Commentary by J. Scott Armstrong on
Fildes *et al.*: "Generalizing about univariate forecasting methods: further empirical evidence,"
International Journal of Forecasting, 14 (1998), 339-358,

Reprinted with permission from
International Journal of Forecasting,14 (1998), 359-361

Fildes, Hibon, Makridakis and Meade (1998), which will be referred to as FHMM, extends two important published papers. The idea of taking findings from each study and testing them against the data used in the other study is a good one. Such replications and extensions are important in the effort to develop useful generalizations and publication of this paper reflects the commitment of International Journal of Forecasting to replication research. In addition the study examines procedures for estimating smoothing parameters, and it evaluates the need for using multiple starting points when evaluating forecasting methods.

On the negative side, FHMM does not fully describe the conditions under which one might expect a given extrapolation method to provide more accurate forecasts than competing methods. This limits the generalizability of its findings. In addition, I believe that the FHMM generalizations are even more limited than they might appear at first glance.

1.1. Defining conditions

Generalization can be viewed as identification of the procedures under which a particular result is expected to hold. Thus, to develop generalizations about the best method for time series extrapolation, one needs good characterizations of the time series. FHMM provides some descriptors to characterize their series, but they do not go as far as the current literature would permit. For example, their Figure 2 shows that for each data set, most of the series involved a strong trend. As to differences between the series, we learn that the telecommunications data involve primarily downward trending series (about 98% of the series), while the M-competition involves mostly upward trending series (perhaps 80%). Also the telecommunications data display much less variability than do the M-competition data, although they have substantially more outliers. The telecommunications data are all monthly, while the M-competition data include monthly, quarterly, and annual data. But more information could have been used in their analysis. The telecommunications series, for the most part, represent special services, and the expectation was that they would go down because they were being replaced by other services. They are "decay" series, using Armstrong and Collopy (1993) terminology for causal forces. In contrast, the M-competition data are comprised of data with primarily growth forces, although some are opposing or decay. An analysis of a sample of the annual series in the M-Competition by Collopy and Armstrong (1992), showed 68% growth, 18% unknown, 7% opposing, and 2% decay.

Researchers should measure and report conditions that might be related to the selection or weighting of an extrapolation method. If this were done, it would aid our ability to draw generalizations from prior studies and extensions. Drawing upon expert opinion and prior research, Collopy and Armstrong (1992) used a broad array of features to describe the conditions for a time series. These include causal forces (growth, decay, supporting, opposing, regressing, or unknown), short- and long-term trends (direction and statistical significance), uncertainty (coefficient of variation about the trend line and whether the long- and short-term trends were in the same direction), instability (including such things as level discontinuities and outliers), cycles, and functional form. This descriptive scheme has also been used in subsequent studies; Armstrong and Collopy, 1992; Vokurka et al., 1996; Adya et al., 1998). It is disappointing that FHMM fail to consider many of these conditions.

It is difficult to generalize about an entire set of data when it involves multiple conditions. Ideally, series should be classified by type. Thus, we might hypothesize that for decay series with low uncertainty and low instability, such as for most of the telecommunications data, a full trend model would be preferred to a damped trend. As another example, if any of the annual series were ‘contrary,’ which is to say that the statistical trend forecast is opposite in direction to the prior expectation of experts in that area (Armstrong and Collopy, 1993), then one would expect to encounter outliers if the forecasts are produced by trend extrapolation; as a result, damped trend should be expected to be more accurate than full trend models.

Hypotheses have been developed about how these conditions relate to the relative accuracy of various methods (Collopy and Armstrong, 1992). Ideally then, one should not ask management to describe what their data are like on average. Rather, each series should be classified, the series should be grouped by type, and a forecasting method should be selected for each type. Of course, in practice, one might be restricted to using only one method, and general descriptors might be of some value.

FHMM refer to the telecommunications data as being ‘homogeneous’ with the implication that homogeneity aids in generalization. However, while homogeneity is an aid to reliability, it is an enemy of generalization in that one learns relatively less that is applicable to other sets of time series. Consider the effects if all series were exactly the same. One could then learn as much by looking at a single series as by looking at many. So by adding a narrow data set to the broad data of the M-competition, one adds only modestly to the generalizability of the findings. On the other hand, homogeneity is advantageous when trying to determine which method is most appropriate to the conditions, so the telecommunications data play a useful role in this study. We are unlikely to find methods that are best in all conditions, so future studies might better study sets of homogeneous time series. This benefit could also be achieved by segmenting the M-competition time series. When heterogeneous time series are combined, such as has been done in some M-competition analyses (e.g., combining monthly, quarterly, and annual data), it is difficult to draw generalizations.

1.2. Some limitations

The analysis in FHMM is sound. The procedures and data are fully disclosed and are available to other researchers. The design is objective. A variety of forecast errors were used to check the robustness of the conclusions threats. Furthermore, the test on the value of multiple starting points was useful; it showed that this is an important procedure even when the analysis involves a large number of time series. This finding is surprising and useful, in my opinion.

FHMM have tested generalizations for the selection of the most appropriate method. But their generalizations are restricted because they do not use domain knowledge. Managers often have useful information and it is not obvious that the FHMM conclusions will generalize when such information is used.

The use of an additive model for extrapolation seemed to be unfortunate, especially for the telecommunications data. Given that the series are long, uncertainty is low, instability is moderate, and that it makes sense to think of sales data as changing in percentage terms, the use of a multiplicative trend seems relevant. The multiplicative form also has the nice feature that it does not allow for negative forecasts. On the other hand, the fact that only monthly data were examined makes my recommendation less important. This is an empirical issue and it could be tested.

Little empirical evidence exists on the value of alternative estimation procedures for the parameters for exponential smoothing, so it was refreshing to see this issue addressed. On the other hand, some prior research has been done (e.g., Dalrymple and King, 1981). It seemed to me that an alpha of 0.1 was low for such data (the telecommunication data was used for this analysis). The beta value of 0.01 (for the trend) is enough to render the model to be unresponsive to recent data; it would correspond to a 199-month moving average. Thus, I believe their conclusion that the parameter estimates should not be set by judgment deserves further study. This need for further study is reinforced by their results showing that parameters estimated for similar series produced the most accurate forecasts.

Their conclusion that updating the parameter estimates improves accuracy seems sensible, especially because it can provide a larger sample size. Furthermore, their results are consistent with findings in Dalrymple and King (1981).

The authors' favorable interpretation of the findings on robust trend puzzled me. On the surface, robust trend was not especially accurate when used on the M-competition data. I believe that the findings are consistent with what one might expect. Robust trend is expected to have a relative advantage when the historical time series contain outliers. Here, FHMM provide information about conditions: outliers were much more common in the telecommunications data than in the M-competition. Thus, one might have expected a priori that the robust trend model is less appropriate to the M-competition data.

1.3. Conclusions

This paper adds to our confidence about conclusions from prior research. It also helps to better define the conditions under which robust trend is appropriate. However, it does little to advance our knowledge of the reasons why an extrapolation method should provide accurate forecasts. Further research on generalizations about univariate forecasting methods should rely upon careful identification of the conditions.