limits, while 8.8% fell below the lower 95% limits (see Exhibits 3 and 4 in Makridakis *et al.,* 1987). One should expect about 2.5% of the errors to be above the upper limit and 2.5% to be below the lower limit. Although Makridakis *et al.* (1987) appear to have used an incorrect procedure for expanding the prediction intervals over the forecasting horizon (Koehler, 1990), their conclusion on asymmetry remains valid. Results were similar for the other extrapolation methods they tested. Corresponding figures for Brown's exponential smoothing, for example, were 28.2% on the high side and 10.5% on the low side. Errors for combined forecasts were even more asymmetric.

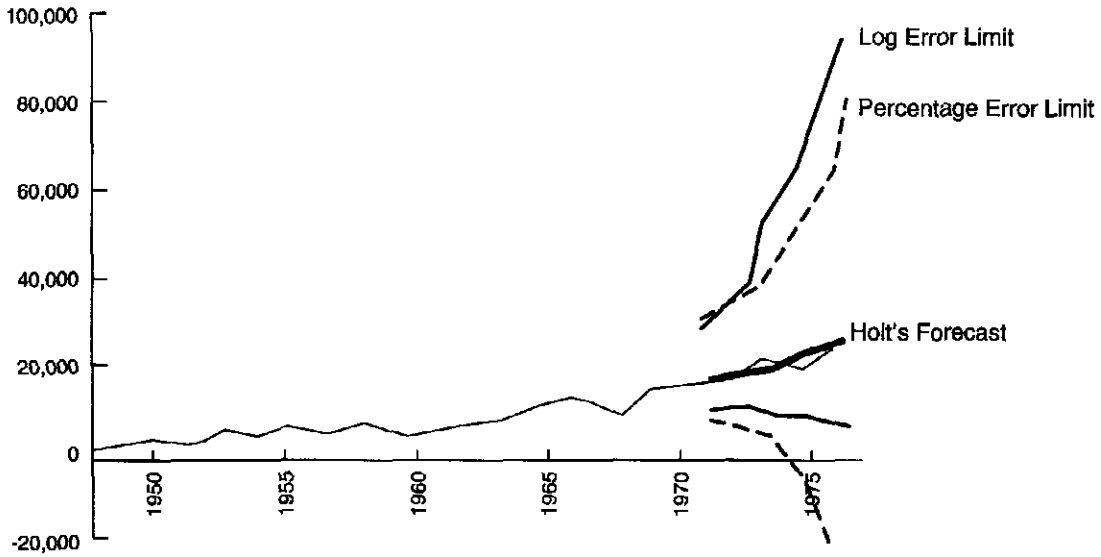
When Forecast Errors Are Asymmetric

Economists typically describe economic behavior using constant elasticities. Thus, economic models often use logs to represent variables. Sutton (1997) explains why it makes sense to think of economic behavior in these multiplicative terms (constant elasticities) rather than additive ones (marginal propensities).

If the underlying phenomena have constant elasticities, errors from an arithmetic forecasting model are expected to be asymmetric. The fact that the underlying phenomena have multiplicative effects does not mean that the forecasting model should necessarily be multiplicative. Those who forecast economic series face many considerations when selecting a model. If the series are unstable or highly uncertain, analysts often use additive rather than multiplicative trends to avoid large errors; rightly so in our opinion. Most analysts who made forecasts for the M-Competition used additive trends (Makridakis *et al.,* 1982).

Our interest is in testing the extent to which causal forces can be used to identify series that are likely to exhibit asymmetric errors when conventional extrapolation methods are used. For series with non-negative values, asymmetries are expected when errors from extrapolation methods are calculated using original units (e.g., as in Makridakis *et al.,* 1987). We needed first to reduce that source of asymmetries. We followed the common convention of using log transformations.

Figure 1. Illustration of shift when log intervals are used (M-Competition series 6: "Ford Automobile Net Sales")



We illustrate the application of log-symmetric intervals in Figure 1. These predictions of Ford automobile sales were from Holt's extrapolation as provided in the M-Competition study (Makridakis *et al.,* 1982). It is one of the series used in our test. The dotted lines show the standard 95% prediction intervals estimated by the average *ex ante* forecast errors for each time horizon by using successive updating over a validation period up to 1967. This successive updating procedure produced 28 one-ahead forecasts, 27 two-ahead, and so forth to 23 six-ahead forecasts. Note that the prediction intervals calculated in percentage errors are unreasonable for the longer forecast horizons because they are negative. In contrast, the prediction intervals calculated assuming symmetry in the logs seem more reasonable. The lower level is not negative and both the lower and upper limits are higher.

Forecasted Trends Conflict With Expectations

Armstrong and Collopy (1993) proposed using *causal forces* to summarize a domain expert's assessment of the net directional effect of the various factors expected to affect the trend over the forecast horizon. In thinking about a time series at a given point in time, it is useful to consider all of the important factors that affect trends in the series before deciding on the causal forces. For example, in forecasting unit sales of computers, one might consider forces such as rising incomes, increasing population, improvements in capabilities, and reductions in prices.

Conceptually, causal forces are independent of the historical trends in the time series. The only exception to this would be when information about the trend has an effect on behavior, such as the tulip bulb craze in Holland in the seventeenth century and Internet stocks in the twentieth.

If the causal forces tend to drive the series up, they are called **growth**. For example, when products are actively marketed in growing markets, the forces would be classified as growth. Unit sales of computers have growth forces.

If the forces tend to drive a series down, they are called **decay**. An example would be the costs of producing technical products such as computers. The historical trends for a such a series might fluctuate or be subject to cycles, but as long as the underlying forces are downward, the series is classified as decay.

If the forces are expected to move against the historical trend, they are **opposing**. An example would be inventory levels relative to sales. When inventories get large, holding costs lead managers to reduce their level. When they are too small, service suffers, causing decisions to increase inventories.

If the forces tend to move the series toward some mean, they are **regressing**. An example would be a measure of the performance of a professional athlete, such as a batting average; his average for the first three games of the current season would tend to regress toward his historical average. If he were a new player, the average might regress towards a typical average for new players.

If the forces reinforce the historical trend, they are called **supporting.** This might occur over specific periods for fashion crazes, fads, or market prices. For example, if real estate prices are going down, the perceived value of the neighborhood might go down. If prices are going up, people may perceive this as the place to live. Nevertheless, it is difficult to believe that a trend is primarily due to such forces. In our years of experience with hundreds of time series, we have yet to encounter a supporting series.

When one has little information about the factors that would affect a series over the forecast horizon, it is best to code the causal forces as **unknown**. Also, in cases where the factors are known, but operate in conflicting ways so that their net effect is unknown, it is best to code the series as unknown.

In summary:

Growth: values tend to get larger irrespective of current trend.

Decay: values tend to get smaller irrespective of current trend.

Opposing: values tend to go counter to the current trend.

Regressing: values tend to revert to some long-term mean value.

Supporting: values tend to move in the direction of the current trend.

Unknown: not enough is known to identify the forces, or the dominant forces conflict with one another.

When a domain expert's expectation about a future trend conflicts with traditional statistical extrapolation, we call the series "contrary." For example, if the causal force for a series is growth (that is, the domain experts expect the series to go up) and the forecasted trend (based, say, on Holt's estimate) is downward, the series is contrary. For such series, we hypothesized that the actual values will depart from the extrapolation in the direction of the causal forces. No prior research has addressed this hypothesis directly.

Collopy and Armstrong (1992a) examined the value of using causal forces to combine annual extrapolations of economic and demographic time series. A heavier weight was placed on an extrapolation method when the predicted direction was consistent with the causal force assessment. This led to more accurate *ex ante* forecasts than those from either traditional methods or from an equally weighted combination of forecasts. The improvements in accuracy were greatest when (a) experts could clearly specify causal forces, and (b) strong causal effects were likely, as in long-range forecasts. In the current study, we consider whether these same forces can be applied to the problem of asymmetric confidence intervals.

Design of the Study

Because of budget constraints and no theoretical reason to expect different results with different extrapolation methods, we used a single method in this study. Holt's exponential smoothing was selected because it is a widely-used procedure that has performed well in empirical studies.

We implemented Holt's method following the procedures described in Makridakis *et al.* (1982). Smoothing values were established by optimizing one-step-ahead forecast errors using a mean squared error criterion. Initial values were estimated by backcasting. We modified the procedure to stipulate that no forecast could be negative as such forecasts would be nonsensical for these data. Forecasts of negative values were replaced with a value that was half that of the most recent value. Sixteen forecasts (all from one series) were thus adjusted. These represent about four-tenths of 1% of the forecasts and had negligible effects on our results.

Data

We used annual data from the M-Competition and a similar set of economic and demographic data which we call the Weatherhead data. We used annual data for this test because causal forces can be expected to have a substantial effect on them.

M-Competition

We selected a probability sample of 18 annual series from the M-Competition (Makridakis *et al.*, 1982), specifically those whose identifiers end with the digit 6. They represent a 10% stratified subsample of the annual data in the M-Competition. These 18 series were from several countries for varying years over the period 1921 to 1976. The median length of these series is 21 years. Assessed at the point where six observations were held back for validation, there were significant long-term trends for 74% of the series (i.e., where the t-statistic for a simple linear trend was greater than 2.0). The trend from Holt's was upward for 76% of the series.

Weatherhead

Data on 26 annual economic series were collected under a project supervised by the second author. These data have characteristics similar to those of the M-Competition. They cover varying years over the period from 1921 to 1988 and their median length is 29 years. Assessed at the six-year holdout point, there were significant long-term

trends for 82% of the series, which is similar to the 74% for the M-Competition. The trend from Holt's was upward 79% of the time, again similar to the M-Competition's 76%.

Causal forces

The causal force codings were taken from an earlier study that was used to assess how to combine forecasts (Collopy and Armstrong, 1992a). The primary basis was the title of the series which was often sketchy. For example, the force for the M-Competition series "Production of cement in Turkey" was called growth.

We coded each series once and assumed that the causal forces remained constant over the forecast horizon. Collopy and Armstrong (1992a) provided evidence that this coding procedure is reliable. Because our knowledge of the series was limited, this study provides a weak test of the contrary series effect. Given better domain knowledge, we would expect to find stronger effects.

To determine whether a series is contrary at a given time, we compared the direction implied by the causal forces and the trend component from Holt's forecast. Growth series are contrary when Holt's forecasts a downward trend. Decay series are contrary when Holt's forecasts an upward trend. Opposing series are contrary irrespective of Holt's trend forecast.

Of our codings for the 18 M-Competition series, eleven were growth, two were opposing, one was regressing, and four were unknown. The Weatherhead data were similar to the M -Competition data, especially in that most series were growth. However, 19% of the Weatherhead data were decay series, whereas the M-Competition contained none. Neither data set contained supporting forces.

Forecasting Procedures

To analyze errors, we used successive updating. For the M-Competition data, we used the first three annual observations to estimate the parameters of the model. We then produced forecasts for the next six years and calculated errors. The fourth year was then added to the historical data, the model parameters were re-estimated, forecasts were made for the next six years, and errors were calculated. The procedure was repeated until the estimation data included all but the last year; this provided 1842 forecasts. A similar procedure was followed for the Weatherhead data. Because the series were longer, we used an initial fit of 12 years; this yielded 2220 forecasts. We made all of these decisions prior to any examination of the errors.

Results

For the 671 forecasts involving contrary series, 81% of the forecast errors were in the direction of the causal forces. For example, if the expectation was growth and Holt's predicted a downward trend, the actual was more likely to exceed the forecast. So, while we found no directional asymmetry on average, the use of causal forces allowed us to identify series where directional asymmetry was likely to occur. These results were statistically significant when compared against the null hypothesis that assumes that as many forecasts will be above the actual as below. Table 1 on the next page summarizes the results.

The bias in the direction of the causal forces became stronger as the forecast horizon increased. This was consistent with our expectation that causal forces would have stronger effects in the long term. The percentage of correct predictions increased consistently over the six-year period. This agreement with expectations is statistically significant at $p < 0.05$ using the Spearman rank correlation.

Asymmetry was most pronounced for contrary series based on growth or decay forces; overall, 85% of these errors were in the expected direction. The corresponding unweighted average for opposing series was 71%. This was expected because growth and decay forces were easier to identify than opposing forces.

Table 1. Percentage of errors in the expected direction for contrary series (number of forecasts; number of series)

	Horizon						Average
	1	2	3	4	5	6	
Growth							
M-Competition (63; 4)	73	91	91	91	90	89	87**
Weatherhead (172; 8)	74	80	86	86	89	89	84**
Decay							
Weatherhead (157; 5)	68	72	85	92	100	100	85**
Opposing							
M-Competition (162; 2)	56	67	71	73	77	77	62*
Weatherhead (117; 1)	82	81	85	79	88	88	84**
Average (671; 20)	69	76	82	83	89	89	81

** $p < 0.001$
* $p < 0.05$

Although we used a nonparametric sign test, the results overstate statistical significance because the forecasts were not all independent. What is most instructive, however, is that prior research about causal forces led to hypotheses about directional effects, and the results were consistent with these expectations.

The asymmetries for contrary series were large, as shown in Table 2. For series whose values were expected to go up based on our domain knowledge but where Holt's forecasted a downward trend, the median bias ranged from 20% to 56% low. If forecasts of a growth series were used as estimates of demand for a contrary series, the typical prediction intervals would be centered on a value that was too low by 20% (using the median). For series that were expected to go down, but whose extrapolations went up, the prediction interval would be centered up to 26% too high on the (median column). Although the magnitude of some of these results may be questionable because they are based upon small numbers of series, the patterns are consistent across all the data and conditions that were examined. The results were in the expected direction for each of the six conditions.

Table 2. Asymmetry of forecast errors for contrary series (number of forecasts; number of series)

	Bias (%)	
	Mean	**Median**
Expected up		
Growth		
M-Competition (63; 4)	-22	-20
Weatherhead (172; 8)	-38	-20
Opposing with trend down		
M-Competition (86; 1)	-58	-56
Expected down		
Decay		
Weatherhead (157; 5)	34	26
Opposing with trend up		
M-Competition (76; 1)	2	12
Weatherhead (117; 1)	10	9

One way to deal with contrary series is to select a more appropriate forecasting method (Armstrong and Collopy, 1993). For example, the naive forecast could be used for contrary series. To examine this, we substituted

naive forecasts for the 225 contrary forecasts from the M-Competition. This had only a modest effect; the random walk forecast errors were slightly more symmetric than those produced by Holt's. Unfortunately the literature to date offers no guidance as to what type of model might better estimate prediction intervals for contrary series.

Limitations

In an actual situation, those who do the coding would have access to historical data, and this is how we designed our study. However, because these events had already occurred, this raises the possibility that we were influenced by knowledge of actual outcomes. We think this possibility is remote. First, the coding was done rapidly and did not allow much time for reflection about what had happened. Second, we were clearly ignorant about what happened for almost all series (e.g., what happened to the production of virgin aluminum in England over any period, much less the horizon in question). Third, our endings can be replicated by those who lack information about the time periods involved. To test this, we gave a description of causal forces, based on that used in this paper, to eight people (five academicians and three practitioners). We asked them to code the 18 series based only on the names provided in the M-Competition data set. They did not see the time series. They were asked to specify the forecasts over a six-year horizon, but we specified no historical time period. (This matches our assumptions that the causal forces are independent of the historical data and that they tend to remain constant over time.) The modes of their codings were in agreement with our codings on 13 of the 18 series. The differences occurred for series where uncertainty was high (based on the use of unknown ratings and on many differences among the coders). This agreement represents excellent reliability, given that a series could be coded in one of six categories.

Our findings are limited in that we examined only annual data. We expect similar findings for quarterly and monthly data, but the strength of the effect should be weak when the interval is short. We do not claim to have identified all situations where log errors might be asymmetric. Factors other than contrary series might cause asymmetric errors.

Other transformations might yield further improvements in calibration. We used logs because this is a common approach to the treatment of economic data and because it is easy to do. Tests of alternative formulations would require much larger samples of forecasts.

Our study was concerned with obtaining a better understanding of the conditions under which prediction intervals are asymmetric. Further research is needed on how our recommendations affect the calibration of the prediction intervals. For example, what percentage of the actual values fall within the 95% prediction intervals when log errors are used? We suggest that these tests be conducted separately for contrary series and other series.

One approach that we believe to be promising would be to ask domain experts to prepare judgmental extrapolations for contrary series. We would expect this method to reduce asymmetry, but we were unable to test it with our data.

Discussion

When we asked four experienced forecasting consultants about industry practices, they told us that organizations almost always use prediction intervals that are symmetric in the original units. So the asymmetric errors problem exists in practice, except where forecasting models themselves use logs. Few forecasting packages advise the user to construct asymmetric prediction intervals.

In terms of decision making, it is sensible that analysts present prediction intervals in terms of the units of the series being forecast. However, intermediate analyses need not use original units. Analysts could use other transformations, such as logs, to calculate prediction intervals and transform the results back to original units.

Our recommendation is that the analyst select the appropriate forecasting method with little concern for the asymmetry of errors. If an additive trend model is used, one can convert the forecasts and actuals to logs, analyze errors to construct symmetric intervals, and then report prediction intervals in the original units. This would have

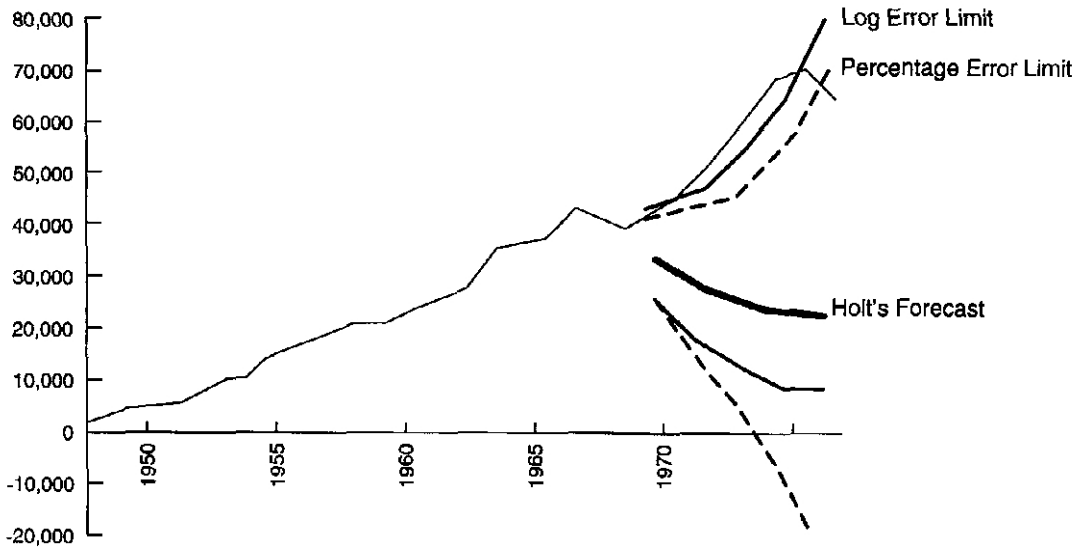
reduced much of the problem of asymmetric errors for annual M-Competition data. We plotted errors from about 4000 annual forecasts and the distribution of the errors was relatively symmetric. From the viewpoint of the forecast user, the prediction intervals, as recast in original units, would be asymmetric.

Despite best efforts at model selection, log errors for contrary series are expected to be asymmetric. We do not know how best to construct prediction intervals for contrary series. We speculate that empirical adjustment factors might be used to shift the prediction intervals in the direction of the causal forces. For example, for the contrary growth series, we would recommend shifting the intervals in the direction of the forces by about 20% (using our median estimate from Table 2). In any case, one should flag these contrary series, recognize that the prediction intervals are not appropriate, and consider other ways to forecast them and to estimate their prediction intervals. In summary then, we suggest the following procedure:

- Select a forecasting method and model.
- If an additive trend model was used, convert forecasts and actuals to logs.
- Analyze the historical errors to produce a symmetric prediction interval.
- If the series is a contrary one
  - Use a forecast with no trend, and
  - Shift the prediction interval in the direction of the causal forces.
- If an additive trend model was used, convert the prediction interval back into original units (they will look asymmetric).

To illustrate, we provide an example of a contrary series for French automobiles (Figure 2). Holt's forecast is from the M-Competition. The 95% prediction intervals were estimated from *ex ante* forecast errors by using successive updating to produce twenty forecasts per horizon through 1968. As before, the use of log intervals shifted the prediction interval upwards and, of course, it avoided negative values for the lower limit. In this case, however, some actual values fell above the upper limit. Had we shifted the upper limit up by 20%, as suggested above, all of the actuals would have fallen within the prediction intervals.

Figure 2. Illustration of remaining asymmetry when log intervals are used for contrary series (M-Competition Series 3: "Production of Main Car Constructors, France")



The results from the Weatherhead data were similar to those from analysis of the M-Competition data. This adds support to the conclusions from prior research, including Collopy and Armstrong (1992b), Fildes *et al*. (1998), and Makridakis *el al*. (1993, 2000), that findings from analyses of M-Competition data can be generalized.

Conclusions

When the forecasted trends are contrary to the causal forces, the logged errors are expected to be asymmetric. For such series, we recommend that non-trended forecasts be used. However, this will not eliminate the asymmetry, so we also recommend that the prediction intervals be shifted in the direction of the expected bias. In any event, forecasts involving contrary series should be flagged and treated separately because of the difficulty in estimating their prediction intervals.

This study provides another test of the value of causal forces. Previously they have been shown to be useful for weighting forecasts and for selecting forecasting methods. Here they are also useful for identifying series in which the forecasted (logged) errors can be expected to be asymmetric.

Acknowledgements

Partial support for this research has been provided by the US Navy Personnel R&D Center and by the Office of Naval Research (under grant number N00014-92-J-1544). Helpful comments on early drafts were received from P. Geoffrey Allen, Christopher Chatfield, Ulrich Kuesters, Robert Fildes, Anne Koehler, Cam Rungie, Leonard J, Tashman, and J. Thomas Yokum. John Carstens, Mary Haight, Vanessa Lacoss, and Suzanne Berman provided editorial assistance.

References

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

Chatfield, C. (2001), "Prediction intervals for time series," in *Principles of Forecasting: A Handbook for Practitioners and Researchers,* Armstrong, J. S. (ed.), Norwall, MA: Kluwer Academic Publishers, pp. 475-494.

Collopy, F. and J. S. Armstrong (1992a), "Rule-based forecasting: Development and validation of an expert systems approach to combining time series extrapolations," *Management Science*, 38, 1394-1414.

Collopy, F. and J. S. Armstrong (1992b), "Generalization and communication issues in the use of error measures: A reply," *International Journal of Forecasting,* 8, 107-109.

Fildes, R., M. Hibon, S. Makridakis, and N. Meade (1998) "Generalizing about univariate forecasting methods: Further empirical evidence (with commentary)," *International Journal of Forecasting,* 14, 339-366.

Gardner, E.S., Jr. (1988) "A simple method of computing prediction intervals for time series forecasts," *Management Science,* 34, 541-546.

Koehler, A. B. (1990), "An inappropriate prediction interval," *International Journal of Forecasting,* 6, 557-558.

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

Makridakis, S., C. Chatfield, M. Hibon, M. Lawrence, T. Mills, K. Ord, and L. F. Simmons (1993), "The M-2 competition: A real-time judgmentally-based forecasting study*, International Journal of Forecasting,* 9, 5-39 (includes commentary).

Makridakis, S., M. Hibon, E. Lusk, and M. Belhadjali (1987), "Confidence intervals," *International Journal of Forecasting,* 3, 489-508.

Makridakis, S. and R. Winkler (1989), "Sampling distributions of post-sample forecasting errors," *Applied Statistics,* 38, 331-342.

Sutton, J. (1997,) "Gibrat's legacy," *Journal of Economic Literature,* 35, 40-59.