ELSEVIER

International Journal of Forecasting 16 (2000) 437–450

international journal of forecasting

www.elsevier.com/locate/ijforecast

Out-of-sample tests of forecasting accuracy: an analysis and review

Leonard J. Tashman*

School of Business Administration, University of Vermont, Burlington, Vermont 05405, USA

Abstract

In evaluations of forecasting accuracy, including forecasting competitions, researchers have paid attention to the selection of time series and to the appropriateness of forecast-error measures. However, they have not formally analyzed choices in the implementation of out-of-sample tests, making it difficult to replicate and compare forecasting accuracy studies. In this paper, I (1) explain the structure of out-of-sample tests, (2) provide guidelines for implementing these tests, and (3) evaluate the adequacy of out-of-sample tests in forecasting software. The issues examined include series-splitting rules, fixed versus rolling origins, updating versus recalibration of model coefficients, fixed versus rolling windows, single versus multiple test periods, diversification through multiple time series, and design characteristics of forecasting competitions. For individual time series, the efficiency and reliability of out-of-sample tests can be improved by employing rolling-origin evaluations, recalibrating coefficients, and using multiple test periods. The results of forecasting competitions would be more generalizable if based upon precisely described groups of time series, in which the series are homogeneous within group and heterogeneous between groups. Few forecasting software programs adequately implement out-of-sample evaluations, especially general statistical packages and spreadsheet add-ins.

Keywords: Out-of-sample; Fit period; Test period; Fixed origin; Rolling origin; Updating; Recalibration; Rolling window; Sliding simulation; Forecasting competitions

1. Introduction

In this paper, I discuss the implementation of out-of-sample tests of forecasting accuracy. Section 2 summarizes the rationale for out-of-sample testing. Section 3 compares fixed-origin and rolling-origin procedures. Section 4 examines the application of out-of-sample testing to an individual time series: Issues addressed are rules for splitting the series between fit and test periods, updating versus recalibrating model coefficients, single versus multiple test periods and the use of rolling windows. Section 5 considers the role of out-of-sample testing in method selection. Section 6 describes the extension of out-of-sample testing from an individual time series to multiple time series and forecasting competitions. Section 7 evaluates the adequacy of out-of-sample tests in forecasting software. Section 8 contains my conclusions and recommendations.

*Tel.: +1-802-425-3805; fax: +1-802-425-3806.

E-mail address: lentashman@compuserve.com (L.J. Tashman).

0169-2070/00/$ – see front matter
PII: S0169-2070(00)00065-0

2. In-sample versus out-of-sample evaluation

Forecasters generally agree that forecasting methods should be assessed for accuracy using out-of-sample tests rather than goodness of fit to past data (in-sample tests). 'The performance of a model on data outside that used in its construction remains the touchstone for its utility in all applications; (Fildes and Makridakis, 1995, p. 293).

The argument has two related aspects. First, for a given forecasting method, in-sample errors are likely to understate forecasting errors. Method selection and estimation are designed to calibrate a forecasting procedure to the historical data. But the nuances of past history are unlikely to persist into the future, and the nuances of the future may not have revealed themselves in the past.

Overfitting and structural changes may further aggravate the divergence between in-sample and post-sample performance. The M-competition (Makridakis et al., 1982) and many subsequent empirical studies show that forecasting errors generally exceed in-sample errors, even at reasonably short horizons. As well, prediction intervals built on in-sample standard errors are likely to be too narrow (Chatfield, 1993, p.131).

Moreover, common extrapolative forecasting methods, such as exponential smoothing, are based on *updating* procedures, in which one makes each forecast as if one were standing in the immediately prior period. For updating methods, the traditional measurement of goodness-of-fit is based on *one step-ahead errors* — errors made in estimating the next time period from the current time period. However, research shows (e.g., Schnaars, 1986, Exhibit 2, p.76) that errors in forecasting into the more distant future will be larger than those made in forecasting one step ahead.

The second aspect to the argument is that methods selected by best in-sample fit may not best predict post-sample data. Bartolomei and Sweet (1989) and Pant and Starbuck (1990) provide particularly convincing evidence on this point.

One way to ascertain post-sample forecasting performance is to wait and see in real time. The M2-competition (Makridakis et al., 1993) did exactly this. In one phase, forecasts (for 1–15 months ahead) made in September 1987 were evaluated at the conclusion of 1988.

Real time assessment has practical limitations for forecasting practitioners, since a long wait may be necessary before a reliable picture of a forecasting track record will materialize. As a result, tests based on *holdout* samples have become commonplace. The *fit period* is used to identify and estimate a model (or method) while the *test period* is reserved to assess the model's forecasting accuracy.

If the forecaster withholds all data about events occurring after the end of the fit period, the forecast-accuracy evaluation is structurally identical to the real-world-forecasting environment, in which we stand in the present and forecast the future. However, 'peeking' at the held-out data while selecting the forecasting method pollutes the evaluation environment.

3. Fixed-origin versus rolling-origin procedures

An out-of-sample evaluation of forecasting accuracy begins with the division of the historical data series into a *fit period* and a *test period*. The final time in the fit period (T) — the point from which the forecasts are generated — is the *forecasting origin*. The number of time periods between the origin and the time being forecast is the *lead time* or the *forecasting horizon*. The longest lead time is the N step-ahead forecast. Equivalently, N denotes the length of the test period.

3.1. Fixed-origin evaluations

In performing an out-of-sample test, we can use either a single forecasting origin or multiple forecasting origins. The former can be called a *fixed-origin* evaluation: Standing at origin (T), we generate forecasts for time periods $T+1$, $T+2, \ldots T+N$. By subtracting each of these forecasts from the known data values of the test period, we determine the forecast errors. We can average the errors in various ways to obtain summary statistics.

Applied to a single time series, the fixed-origin evaluation has several shortcomings. Because it yields only one forecast (and hence, only one forecast error) for each lead time, it requires a fairly long test period to produce a forecasting track record. Second, forecasts generated from a single origin are susceptible to corruption by occurrences unique to that origin. Third, in the usual software implementation of a fixed-origin evaluation, summary error measures are computed by averaging forecasting errors across lead times. The resulting summary statistic is a mélange of near-term and far-term forecast errors.

We can partly overcome the three problems by successively updating the forecasting origin. We can also mitigate the problems by using *multiple time series.* Still, even within a single-series context, the fixed-origin evaluation can play a useful role: it is the only way we can assess the post-sample accuracy of forecasts, such as judgmental forecasts, when we do not know or can not replicate the underlying forecasting methodology.

3.2. Rolling-origin evaluations

In a *rolling-origin evaluation*, we successively update the forecasting origin and produce forecasts from each new origin. One of the first explicit descriptions of the procedure was Armstrong and Grohman's (1972). Armstrong (1985, p. 343) provides a schematic illustration of the rolling-origin procedure.

When $N=4$, for example, the fixed-origin evaluation results in four forecasts, all from origin T. The rolling-origin evaluation also generates four forecasts from this origin, but then supplies an additional three forecasts from origin $T+1$, two from origin $T+2$, and one from origin $T+3$, for a total of 10 forecasts. The total number of forecasts grows from 4 to 10. In general, the rolling-origin procedure provides $N(N+1)/2$ forecasts, against N from the fixed-origin. With eight time periods forming the test set, for example, the rolling-origin evaluation supplies 36 forecasts, a multiple of 4.5 times N.

3.3. Analysis of forecasting errors by lead time

In contrast to the fixed-origin evaluation, the rolling out-of-sample evaluation produces multiple forecasts for every lead time but the longest, N. As a result, it permits us to assess the forecasting accuracy of an individual time series at each lead time. Moreover, the errors for a given lead time form a coherent empirical distribution, one we can profitably analyze for further distributional information, such as outliers. Makridakis and Winkler (1989) describe such analysis.

4. Issues in implementing out-of-sample evaluations

In designing an out-of-sample test for an individual time series, the most fundamental choice is how to split the series between fit and test periods. This decision determines the amount of data that will be available to identify and fit a forecasting model and the number of forecasts generated for the out-of-sample evaluation of the model's performance.

4.1. Series splitting rules

In deciding upon the appropriate number of periods N to withhold from the time series, we can be guided by several considerations, the most important of which is the longest-term forecast required. Denote this maximal length forecast by H. Manifestly, N must be at least as large as H.

However, we may wish to increase the length of the test period to insure a certain minimum number of forecasts M at lead time H. We would then set the length of the test period to equal $H + M - 1$ forecasts. If this minimum is $M = 3$, we should design a rolling-origin evaluation with a test period of length $H + 2$. For example, if the longest-term forecast required is a five-year ahead forecast ($H = 5$), we would specify a test period of seven years, thus insuring that the assessment of accuracy in forecasting five years ahead is based on a minimum of three forecasts. We would need a much larger number of forecasts than this to examine a *distribution* of forecast errors, rather than simply measures of average error.

Short time series impose restrictions on the length of the test period, since truncating the data could leave too few observations to fit the model. In this circumstance, we might profit from the efficiency of the rolling-origin procedure and still be able to examine one-step ahead forecast errors without greatly truncating the period of fit.

4.2. Updating versus recalibrating

In the rolling-origin evaluation, each update of the forecasting origin leads to a revision of the forecasting equation.The successive revisions to the forecasting equation may arise simply from the addition of a data point to the fit period, or may arise as well from recalibration (reoptimization) of the smoothing weights as the new data point comes in.

Recalibration is the preferred procedure. Updating without recalibrating imposes an arbitrary handicap on the forecasting method. Recalibration, moreover, desensitizes error measures to events unique to the original fit period. However, recalibration is more computationally intensive than simply updating, and only two of 15 forecasting software packages examined by Tashman and Hoover (2001) recalibrate as they update the forecasting origin.

When it is a (causal) regression model under evaluation, failure to recalibrate transforms a rolling-origin evaluation into a fixed-origin evaluation at one step ahead and into meaningless figures at longer horizons. Without recalibration, the addition of a new data point changes neither the inputs to nor the coefficients of the forecasting equation.

For extrapolative methods, research is lacking on the extent to which recalibration of the smoothing weights across the test period influences the reported absolute and relative accuracy of forecasting methods. Fildes, Hibon, Makridakis and Meade (1998) provide evidence that recalibrating weights in *fitting* an exponential smoothing method improves the out-of-sample accuracy of the method. However, they did not recalibrate the smoothing weights *within the test period* of a rolling-origin evaluation.

Similarly, no one has examined the empirical significance of recalibration in the context of out-of-sample evaluations of regression models. If the model contains dynamic terms, such as a lagged dependent variable or a lagged error, each forecast will adjust as the origin is successively updated. Unless the sample size is small, these effects may be more substantial than the changes that arise from recalibrating the regression coefficients.

4.3. Multiple test periods

Fildes (1992, p.82) observed that replacing a fixed-origin design with a rolling-origin design

removes 'the possibility that the arbitrary choice of time origin might unduly affect the [forecasting accuracy] results' Distinguishing *sensitivity to outliers* in the test period from *sensitivity to the phase of the business cycle*, however, is useful. The test period marks a single calendar interval. Especially for monthly and quarterly data, therefore, it is likely to reflect a single phase of the business cycle or single period of business activity. To attain cyclical diversity in analyzing an individual time series, we should use *multiple test periods*.

Pack (1990) illustrated the virtues of multiple test periods using a retail sales series of 95 consecutive months. For each of three forecasting methods, he designated three distinct test periods, and performed a rolling-origin evaluation for each test period. Table 1 is a portion of his Exhibit 5 (p. 217).

The MAPEs are sensitive to the choice of test period. For lead time 4, for example, forecasting method A earned a MAPE of 3.1 percent over test period 61–71; however, the same measure applied to test period 73–83 yielded a MAPE of 5.8 percent, nearly twice as high. At lead time 1 in test period 85–95, the three methods appear about equally accurate (MAPEs of 3.1%, 3.3% and 3.4%), while, in test period 73–83, method B looks significantly worse (at both lead times) than the others.

Diversifying into multiple test periods seems prudent. Perhaps individual test-period MAPEs should be averaged. The average MAPE for Method A at four-steps head is 4.5 percent, which is the most broad-based indication of this method's expected accuracy in forecasting four months into the future.

Table 1
How the MAPE varies by lead time and test period in comparing three methods

Lead time	Method	Test periods			Average
		61–71	73–83	85–95	
1	A	3.0	4.1	3.1	3.4
	B	3.2	5.0	3.3	3.8
	C	2.3	2.7	3.4	2.8
4	A	3.1	4.6	5.8	4.5
	B	5.3	7.4	6.0	6.2
	C	3.5	3.9	7.0	4.8

Fildes et al. (1998) used multiple test periods, which they called *multiple origins,* to compare the accuracy of five designated extrapolative methods on a batch of monthly telecommunications time series. While they found that one method was uniformly most accurate (across lead time and for every test period), the relative accuracy of three of the other methods was not consistent across test periods.

Schnaars (1986) examined the cyclical sensitivity of forecast error measures by sorting all one year-ahead forecast errors by calendar year (1978–1984). He then compared forecast errors for (a) years in which cyclical turning points occurred and (b) years in which the overall direction of the economy did not change. For almost all of the methods included, he found that one-year-ahead forecasting accuracy was poorer during the years of cyclical turning points.

Using multiple test periods may be particularly beneficial when we are limited by software to fixed-origin evaluations. However, the procedure requires a long time series.

4.4. Rolling windows

In a rolling-origin evaluation, each update of the forecasting origin adds one new observation to the fit period. Alternatively, in some studies, researchers have maintained a fit period (or *sample* or *window*) of constant length. They do this by pruning the oldest observation at each update, much as we would in taking a moving average. The procedure is called a fixed-size, *rolling window* (Swanson and White, 1997) or fixed-size *rolling sample* (Callen, Kwan, Yip and Yuan, 1996).

Why prune the fit period at each update of the forecasting origin? One reason is to 'clean out old data' in an attempt to update model coefficients. Doing so may be unnecessary in common time-series methods, however, because the weighting systems in these methods mitigate the influence of data from the distant past.

Swanson and White (1997) discussed the usefulness of rolling windows in econometric modeling, particularly in determining how econometric models evolve over time to fixed specifications.

For out-of-sample testing, the principal purpose of a rolling window is to level the playing field in a multiperiod comparison of forecasting accuracy. We might analyze whether a particular method's performance deteriorates between an earlier and later test period. The comparison would be confounded if the second fit period were longer than the first.

Swanson and White (1997) further pruned their rolling windows to generate the same frequency of forecasts at each horizon of the *test period.* They wished to ensure equality between the number of one step-ahead forecasts and the number of four step-ahead forecasts. That procedure, however, results in a different calendar fit period for each forecast horizon: the fit period for a four- step-ahead forecast will begin and end three periods earlier than the fit period underlying the one step-ahead forecasts. As a result of the calendar shift, the evidence on how forecasting accuracy of any method deteriorates as the forecasting horizon increases may be confounded.

5. 'Sliding simulations'

Makridakis (1990) extended the rolling-origin design to serve as a process for method selection and estimation. He called this process a *sliding simulation*. (He did not intend the term *simulation* to mean a resampling or Monte Carlo process; he used it rather as a synonym for out-of-sample analysis.) Fildes (1989) also used the procedure — under the name *rolling horizon* — to compare the efficacy of various method-selection rules.

The sliding simulation requires a *three-way division* of the time series. N observations withheld from the time series serve as a test set. The remaining period of fit is subdivided between the first T observations, which represent the *in-sample fit period* and the remaining P observations, $T+1$ to $T+P$, which constitute the *post-sample fit period*.

For each method under consideration, the sliding simulation entails a pair of rolling out-of-sample evaluations. In the first, we optimize the smoothing weights to the *post-sample fit period,* and select a best method for each lead time. The second is performed on the test set, with the traditional purpose of evaluating the accuracy of the forecasts made with this method.

In the same spirit, Weiss and Anderson (1984, p.485) proposed that, for cumulative forecasts, a model be calibrated to minimize a cumulative post-sample error measure.

Makridakis (1990) applied variants of the sliding simulation to a subsample of 111 time series used in the M-competition (Makridakis et al., 1982). For each of three exponential smoothing methods, post-sample forecasting accuracy improved when he calibrated smoothing weights to minimize a post-sample error measure instead of calibrating weights in-sample, as is traditional.

Results reported in the M2-Competition (Makridakis et al., 1993) were not so positive for the sliding simulation process. There, the method chosen as best — from among simple, damped, and linear-trend smoothing — did not systematically outperform any individual smoothing method (Exhibit 3, p.9). In fact, two of the three smoothing methods performed more poorly when calibrated post-sample, the linear trend being the exception.

Fildes (1989) used the sliding simulation to

compare *individual-selection* and *aggregate-selection* rules. When following an individual-selection rule, we identify a best method for each time series in a batch. When following aggregate-selection rule, we apply to every series in the batch the method that works best in the aggregate.

Fildes considered two extrapolative methods, both involving damping of trends and smoothing of outliers. He calibrated each method to a post-sample fit period and chose the better of the two methods based on post-sample fit. He concluded that the extra effort needed in individual rather than aggregate selection was not worth the small potential gain in accuracy for forecasting one month ahead, the most important horizon when forecasting for inventory control. At longer lead times, individual selection has more potential to improve accuracy.

6. Multiple time series: forecasting competitions

For a single time series, desirable characteristics of an out-of-sample test are *adequacy*, enough forecasts at each lead time, and *diversity,* desensitizing forecast error measures to special events and specific phases of business. To achieve these goals with an individual time series, we must use rolling origins and multiple test periods.

Alternatively, we can attain adequacy and diversity by using multiple time series. To promote adequacy, we need to select component series that are homogeneous in some relevant characteristic. For diversity, we should collect time series that are heterogeneous in both nature and calendar time, thus establishing a broad-based track record for a forecasting method.

Diversity was the primary motivation in the early forecasting competitions. Newbold and Granger (1974) amassed 106 economic series, a mixture of monthly and quarterly as well as of micro-level and macro-level data. The M-competition (Makridakis et al., 1982) included 1001 time series, a compendium of annual, quarterly, and monthly as well as firm, industry, macroeconomic, and demographic data. ''Although the [M-competition] sample is not random, efforts were made to select series covering a wide spectrum of possibilities. This included different sources of statistical data and different starting/ending dates.'' (p.113).

In contrast, *selectivity* was the principal objective for Schnaars (1986). Schnaars wished to ''discover how well extrapolations are able to perform on a specific type of data series — annual unit sales by industry — rather than a wide assortment of potentially disparate series.'' (p.72). Selectivity was also an objective for the M2-competition (Makridakis et al., 1993). Of its 29 time series, 23 were monthly firm-level series, chosen to compare the accuracy of designated methods in forecasting for budgeting and capital investment.

The diversity objective for the M-competition returns with the M3-competition (Makridakis and Hibon, 2000), in which the database is enlarged from 1001 to 3003 time series. Again, the authors chose time series to represent data of different periodicities (yearly, quarterly, monthly, and other) and types (micro, industry, macro, finance, demographic, and other). The selection process was essentially downloading a convenience sample of data from the Internet.

The emphasis in a forecasting competition affects both the selection of time series and the implementation of the out-of-sample tests. With the emphasis on *diversity*, the authors of the M-competition and the M3-competition amassed a large collection of heterogeneous time series, but relied on fixed-origin evaluations and a single test period per series to obtain post-sample error measures. In emphasizing *selectivity,* Schnaars and the authors of the M2-competition employed a relatively small number of homogeneous series and used rolling-origin evaluations (Schnaars) and multiple test periods (M2-competition) for diversity.

The reliance on fixed-origin rather than rolling-origin evaluations in the three M-competitions was probably also essential for keeping the forecasting process manageable. In these studies, participants provided forecasts to the researchers, who had withheld the test period data. To implement a rolling-origin evaluation, the participants would have had to be shown the test period data, so that they could successively update the forecasting origins. In contrast, Schnaars (1986) produced his own forecasts.

In principle, a synthesis of the diversity and selectivity strategies is to be recommended. Ideally, a forecasting competition would begin with precisely described groups of time series, in which the series are homogeneous within group but heterogeneous between groups. Randomized selection could then be used to obtain a sample of series from each group.

Armstrong et al. (1998, p. 360) observed that *within-group homogeneity* abets method selection by helping the forecaster to determine which methods are best suited to the specific characteristics of the data. Within-group homogeneity can also be of value for forecasting product hierarchies. At the same time, the forecaster needs *heterogeneity among groups* to draw general inferences about the relative forecasting accuracy of different methods.

In practice, it is difficult to implement a random-sampling design. Time series are multi-attributed: *periodicity* and *type* were the two explicit attributes in the forecasting competitions. However, *type* is really a catchall descriptor, comprising *level of aggregation* (item, product, brand, company, industry, economy), *domain* (financial, marketing, operations), *geographic area* (country, region) and *data characteristics* (seasonal versus nonseasonal, stable versus volatile, trended versus untrended). Another dimension of importance is calendar time interval: Series differ in starting date, ending date, and length, and span different stages of economic cycles and product life cycles. Moreover, the attributes are interdependent in many ways: Seasonality is likely to be most pronounced in quarterly and monthly data, volatility greatest in micro level series, and trends strongest in macroeconomic data.

A perfectly stratified random sample, hence, is not a realistic possibility. Nevertheless, the competitions can be faulted for a lack of formality in the collection of data. Series were collected and *retrospectively* classified by attribute. For this reason alone, tabulations based on 'all series' are suspect.

6.1. Pooled data structure

The use of multiple time series, as in a forecasting competition, creates a pooled data structure: S time series, $s = 1$ to S, and up to $T + N$ time periods per series. Individual time series need not be of equal length nor need they cover the same calendar period. Hence, the periods of fit can vary in both length and calendar interval.

The length of the *test period*, however, is normally fixed for all time series of a given periodicity. For example, Schnaars (1986) withheld the last five years from all the historical series. In the three M-competitions, the test period was specified to be six years, eight quarters and 18 months for annual, quarterly and monthly data respectively.

Fixing the length of the test period is partly a matter of statistical convenience: it simplifies the calculation and presentation of forecast-error averages. Still, considerable obfuscation can result if the forecast error measures are tabulated for an aggregate of series of different periodicities. For the M-competition results, the 'all data' tables combined monthly, quarterly and annual series. Thus, a one step-ahead error figure blended the one-month-ahead, one-quarter-ahead and one-year-ahead forecast errors. The M2-competition and M3-competition have avoided this confusion by separately reporting results for series of different periodicities.

6.2. Pooled averages

To calculate forecast error statistics in a multiseries data set, we can average errors across time series, Σ_s; across lead time, Σ_n; or both, Σ_{sn}. Precisely how the averaging is done can be important.

6.2.1. Choice of error statistic for averaging over series

Much has been written about the choice of forecast-error statistics. A good overview is provided in a series of articles and commentaries in the *International Journal of Forecasting* (Armstrong and Collopy, 1992; Fildes, 1992; Ahlburg et al., 1992).

There are two arithmetic issues. One concerns the choice of *error measure*: Should we be averaging squared errors, percent errors or relative errors? The second deals with the appropriate statistical operator: should we use a median, an arithmetic mean or geometric mean?

The lessons from the research are at least threefold: When averaging over series Σ_s, we should:

1. Avoid scale dependent error measures, such as root mean squared error RMSE or mean absolute deviation MAD. With these, if you rescale the measurement (for example, from one currency to another or from millions of units to thousands of units), you alter the numerical value of the error measure. Moreover, a subset of the time series with large numerical values may dominate the error measures, and that subset would change with the scaling.
2. Use percent error measures instead, such as the absolute percent error (APE), because (for data with a natural zero) they are scale independent. However, the distribution of percent errors can be badly skewed, especially if the series contains values close to zero. In this case using the median absolute percent error (MdAPE) may be preferable to using the MAPE. MdAPE is the principal error statistic used in Vokurka, Flores and Pearce (1996). Still another alternative to the MAPE is the symmetric MAPE (Armstrong, 1985, p. 348), which makes underforecasts and overforecasts of the same percent equal. This statistic is being featured in the M3-competition.
3. Use relative error measures when it is necessary to average over time series that differ in volatility. Collopy and Armstrong proposed a ratio of an absolute error from a designated method to the analogous absolute error from a *naïve* method, which they call the *relative absolute error*, RAE (Collopy and Armstrong, 1992, p.71). They showed that the RAE is not only scale independent but also serves to standardize the component series for degree of change and, hence, degree of forecasting difficulty. Tashman and Kruk (1996) used relative error measures to compare the accuracy of a forecasting method between distinct *groups* of time series. Recommended operators for averaging relative errors are the median (MdRAE) and geometric mean (GMRAE). Fildes (1992, p.84) endorses a variant (calculable only in a rolling-origin evaluation) called the relative geometric root mean square.

By using a single summation Σ_s, we obtain an average error for an individual method at a specific horizon. In reporting the M-competition results, the authors refer an average of absolute percent errors (APEs) as an *average MAPE* (Makridakis et al., 1982, Table 2). For an individual lead time, however, it may be called simply a MAPE, without the preceding *average*, since we are averaging a single APE per time series.

6.2.2. Cumulating over lead times

For cumulative lead time error measures, such as 1–4 quarters or 1–12 months ahead, we

can use a double summation Σ_{sn}, summing individual APEs over both the series and the lead times. Doing so gives equal weight to errors at short and long lead times. Alternatively, we can start with each individual lead time MAPE and then take an average or weighted average across lead times, Σ_n MAPE. The latter properly requires a modifier such as *average* MAPE.

The route taken for calculating cumulative lead time error measures can make a difference. Using the Σ_n approach maintains the distinctiveness of the individual lead times and thus permits flexibility in assigning weights to reflect the relative importance of the individual horizons. Moreover, in a rolling-origin evaluation, the alternative Σ_{sn} approach would assign greater weight for the first lead time, successively smaller weights for each longer lead. If equal weighting of each lead time is desired, the Σ_n MAPE calculation is preferred.

Sensitivity to outliers can be mitigated in both approaches. With the doubly summed measure, we can calculate a median absolute percent error MdAPE or we can employ the median MAPE, as do Tashman and Kruk (1996, Table 7).

For measuring forecast accuracy over a cumulative lead time, Collopy and Armstrong propose the cumulative RAE (Collopy and Armstrong, 1992, p. 75–76).

6.3. Stability of error measures across forecasting origins

Pooling time series and cross-sectional data can create analytical and interpretational difficulties. Normally, as a precondition of pooling, we perform tests to see if the parameters of cross-sectional models are stable over time.

Fildes et al. (1998) used a data set of 263 telecommunications series to examine the stability of error measures across forecasting origins. Their results, similar to those reported earlier from Pack (1990), indicate that the relative accuracy (ranking) of different forecasting methods changed appreciably as the forecasting origin varied. Such instability, they concluded, should discourage forecasters from using a single forecasting origin.

Whether their concern extends to the forecasting competitions is uncertain. Their time series were of equal length and had identical starting and ending dates. The series in the M-competition and in the M3-competition have considerable diversity in length and calendar dates.

Calendar diversity plays the same role in multiseries evaluations that multiple test periods play in individual-series evaluations: Both mitigate the sensitivity of forecast error measures to the phase of the business cycle.

6.4. Method selection rules

In the forecasting competitions, every forecasting method was applied to every time series, whether or not the method was appropriate for the series. For example, Holt's exponential smoothing method was applied to nontrended series, and simple exponential smoothing was applied to trended series. Tashman and Kruk (1996, p. 5) call this *unselective application* and argue that, by fusing appropriate and inappropriate cases, unselective application tends to denigrate a method's expected performance. The alternative is to first screen out those series for which a method is judged inappropriate. Effective screening, however, requires a reliable method-selection rule.

Fildes (1989) articulated the distinction between (a) knowledge of a method's forecasting accuracy after a test and (b) the ability to select a best method in advance. 'Forecasting competitions, such as the M-competition, only offer the forecaster information on the relative accuracy of (methods) A and B, ex post; these show which of the two turned out to be better; but they do not demonstrate how to pick a winner' (1989, p. 1057).

Effective method selection, ex ante, requires effective *method-selection rules*. Among the forecasting competitions, the M3-competition (Makridakis and Hibon, 2000) is the first to examine *automatic forecasting systems*, many of which incorporate method-selection rules. Although the M3-competition summary tables do not include a direct comparison of the category of automatic forecasting systems against the aggregate of single-method procedures, automatic systems were found to be among the methods that give best results for many types of time series.

This result is more promising than prior research would have suggested. Gardner and McKenzie (1988) offered selection rules for choosing among exponential smoothing procedures. Tashman and Kruk (1996) compared the Gardner–McKenzie protocol with two other protocols for method selection. They found that (1) none of method-selection protocols effectively identified an appropriate smoothing procedure for time series that lacked strong trends, (2) the protocols frequently disagreed as to what constituted an appropriate method, and (3) even when they agreed on an appropriate method, following their advice did not ensure improved forecasting accuracy (1996, p. 252).

6.5. Product hierarchies

While the authors of the forecasting competitions have classified time series by periodicity and level of aggregation, they have not incorporated hierarchical data structures. New techniques for demand forecasting have emerged in the past decade that link forecasts for one item (stock keeping unit) to the product class to which the item belongs. For example, Bunn and Vassilopoulis (1993) showed how the seasonal pattern in the product class aggregate could be applied effectively to forecast the seasonality in individual items. Several forecasting programs permit automatic adjustment of forecasts for individual items to reconcile them with the product-class aggregate, thus effectively imposing the structure of the product-class series on the individual components. Doing so is appealing when individual item series are short and irregular.

Testing product hierarchy methodologies should be a high priority for future research.

7. Out-of-sample evaluations in forecasting software

In a review of 13 business-forecasting programs with automatic forecasting features, Tashman and Leach (1991) reported that only six programs included post-sample tests of forecasting accuracy. Of these, moreover, all but two were limited to fixed-origin evaluations on a single series. In the two packages that offered rolling-origin evaluations, the implementation was based on a single series in a single test period and model coefficients that were held fixed rather than recalibrated through the test period. While the authors warned forecasting practitioners to evaluate those methods the software selected automatically, the forecasting software of the early 1990s did not facilitate this process.

Has out-of-sample testing in forecasting software been upgraded during the past decade? Of the 13 programs Tashman and Leach investigated, 10 have ceased to exist. In the remaining three, *Autobox, Forecast Pro* and *SmartForecasts*, the developers have enhanced their post-sample testing options All three now offer rolling out-of-sample evaluations and a variety of forecast error measures.

During the 1990s, the forecasting software market has seen many new entrants. Tashman and Hoover (2001) examined 15 forecasting software programs, of which 9 had their roots in the 1990s. They divided the forecasting packages into four categories: spreadsheet add-ins, forecasting modules of general statistical programs, neural-network programs, and dedicated

business-forecasting programs. The last category included the three aforementioned packages plus *Time Series Expert* and *tsMetrix*.

Tashman and Hoover (2001, Table 4) reported that only one of the three spreadsheet add-ins and one of the four general statistical programs effectively distinguished *within-sample* from *out-of-sample* forecasting accuracy. In contrast, two of the three neural-network packages and three of the five dedicated business-forecasting programs made this distinction effectively.

In my further analysis of the 12 non-neural network programs (software references are at the end of the paper), I found that none of the four general statistical programs and none of the three spreadsheet add-ins offered a rolling out-of-sample evaluation. In addition, most of these include a limited set of error measures: their developers essentially ignore the recent literature on forecast error measurement.

Within the category of dedicated business-forecasting software, *tsMetrix* comes closest to providing the opportunity for systematic out-of-sample tests on individual series. Once the user selects a test period, the program will perform a rolling-origin evaluation, recalibrating the coefficients of the forecasting equations at each update of the origin. This option is available for smoothing, ARIMA, and regression methods. Users can define multiple test periods; however, the program does not integrate error measures across test periods.

The post-sample procedure in *Autobox* matches that in *tsMetrix,* although it is available only for ARIMA modeling. The *Forecast Pro* procedure is also similar, except that it does not recalibrate coefficients with each update of the forecasting origin.

A major growth segment of the forecasting software market has been *demand planning* packages, which incorporate automatic batch forecasting for large product hierarchies. Unfortunately, few reviews and evaluations of this market segment have been published. Developers of demand planning packages have focused on the technology of managing forecasting databases and automating forecasting methods. This focus has come at the expense of transparency regarding how forecasts are made and what forecast errors to expect. Useful out-of-sample tests are seldom included in this type of program.

Forecast Pro, *SmartForecasts* and *Autobox*, which can serve as *forecasting engines* in a demand planning package, are major exceptions. These programs enable users to view average forecast errors made on an entire batch of time series. The programs perform rolling-origin evaluations on individual time series, sorts the forecasting errors by lead time and then report averages of the forecast errors across time series.

8. Summary

For an individual time series, out-of-sample testing of forecasting accuracy is facilitated by use of rolling-origin evaluations. The rolling-origin procedure permits more efficient series-splitting rules, allows for distinct error distributions by lead time, and desensitizes the error measures to special events at any single origin. Applying the procedure across multiple test periods is desirable to mitigate the sensitivity of error measures to single phases of the business cycle. In an implementation of a rolling-origin evaluation, recalibration of the parameters of a forecasting equation can be important in general and is essential in the context of a regression model.

Forecasting software does not always nurture the proper implementation of post-sample tests. Many programs permit only fixed-origin evaluations and report few error measures. Those that offer rolling-origin evaluations often restrict them to certain methods, usually extrapolative.

Few demand planning packages incorporate useful out-of-sample evaluations.

Forecasting competitions would be more generalizable if based upon precisely described groups of time series, in which the series were homogeneous within group and heterogeneous between groups. Even a large collection of time series does not automatically ensure diversity of forecasting situations, especially if calendar dates are more or less coterminous. Measures based on a single cross-section can be unstable over time. Error statistics that are calculated by applying every method to every time series may give misleading results. Evaluating methods used in forecasting product hierarchies remain an important avenue for further research.

References

Ahlburg, D. A., Chatfield, C., Taylor, S. J., Thompson, P. A., Winkler, R. L., Murphy, A. H., Collopy, F., & Fildes, R. (1992). A commentary on error measures. *International Journal of Forecasting 8*, 99–111.

Armstrong, J. S. (1985). *Long-range forecasting*, Wiley-Interscience, New York.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalising about forecasting methods: empirical comparisons. *International Journal of Forecasting 8*, 69–80.

Armstrong, J. S., & Grohman, M. C. (1972). A comparative study of methods for long-range market forecasting. *Management Science 19*, 211–221.

Armstrong, J. S., Koehler, A. B., Fildes, R., Hibon, M., Makridakis, S., & Meade, N. (1998). Commentaries on 'Generalizing about univariate forecasting methods: further empirical evidence'. *International Journal of Forecasting 14*, 359–366.

Bartolomei, S. M., & Sweet, A. L. (1989). A note on a comparison of exponential smoothing methods for forecasting seasonal series. *International Journal of Forecasting 5*, 111–116.

Bunn, D. W., & Vassilopoulis, A. I. (1993). Using group seasonal indices in multi-item short-term forecasting. *International Journal of Forecasting 9*, 517–526.

Callen, J. L., Kwan, C. C. Y., Yip, P. C. Y., & Yuan, Y. (1996). Neural network forecasting of quarterly accounting earnings. *International Journal of Forecasting 12*, 475–482.

Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics 11*, 121–135.

Collopy, F., & Armstrong, J. S. (1992). Rule-based forecasting. *Management Science 38*, 1394–1414.

Fildes, R. (1989). Evaluation of aggregate versus individual forecast method selection rules. *Management Science 35*, 1056–1065.

Fildes, R. (1992). The evaluation of extrapolative forecasting methods. *International Journal of Forecasting 8*, 81–98.

Fildes, R., Hibon, M., Makridakis, S., & Meade, N. (1998). Generalising about univariate forecasting methods: further empirical evidence. *International Journal of Forecasting 14*, 339–358.

Fildes, R., & Makridakis, S. (1995). The impact of empirical accuracy studies on time series analysis and forecasting. *International Statistical Review 63*, 289–308.

Gardner, Jr. E. S., & McKenzie, E. (1988). Model identification in exponential smoothing. *Journal of the Operational Research Society 3*, 863–867.

Makridakis, S. (1990). Sliding simulation: a new approach to time series forecasting. *Management Science 36*, 505–512.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, P., & Winkler, R. (1982). The accuracy of extrapolation (time series) methods: results of a forecasting competition. *Journal of Forecasting 1*, 111–153.

Makridakis, S., Chatfield, C., Hibon, M., Lawrence, M., Mills, T., Ord, J. K., & Simmons, L. F. (1993). The M2 competition: a real life judgmentally-based forecasting study. *International Journal of Forecasting 9*, 5–29.

Makridakis, S., & Hibon, M. (2000). The M3-competition: results, conclusions and implications. *International Journal of Forecasting 16*, 451–476.

Makridakis, S., & Winkler, R. L. (1989). Sampling distribution of post-sample forecasting errors. *Applied Statistics 38*, 331–342.

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

Pack, D. J. (1990). In defense of ARIMA modeling. *International Journal of Forecasting 6*, 211–218.

Pant, P. N., & Starbuck, W. H. (1990). Innocents in the forest: forecasting and research methods. *Journal of Management 16*, 433–460.

Schnaars, S. P. (1986). A comparison of extrapolation procedures on yearly sales forecasts. *International Journal of Forecasting 2*, 71–85.

Swanson, N. R., & White, H. (1997). Forecasting economic time series using flexible versus fixed specification and linear versus nonlinear econometric models. *International Journal of Forecasting 13*, 439–461.

Tashman, L.J., and Hoover, J.H. (2001). Diffusion of forecasting principles: an assessment of forecasting software programs. In J. Scott Armstrong, Principles of forecasting: a handbook for researchers and practitioners. Norwell, MA: Kluwer Academic Publishers (in press).

Tashman, L. J., & Kruk, J. M. (1996). The use of protocols to select exponential smoothing methods: a reconsideration of forecasting competitions. *International Journal of Forecasting 12*, 235–253.

Tashman, L. J., & Leach, M. L. (1991). Automatic forecasting software: a survey and evaluation. *International Journal of Forecasting 7*, 209–230.

Vokurka, R. J., Flores, B. E., & Pearce, S. (1996). Automatic feature identification and graphical support in rule-based forecasting: a comparison. *International Journal of Forecasting 12*, 495–512.

Weiss, A. A., & Anderson, A. P. (1984). Estimating time series models using relevant forecast evaluation criteria. *Journal of the Royal Statistical Society (A) 147*, 484–487.

Software References

Autobox for Windows, Version 5 (1999). AFS Inc., PO Box 563, Hatboro, PA 19040

CB Predictor: forecasting software for Microsoft Excel, Version 1 (1999). Decisioneering, Inc., 1515 Arapahoe Street, Suite 1330, Denver, CO 80202

Forecast Pro, Version 4 (1999) and *Forecast Pro Unlimited* (1999). Business Forecast Systems, Inc., 68 Leonard Street, Belmont, MA. 02178

SAS/ETS, Version 7 (1997–99). SAS Institute, Inc., SAS Campus Drive, Cary, NC 27513-2414

Insight.xla: business analysis software for Microsoft Excel, Version 1 (1998). Sam Savage, Duxbury Press.

Minitab, Release 11 (1997). Minitab, Inc., 3081 Enterprise Drive, State College, PA 16801-3008

SmartForecasts for Windows, Version 5 (1999). Smart Software, Inc., 4 Hill Road, Belmont, MA 02178

Soritec for Windows, Version 1 (1998). Full Information Software, Inc., 6417 Loisdale Road, Suite 200, Springfield, VA, 2215-1811

SPSS Trends, Version 8 for Windows (1998). SPSS, Inc., 444 North Michigan Avenue, Chicago, IL 60611

Time Series Expert, Version 2.31 (1998). Statistical Institute of the Free University of Brussels (Contact person: Professor Guy Melard, gmelard@ulb.ac.be)

tsMetrix, Version 2 (1997). RER, Inc.,12520 High Bluff Drive, Suite 220, San Diego, CA 92130

Biography: Len TASHMAN is on the faculty of the School of Business Administration of the University of Vermont. He has contributed articles to several forecasting journals and has published many evaluations of forecasting software.