COMMENTARIES ON THE COMPETITION
EVERETTE S GARDNER JR
Journal of Forecasting (pre-1986); Jul-Sep 1983; 2, 3; ABI/INFORM Global
pg. 263

COMMENTARIES ON THE COMPETITION

The Trade-offs in Choosing a Time Series Method

EVERETTE S. GARDNER, JR.
Commander, Supply Corps, U.S. Navy, Management Information Systems Officer, U.S. Atlantic Fleet Headquarters, Norfolk, Virginia 23511, U.S.A.

How can the results of the M-Competition be used in practice? This paper attempts to answer that question by generalizing about the relative accuracy of the methods tested. Although there are many objections to such generalizations (see Jenkins, 1982, for example), I can see no other way to develop some principles for model selection. There is certainly no generally accepted theory to guide the applied forecaster.

The first section of the paper reviews the accuracy criteria used in the M-Competition. Next, the performance of each forecasting method is evaluated. Within the group of exponential smoothing methods, I contrast the results with what should be expected—both from simulation work and from theoretical studies of frequency and impulse response functions.

ACCURACY CRITERIA

Median APE vs. MAPE

The median APE is the most descriptive measure of the central tendency of errors in the forecasting competition. The error distributions from all methods are badly skewed, which distorts the MAPE. The median APE is less affected. By definition, the median APE falls between the mode and the MAPE in a skewed distribution.

The MSE criterion

The ability of a forecasting method to avoid large errors is often more important than the central tendency of errors. The only accuracy criterion used in the M-Competition that gives extra weight to large errors is the MSE. Although the MSE results are difficult to interpret, they should not be overlooked. Several examples will illustrate why the MSE was ranked as the most important accuracy criterion by practitioners in the Carbone and Armstrong (1982) survey.

Consider forecasting for production planning. Large errors can be disastrous, since physical production capacity (plant and equipment) is fixed in the short run. Some flexibility usually exists to adjust the output rate for forecast errors by overtime, layoffs, subcontracting, and so forth; but large positive errors can result in a significant loss of market share before capacity can catch up to demand. Large negative errors can drive output below the break-even point for a given capacity level.

Forecasting for budgeting is another case where it is important to avoid large errors. Anyone who has had to manage an operation on a fixed budget can attest to the disruption caused by large forecasting errors.

Forecasting for inventory control is the most frequent application of time series methods. Again the MSE is the most important accuracy criterion. Safety stocks are based on the variability of the forecast errors, as computed by the MSE or equivalent measures.

The MSE results in the forecasting competition are summed across all series, although the levels of the series vary widely. The series might be sorted into groups with similar levels to make the MSE results easier to interpret, perhaps by using the root MSE. Despite these interpretation

Revised October 1982 and March 1983

problems, the MSE results in their present form do give some idea of the stability of each forecasting method. Some tentative conclusions using the MSE criterion are discussed below.

SOPHISTICATED METHODS

Bayesian forecasting

Among other objectives, the Bayesian multi-state model was designed to avoid large errors. Although its performance was mediocre on other accuracy criteria, the Bayesian model gave the best MSE (average of all horizons) on both 1001 and 111 series.

Lewandowski

Lewandowski was the best overall choice on the median APE criterion in the 111 series, was second only to Bayesian forecasting in MSE, and was the best long-range forecaster on any criterion. The major reason for these successes is the manner in which Lewandowski searches among several nonlinear possibilities for trend. This search usually produced a steadily decreasing rate of trend in the forecasts. Most other methods, particularly those based on a linear trend, had a tendency to overshoot the data at longer horizons.

Parzen

Parzen may be the most robust method tested, considering all accuracy criteria, types of data, and horizons. It is unfortunate that this method was run on only 111 series. Robustness would be more convincing on all 1001 series.

Box–Jenkins

I can see no important advantage for Box–Jenkins anywhere in the M-Competition. Although Makridakis places Box–Jenkins in a group of eight unusually robust methods (footnote to Table 33), Parzen is equally robust and has the considerable advantage that it can be used completely automatically.

COMBINING METHODS

Makridakis recommends the Combining A method over any of its components used individually. I disagree. Combining A is superior in MAPE and average ranking to its components, but not in median APE or MSE. Holt or Holt–Winters do about the same as Combining A in median APE at all horizons, using 111 or 1001 series. Any of Holt, Holt–Winters, or Brown consistently beats Combining A in MSE.

The MSE comparisons are surprising. It seems intuitive that a combined forecast would avoid large errors. The problem is that single smoothing and ARRES are poor choices on the MSE criterion. They inflate the MSE of Combining A to the point where it is worse than the MSE of the other components.

Considering the start-up and maintenance problems associated with running six different methods at once, I find it difficult to justify combining methods. Maintenance problems are compounded by the fact that four of the six methods use fixed parameters. If repetitive forecasts are made over time, the fixed-parameter methods would have to be refitted periodically. Between refittings, these methods would have to be monitored with tracking signals to adjust for outliers and bias. All this bother is unreasonable in view of the accuracy comparisons. (*Editor's note:* for alternative viewpoints, see the Commentary by Geurts and the Reply by Winkler.)

SIMPLE METHODS

Moving averages, quadratic exponential smoothing, linear regression

These methods were the worst forecasters overall. In most cases, one could do better using Naïve 2. It is surprising that single exponential smoothing did so much better than moving averages since the two are closely related. Previous research (see Armstrong, 1978b, for a review) has found little difference between exponential smoothing and moving averages.

Automatic AEP

Automatic AEP is presently the only reasonable alternative to exponential smoothing in applications where simplicity is important. There is little difference in accuracy between AEP and Holt in non-seasonal data. AEP may be more attractive for large applications in non-seasonal data, since it requires no maintenance. For seasonal data, AEP was one of the worst methods tested.

Single smoothing: fixed vs. adaptive parameters (ARRES)

Single smoothing was a good choice for one-step-ahead forecasting on all criteria except the MSE. The trend-adjusted smoothing models gave a better MSE.

Models with adapative smoothing parameters such as ARRES appear to be widely used in practice. However, the empirical evidence indicates that fixed parameters yield more accurate forecasts. Both the forecasting competition and the simulation study by Gardner and Dannenbring (1980) support this conclusion.

Using either 1001 or 111 series in the M-Competition, the overall median APE and MSE favoured single smoothing with a fixed parameter over ARRES. These comparisons were for one-step-ahead forecasting, which is what all these models are designed to do. Within the 111 series, there were 14 one-step-ahead comparisons in average ranking and median APE. Every comparison favoured single smoothing with a fixed parameter.

In the Gardner and Dannenbring study, 9000 times series were simulated with a variety of noise levels and characteristics (constant mean, constant trend, sudden shifts in mean and/or trend, changes in direction of trend). The simulation results showed that ARRES had a tendency to overreact to purely random fluctuations in the time series. This instability usually offset the response rate advantage of ARRES when a sudden shift in the series occurred.

For time series with a constant mean, smoothing with a fixed parameter in the 0.05 to 0.10 range yielded a significantly smaller MSE than ARRES. For both stable series and those subject to sudden shifts in the mean, there was no significant difference between using a fixed parameter in the 0.30 to 0.40 range and ARRES.

For a discussion of several other empirical studies on adaptive exponential smoothing, see Ekern (1981, 1982). Ekern concludes that there is no evidence that adaptive smoothing models are superior to models with fixed parameters.

Trend-adjusted exponential smoothing: Holt vs. Brown

Both the Holt and Brown trend-adjusted smoothing models are widely used in practice. Analysis of frequency and impulse response functions by McClain and Thomas (1973) and by McClain (1974) shows that Brown should be preferred on theoretical grounds. The Brown formulation is critically damped, which means that it gives the most rapid possible response to a change in the time series without overshoot. The Holt model will oscillate badly when many intuitively appealing values of the smoothing parameters are used. One common situation in which the Holt model oscillates is when its smoothing parameters are equal. For example, with both parameters set at 0.1, the Holt model will oscillate for 72 periods after an impulse signal in an otherwise noise-free series.

There is no evidence that Holt's rather obscene response functions have any effect on forecast accuracy In the Gardner and Dannenbring study, there were rarely any statistically significant differences in MSE between Holt and Brown. However, the Holt model had a small advantage on most series in one-step-ahead forecasting. The reason for this is that the additional parameter in the Holt formulation gives a better fit to many kinds of series. For example, when a series has a negligible trend, the Holt trend parameter can be set near zero. For series subject to sudden changes in level or trend, the corresponding Holt parameter can be increased, while holding the other parameter at a lower, more stable level.

The results of the M-Competition also give Holt a small edge over Brown. Holt's overall median APE is better using both 1001 and 111 series. Brown's MSE is better using 1001 series, but Holt is better using 111 series. Within the 111 series, Holt was better than Brown in median APE on most types of data.

Smoothing on deseasonalized data vs. the Winters method

McClain (1974) makes a strong case for smoothing with deseasonalized data. Using frequency and impulse response functions, he shows that the Winters method of updating the seasonal factors one at a time through exponential smoothing should make the forecasts highly sensitive to noise.

To illustrate, suppose that a large random impulse occurs in a time series being forecasted with Holt–Winters. Depending on the set of smoothing parameters used, some portion of this impulse will be misinterpreted as a change in both mean and trend. Fortunately, the distortion will be removed in a reasonable length of time by the smoothing process.

However, some portion of the random impulse will also be absorbed by that period's seasonal factor. If L is the length of the seasonal cycle, that seasonal factor will not be smoothed again for L time periods. Many years may be necessary to wash out the effects of a single random impulse, which could lead to unstable forecasts.

In the M-Competition, there is no evidence that this problem has any effect on forecast accuracy. Most comparisons give Holt–Winters some margin over deseasonalized Holt. When storage problems are considered, Holt–Winters has an important advantage. To smooth with deseasonalized data, the raw data from several cycles must be stored in order to update the average seasonal factors. With Holt–Winters, only the seasonal factors themselves have to be stored.

CONCLUSIONS

The trade-offs in choosing a time series method can be summarized as follows, using the median APE and MSE criteria:

When simplicity is important in the proposed application, the choices can be reduced to Holt or AEP in non-seasonal data. Although single smoothing is a reasonable choice for one-step-ahead forecasting, there is no apparent penalty for using Holt on all series to give some protection against the development of trends. In seasonal data, Holt–Winters is the best choice.

If a specialist is available to support the forecasting system, several sophisticated methods should be considered. Over all horizons and types of data, Bayesian forecasting or Lewandowski should give the best MSE and Lewandowski the best median. At long horizons, Lewandowski should be the best choice on any criterion. When there is difficulty in finding an adequate model, Parzen should be considered because of its robustness.

There was not much difference in the M-Competition between Holt–Winters and sophisticated methods in seasonal data. However, it would be foolish to overlook sophisticated methods because most can be used completely automatically.

APPENDIX

An erroneous formulation is presented by Makridakis *et al.* for Brown's linear trend model. The Brown model, as presented on p. 144, is:

$$S'_t = \alpha X_t + (1 - a)S'_{t-1}, \tag{1}$$

$$S''_t = \alpha S'_t + (1 - \alpha)S''_{t-1}, \tag{2}$$

$$\hat{X}_{t+1} = a_t + b_t, \tag{3}$$

where

$$a_t = 2S'_t - S''_t \tag{4}$$

$$b_t = (1 - a)^{-1}(S'_t - S''_t) \tag{5}$$

In equation (1), $(1 - a)$ should be $(1 - \alpha)$. In equation (5), $(1 - a)^{-1}$ should be $\alpha/(1 - \alpha)$. These errors are typographical. The authors used the correct model in the computer work.