Review of Guidelines for the Use of Combined Forecasts

Lilian M. de Menezes
Goldsmiths College, University of London

Derek W. Bunn and James W. Taylor
London Business School

European Journal of Operational Research, 2000, Vol. 120, pp. 190-204.

Address for Correspondence:
James W. Taylor
Saïd Business School
University of Oxford
Park End Street
Oxford OX1 1HP, UK
Email: james.taylor@sbs.ox.ac.uk

Review of Guidelines for the Use of Combined Forecasts

Abstract

A large literature has evolved in the thirty years since the seminal work on combining forecasts. Despite this, when evaluating performance we only look at measures of accuracy and thus ignore most of the rigour of time series analysis. Furthermore, the output from a combination of forecasts is just a single point estimate which is insufficient for business planning models which take explicit account of risk and uncertainty. In this paper, we review evidence on the performance of different combining methods with the aim of providing practical guidelines based on three properties of the forecast errors: variance, asymmetry and serial correlation. The evidence indicates that using different criteria leads to distinct preferences, and that the properties of the individual forecast errors can strongly influence the characteristics of the combination's errors. We show that a practical approach to combining also requires a degree of judgement on the attributes of error specification.

*Keywords:*Forecasting; Combinations; Error Specification; Guidelines.

1. Introduction

Combining forecasts is a well-established procedure for improving forecasting accuracy which takes advantage of the availability of both multiple information and computing resources for data-intensive forecasting (Bunn, 1989). Generally, when one evaluates the results of a composite forecast, one only looks at measures of accuracy, for example, mean squared error (MSE). This is a remarkably different procedure from that of analysing results within any other forecasting technique where the forecast error pattern is also considered. It seems that much of the rigour, which is necessary to generate the individual models' forecasts, is forgotten when evaluating the combination. Furthermore, the output from a combination of forecasts is still just a single point estimate, however, this is no longer sufficient for business planning models (e.g. capital budgeting, resource allocation, policy monitoring) that need to take explicit account of risk and uncertainty. Indeed the increasing use of risk analysis (Cooper and Chapman, 1987) places an additional requirement for high quality estimates of predictive distributions. It seems a little surprising that even Bayesian combining methods (e.g. Clemen and Winkler, 1993), which use the distributional properties of the individual forecasts to construct the combination, do not consider the predictive distribution of the resultant composite forecast.

Despite a large literature on combining forecasts, the choice of which method to use is not obvious. Even under a single accuracy criterion, such as mean squared error, guidelines concerning the "best performance" are not straightforward. When additional criteria are considered, the development of practical rules and the interactive use of judgement become harder to formalise. The main aim of this paper is to develop practical guidelines for combining forecasts where the criteria are error variance, distribution asymmetry and serial correlation. Recent research has provided theoretical and empirical evidence regarding the latter two alternative criteria. This paper draws practical guidelines from these studies, and contrasts them with guidelines based purely on an accuracy perspective. The resultant guidelines serve as a set of decision rules, which may be applied not only to select an appropriate combining method but also for switching from one combining method to another, or refining the forecasting model, as new signs develop.

In the decade since Clemen's (1989) extensive bibliographical review of forecast combining, the literature has grown substantially. A further contribution of our work is to provide an update of this review from a practical perspective. In order to produce useful insight, we focus on several of the most widely used combining rules. Three of these are based on Bates and Granger's (1969) initial proposal and so this paper also provides an overview of how far guidelines have developed over thirty years of combining forecasts. First, the combining methods, which we consider in the present study are discussed. We then review the literature to establish the current perspective on their relative performance. This survey reveals that combined forecasts have been evaluated solely on the basis of accuracy. We summarise practical guidelines for combining from an accuracy perspective and then, in the subsequent two sections, we concentrate on the alternative criteria of skewness and serial correlation, for which we provide illustrative examples. Finally, a set of guidelines, possible extensions and further practical implications are discussed.

2. Combining methods

The years since Bates and Granger's seminal article have seen the development of many interesting combining methods. The methods now available to the forecaster range from the robust simple average to the far more theoretically complex, such as state-space methods that attempt to model non-stationarity in the combining weights. An example of one of the more recent developments is the work of Donaldson and Kamstra (1996) on estimating the combining form using artificial neural networks, which is another data-intensive, theory-sparse forecasting approach (Bunn, 1996).

Our analysis, however, concentrates on seven well-established methods that were selected to be good representatives of varying degrees of sophistication. All the methods adopt the linear formulation whereby a vector, f, of n forecasts are combined via a linear weighting vector, w, as $f_c = w'f$.

(1) *Simple average*: This has the virtues of impartiality, robustness and a good "track-record" in economic and business forecasting. It has been consistently the choice of many researchers (see Clemen's review of 1989).

(2) *Outperformance* (Bunn, 1975): This approach develops the forecast combination as $f_c = p'f$ where p is a simplex of probabilities which can be assessed and revised in a Bayesian manner. Each individual weight is interpreted as the probability that its respective forecast will perform the best (in the smallest absolute error sense) on the next occasion. Each probability is estimated as the fraction of occurrences in which its respective forecasting model has performed the best in the past. It is a robust nonparametric method of achieving differential weights with intuitive meaning which performs well when there is relatively little past data and/or when the decision maker wishes to incorporate expert judgement into the combining weights (Bunn, 1985).

(3) *Optimal* (Bates and Granger, 1969): Here the linear weights are calculated to minimise the error variance of the combination (assuming unbiasedness for each individual forecast). Specifically, the vector of combining weights, w, is determined according to the formula

$$w = \frac{S^{-1}e}{e'S^{-1}e} \tag{1}$$

where e is the ($n \times 1$) unit vector and S is the ($n \times n$) covariance matrix of forecast errors. Granger and Ramanathan (1984) showed that the method is equivalent to a least squares regression in which the constant is suppressed and the weights are constrained to sum to one. The problem with this optimising approach is that it requires S to be properly estimated. In practice, *S* is often not stationary, in which case it is estimated on the basis of a short history of forecasts and thus the method becomes an *adaptive* approach to combining forecasts.

(4) *Optimal (adaptive) with independence assumption*: The estimate of *S* in (1) is restricted to be diagonal, comprising just the individual forecast error variances.

(5) *Optimal (adaptive) with restricted weights*: In this case the optimal formula has the additional restriction that no individual weight can be outside the interval [0,1].

(6) *Regression*: In this method the constituent forecasts are used as regressors in an ordinary least squares (OLS) regression with the inclusion of a constant. Granger and Ramanathan (1984) argued that this has the advantage over the popular optimal method that an unbiased combined forecast is produced regardless of whether the constituent forecasts are biased.

(7) *Regression with restricted weights*: Here the least squares regression is performed with the inclusion of a constant but the weights are constrained to sum to one.

3. The current state of preferences: controversies over relative performance

In thirty years of combining forecasts, assessments of the relative performance of various combinations have generally been made under an accuracy criterion, mostly expressed in terms of MSE. In this section we consider empirical and simulation studies, which were specifically designed to investigate the relative accuracy of combining methods. They serve as the basis for some initial guidelines and practical observations, which we discuss in section 4.

3.1. Early empirical evidence

Newbold and Granger (1974) analyzed combinations of three forecasts for 80 time series using different estimates of the optimal weights, though restricting the weights to the interval [0,1]. They concluded that formulations which assume independence between the individual forecast errors perform considerably better than those attempting to account for correlation. They also found that, regardless of the combining method employed, a small improvement in forecast accuracy results from the addition of a third forecast.

A large-scale forecasting competition (1001 time-series), known as the *M-Competition*, was reported by Makridakis *et al*. (1982). Amongst several extrapolation techniques, two combinations of six forecasts were tested: the simple average and a weighted average based on the covariance matrix. Makridakis *et al*. found that the simple average produced the better performance, and indeed outperformed the individual methods included in the average, although Gardner (1983) showed that the ranking of the forecast methods depended somewhat on the choice of error measure used. Subsequently, Winkler and Makridakis (1983) used the 1001 series from the M-Competition and examined five weighting procedures for combining various numbers of forecasts from among ten different methods. Their results confirmed Newbold and Granger's (1974) conclusions regarding the preference for the independence assumption and a relatively small number of individual forecasts.

In a simulation study, Bunn (1985) addressed the relative performance of combining methods as a function of the three statistical properties of the individual forecast errors: variance ratios, correlation coefficient and sample size. He reported on combinations of two forecasts generated from six combining methods, namely: equal weights, optimal, optimal with independence assumption, outperformance, Bayesian probabilities and quasi-Bayes probabilities. Bunn found an overall robustness of optimal with independence assumption, outperformance and quasi-Bayes probabilities. He found that forecasting accuracy was heavily dependent on sample size with the outperformance method dominating for smaller sample sizes (≤6); the optimal procedure with independence assumption was found to be the most efficient for sample sizes varying from 7 to 25, whenever the variance ratio was significantly different from 1 and the data was well behaved; and not surprisingly, the optimal procedure dominated over larger samples, provided well behaved data.

Our view is that there is considerable disagreement and contradiction regarding the "best choice" of combining rule. Using a MSE criterion, Schnaars (1986a) compared performances of seven extrapolation models and three combinations of models over nearly 1500 forecasts of annual sales series. He concluded that combinations generally outperformed the individual models and equal weights were preferred. On the other hand, in a similar study, where most series had less than 30 observations, the same author (Schnaars, 1986b) found that a weighted average performed significantly better than a simple average.

3.2. Evidence on regression-based combining and the issue of bias

Granger and Ramanathan (1984) framed the combination of forecasts as an unrestricted least squares regression with an intercept, and showed that if the individual forecasts are biased, the method will be superior to Bates and Granger's (1969) optimal method. Recently, MacDonald and Marsh (1994) reported that the presence of substantial biases in constituent forecasts led them to use OLS regression to combine exchange rate forecasts. The superiority of the regression method was supported by the work of Guerard (1987) and Holmen (1987). By contrast, Mills and Stephenson (1985), Clemen (1986), Holden and Peel (1986) and Lobo (1991) provided empirical evidence favouring the optimal approach over OLS regression.

As the optimal method can be viewed as a least squares regression with the restrictions of no intercept and slope coefficients summing to one, it is clear that the combined within-sample MSE will be higher for the optimal method than for unrestricted regression. Therefore, an unrestricted regression-based approach would appear to be the natural choice. However, our advice is that some care should be taken. Since the unrestricted regression includes one more forecast among those to be combined, the unconditional mean of the variable being forecast, the variable must be either stationary or made stationary. Another issue, which we discuss more fully in section 5, is that the forecast errors resulting from unrestricted least squares combining are likely to be serially correlated (Diebold, 1988). A third problem with the unconstrained regression approach is multicollinearity which is likely to be an issue as constituent forecasts are often correlated.

Following Granger and Ramanathan's interpretation, Holden *et al.* (1990) concluded that a reasonable approach to combining is to include a constant in the regression and to restrict the weights on the forecasts to sum to one, thus excluding the unconditional mean of the series that is implicitly included in an unrestricted regression approach. The authors argued that this is appealing as for many forecasting situations the unconditional mean cannot be expected to contribute to a combination. They also claim that this procedure is equivalent to debiasing the original forecasts and then applying an optimal formulation. Unfortunately, this is not strictly true because although the constant will correct for *location* bias, it will be insufficient to ensure unbiasedness if the individual forecasts suffer from *scale* bias.

The empirical work of Gunter (1992) and Aksu and Gunter (1992) compared the accuracy of a wide range of restricted least squares combining procedures plus the simple average. The methods that they examined used various mixtures of the following restrictions: constraining weights to sum to one, suppressing the intercept and constraining weights and intercept to be non-negative. They found that restricting weights to be non-negative was as robust and accurate as the simple average, and that both almost always outperform least squares without constraints and least squares with the restriction that weights sum to one but may be negative.

3.3. Evidence on simple average combining and the issue of stability

As it is widely accepted that only "good" forecasts should be included in a combination, strong differences in forecast error variances between the individual forecasts are not to be expected. In such circumstances, it seems unlikely that a weighted average will outperform a simple average combination.

As we discussed earlier, the M-Competition of Makridakis *et al.* (1982) gave strong support for the simple average. More recently, Makridakis *et al.* (1993) reported on their M2-Competition, which aimed at determining post-sample forecasting accuracy of 10 extrapolative methods, including 5 human forecasters. The study considered different time series, which characterised specific contexts where the forecasters would tend to incorporate additional information into their predictions. Among their conclusions, the authors recommended a simple average of smoothing methods on grounds of accuracy and efficiency. Unfortunately, they did not consider other combining approaches. Another opportunity to evaluate the relative success of automatic combining methods will be provided by the M3-Competition which is currently being organised and involves 3003 series.

Many other applications have favoured equal weights (e.g. Bessler and Brandt, 1981; Kang, 1986; Clemen and Winkler, 1987), the argument being that it will perform best or nearly best. These conclusions made Clemen (1989), in his review of over 209 articles, ask "why does the simple average work so well, and under what conditions do other specific methods work better?". Since then, there have been several significant theoretical contributions to the debate. Palm and Zellner (1992) pointed out that the sampling uncertainty in the estimated weights is often overlooked when the theoretical MSE of a weighted average is evaluated. Gunter (1990) identified analytically the conditions under which the simple average outperforms the regression and optimal methods. He considers the range of possible values of the true weight vector for which the simple average will outperform the alternatives, based on an expected MSE criterion. Our view is that, although interesting and intuitive results emerge, it is difficult to draw guidelines for the practical situation where the true weight vector is obviously unknown.

A possible answer for the success of the simple average, as indicated by comments from several authors (e.g. Holden and Peel, 1989), may rely on the potential for unstable weights, which often result from unsystematic changes over time in the variance-covariance matrix of the individual forecast errors. Under these circumstances, a simple average, although having non-optimal weights, may still give rise to better results than time-varying weights. An analogous argument can be developed with respect to the amount of data available to estimate the combining weights.

The issue of structural change has been approached in the literature by the use of time-varying (Diebold and Pauly, 1987; LeSage and Magura, 1992), or more adaptive combining procedures (Bates and Granger, 1969), or simply ignoring part of past data whenever it becomes irrelevant for forecasting purposes (Chatfield, 1988). Miller *et al.* (1992) reported the results of a simulation study investigating the effects of non-stationarity on a range of combining methods. They simulated the non-stationarity by means of a shift in the underlying process generating the forecast errors. Not surprisingly, the simple methods did much better during and immediately after structural changes, however, overall their study indicated no great advantage for the simple average approach.

3.4. Evidence from studies into switching methods

Schmittlein *et al.* (1990) focused on potential policies for switching between combining models. Their work provides useful insight for combining. Using simulated data, they analyzed performance regarding: the size of the forecast history used for estimation, the accuracy and the correlation of the individual forecasts. Combinations of two forecasts were considered using the following methods: equal weights, optimal and optimal with independence assumption. The obtained results (supported by the standard MSE criterion) indicated that for equal weights to be the best alternative, the individual forecasts' accuracy must be similar and large positive correlations must not occur ($\rho<0.5$). They found that the optimal method with independence assumption dominated, when the accuracies were unequal and the absolute value of the correlation was small ($-0.3<\rho<0.4$), and also when the accuracies are similar and correlation was large and positive. The optimal method should be used whenever accuracies are unequal and the absolute value of the correlation is large ($\rho < -0.3$, $\rho > 0.4$). They found that as the forecast history develops, the availability of data will generally support more sophisticated procedures. Hence, the "best" combining method could change upon the arrival of more information concerning the individual forecasts. Finally, they argued that in practice the relative performance of equal weights is likely to improve, since it does not rely on estimates of parameters, which are then subjected to instability.

Deutsch *et al.* (1994) introduced combining models with changing weights derived from switching regression models or from smooth transition regression models. They considered models in which the weights are allowed to change immediately or gradually when there is a change in regime. The switching regime was estimated using two alternative approaches. The first used the lagged forecast errors from the constituent forecasts whilst the second based the regime on a relevant economic variable.

3.5. Evidence on the use of judgement in combining

Flores and White (1989) evaluated subjective versus objective combinations of forecasts. Their experiment used 93 undergraduate students as the forecasters and two different types of time series. They claimed that subjective combinations were as accurate or even more accurate than objective combining methods (simple average or optimal methods). However, they agreed with Newbold and Granger (1974) by recommending that no more than four forecasts should be combined. Indeed, in spite of controversies over the "best" combining method, there has been considerable agreement with respect to the number of individual forecasts to be included in the combination (maximum of 4). However, in section 5 we present motivation for the inclusion of a much larger number of individual forecasts in the combination.

Collopy and Armstrong (1992) presented a rule-based expert systems approach to integrating judgement and quantitative approaches to forecasting. Knowledge elicitation was achieved through a survey of forecasting experts' routines, which revealed an overall preference for combining forecasts instead of developing complex models. Their final modelling stage consists of a combination, where different weights are assigned to the forecasts based on accuracy over the hold-out data period. The combining rule assumes independence between the individual forecast errors, and, in circumstances of high uncertainty, it reverts to equal weights.

4. Practical guidelines from a minimum variance perspective

Under stable conditions with relatively well-behaved data, theoretical and simulation research has shown that the relative performance of combining forecast methods depends upon: the error variance ratios of the forecasts, the correlation between forecast errors and the sample of past data used for estimation. A review of the past thirty years of combining, and recent simulation and case-study results (de Menezes, 1993; de Menezes and Bunn, 1993), have led us to the following practical guidelines for combining forecasts based on a minimum forecast error variance criterion:

(1) Over small samples, use the outperformance as it is a simple method that takes advantage of the dissimilarity in forecast error variances.

(2) Over medium samples with low correlation, use an optimal method with independence assumption.

(3) Over large samples, use an optimal or a restricted regression-based model.

(4) However, if error variances are similar and positive correlation is weak ($\rho<0.5$) or unstable, the simple average is advocated on the basis of simplicity and efficiency.

In the vast majority of cases, preference towards a particular combining method has been based on an accuracy criterion. One exception to this is the work of Reeves and Lawrence (1982, 1991), who considered a multiple objective combining framework where the criteria are forecast accuracy and the ability to correctly predict the direction of change. However, none of the literature has considered the selection of a combining approach under a range of criteria concerned with the specification of the combined forecast error distribution. In the next two sections of this paper we broaden the evaluation criteria by considering the skewness and the serial correlation of the forecast errors.

5. Skewness of the forecast error distribution

We now focus on the shape of the combined forecast error distribution. This is of practical value because different shapes imply distinct attitudes towards risk and uncertainty (see the literature on stochastic dominance, e.g. Bunn (ch.4, 1984) and Levy (1992)). From a practical perspective, there are two main issues of interest:

(1) If different combining rules lead to differently shaped forecast error distributions, then the decision-maker's attitude to risk becomes an additional factor in selecting the combining method.

(2) If combining different individual forecasts results in diverse shapes of error distributions, then the choice of which specific set of forecasts to be included in the combination will also be affected by this criterion.

Our consideration of the shape of the forecast error distribution focuses on the departure of the distribution from the common assumption of normality. Although a more complete analysis would account for kurtosis, we consider only skewness as this feature of the distribution can be perceived even in very small samples (Ramsey and Ramsey, 1990; D'Agostino *et al.*, 1990), whilst inference on the coefficient of kurtosis would require a minimum of several hundred observations. In this section we first present practical guidelines for combining forecasts using a skewness criterion and then provide an illustrative example.

5.1. Practical guidelines from a skewness perspective

Whilst there have been contributions in the literature to estimating the predictive distribution in a combining context (Granger *et al.*, 1989; Taylor and Bunn, 1996, 1998), the only studies that we are aware of, which offer guidelines for combining forecasts based on the shape of the distribution, are those of de Menezes and Bunn (1993, 1998). These case-study and simulation results indicate that, in a context of asymmetric individual forecast error distributions, a skewness criterion suggests the following guidelines for combining forecasts:

(1) Over small samples, use an outperformance method.

(2) Over medium and large samples use:

- An adaptive (optimal) method with independence assumption whenever positive correlation is not large.
- An adaptive (optimal) method with correlation estimation and restricted weights in the presence of significant positive correlation.

(3) When selecting individual forecasts for the combination, consider different (preferably "balancing") shapes.

(4) When combining, include as many different forecasts as available, since they may provide little additional information in variance terms, but are likely to improve the shape of the resulting forecast error distribution, and thus reduce risk.

(5) When analysing results, estimate not only the mean forecast error but also measures that indicate asymmetry (mode, median, etc.).

(6) Whenever a simple average is chosen (e.g. for simplicity or due to an unstable covariance matrix of the forecast errors), be aware that skewness may not be diluted through combining.

Although these guidelines suggest an evolution from a less sophisticated (outperformance) to a more complex (optimal) method, they differ substantially from those supported by a strictly minimum error variance criterion. The differences are expressed not only in terms of the expectations regarding the simple average, but also in the number of forecasts to be included in the model, which indeed can be larger than the maximum traditionally advocated in the literature. Case-study results (de Menezes and Bunn, 1993) indicate that the above suggestions may be extended to kurtosis as a possible measure of increasing risk, although a detailed analysis could only be carried out for large samples.

5.2 An illustrative example: UK Daily Electricity Load

For a number of years the engineers responsible for scheduling electricity generating plants in England and Wales have consulted three separate forecasts that are produced in three hourly intervals throughout the day. Thus, the National Control engineers have made their own forecast as a judgemental synthesis (Bunn, 1987) of: AREA - a synthesis of the regionally produced, area based forecasts; DFS - a weather based regression model from the demand forecasting section; HEUR - a heuristic load curve based approach of their own.

Here, we consider a re-evaluation of a set of 246 observations for consecutive evening peak forecasts from the three methods. First, we examined the variance-covariance structure of the individual forecast errors. Although, the correlation appeared to increase with the forecast history, which would have favoured an optimal method of combination, it was very unstable (as illustrated in Table 1). Thus, we should revert to simpler combining methods. In this example we consider the simple average.

******************* TABLE 1 APPROXIMATELY HERE **********************

The individual forecast error distributions are summarised in Table 2 and indicate that non-normality is a serious concern. The skewness and kurtosis values should ideally be around 0 and 3 respectively for normality. The Bowman Shenton (Bera and Jarque, 1980) is an omnibus test for normality (see Newbold, 1995, page 412). We notice that the forecast with the smallest MSE, DFS, which would generally be given the most weight in a combination, has a significantly skewed and long-tailed error distribution.

******************* TABLE 2 APPROXIMATELY HERE **********************

We focus upon simple averages of all three, and of pairs of the original forecasts (Table 3). All averages have led to more accurate predictions, compared to the best individual forecast (DFS). However, we notice that high levels of skewness and kurtosis of the DFS persisted into combinations. The critical value (5% level) for skewness is about 0.3. We can see that of the individual forecasts, only DFS produced significantly skewed errors. Yet, substantial skewness was transmitted into all averages that contained DFS. Therefore, reliance on the ability of averaging to dilute non-normality seems rather dangerous.

******************* TABLE 3 APPROXIMATELY HERE **********************

In practice, the asymmetric costs of error in electric load forecasting encourage over-forecasting. If we disregard bias, we might conclude that the better shape of the combined pair AREA & HEUR, is preferable to the slightly more accurate average of all three, especially if confidence intervals are to be estimated.

6. Serial correlation in the forecast errors

Although several of the empirical studies on combining forecasts have evidenced some degree of serial correlation, few authors have yet addressed the issue. The implication of the presence of serial correlation in the combined prediction errors is that the estimates of the combining weights are inefficient and their associated standard deviations are inconsistent. Hence, the combined forecasts may no longer be the best unbiased linear combination. In this section we consider the case of serial correlation for regression-based combination methods, we then review the recent work relating co-integration to combining forecasts, and finally we present practical guidelines for combining using a serial correlation perspective followed by an illustrative example.

6.1. Serial correlation from regression-based combining models

A forecaster may believe that well behaved forecast errors (e.g. white noise) are sufficient to guarantee uncorrelated combined disturbances. Whilst this is true for the optimal method, Diebold (1988) showed analytically that unrestricted OLS regression combinations are expected to give rise to serially correlated residuals even if the individual forecasts have serially uncorrelated errors. However, if the regression is constrained so that the combining weights sum to unity, either with or without the presence of a constant, then the combined forecasts' residuals will be serially uncorrelated. This is the case for the optimal method which is equivalent to regression with no intercept and weights summing to one. Diebold went on to show that if individual forecasts are weakly inefficient, exhibiting a small degree of serial correlation, the combined forecast errors will be even more serially correlated. Of course, ideally, serial dependence should be absent in the individual forecasts. However, in practice, some serial correlation may occur. For example, error processes with low time-dependent coefficients are difficult to identify. Also, one often relies on externally produced forecasts, which may be published by different sources that periodically update previous estimates, and thus serial correlation emerges from the forecasting process itself.

Diebold recommended the inclusion of a model of the disturbance process in the combination, however, in practice, stable conditions may not prevail and it may be difficult to adequately estimate the error process. Coulson and Robins (1993) investigated the implications of including lags into combinations, by focusing on the specific case of combining two forecasts whose combination error follows an AR(1) process. They concluded that a parsimonious method for incorporating the dynamics is achieved by using a lagged dependent variable, with no lagged forecasts. They reported improvements in post-sample forecasting, although they failed to report the specification of the resultant forecast errors.

6.2. Integrated processes: a special case

We have not yet addressed the specific case in which the variable to be forecast is integrated. (When a series must be differenced d times before achieving stationarity it is said to be integrated of order d, and denoted $I(d)$.) In fact, integrated series are quite common, particularly when dealing with financial and economic variables. The problem with using integrated variables in an unrestricted regression is that the R^2 and t-values may suggest a good model but instead the regression may be spurious. The key issue in identifying a spurious regression is the behaviour of the residuals in terms of stationarity. This can initially be assessed graphically or on the basis of the Durbin-Watson statistic (*DW*). If the process is non-stationary, then the *DW* statistic tends to zero and the model is misspecified. A stationary process leads to a *DW* value that is significantly different from zero, indicating that the variable being forecast and the actual forecasts are cointegrated and the estimated weights are then super-consistent. (In general, linear combinations of $I(d)$ variables are also $I(d)$, but if there exists one or more linear combinations that are integrated of lower order, then the variables are said to be cointegrated.)

Hallman and Kamstra (1989) were the first to note that cointegration is an issue when combining forecasts of an integrated variable. When a forecasted variable y_t is $I(1)$, any reasonable forecast f_t should be cointegrated with it, with co-integrating vector $(1,-1)'$. If not, the forecast error $(y_t - f_t)$ will not be stationary and the series and its forecast will drift increasingly apart over time. (Indeed, Holly and Tebbutt (1993) made the point that, before combining, it is important to ensure that if y_t is $I(1)$, then each individual forecast is cointegrated with y_t). If the dependent variable in a regression is cointegrated with the independent variables, then the estimated coefficients will have reasonable properties but will in theory be inefficient, unless account is taken of the cointegration. Engle and Granger (1987) indicated that some form of the error-correction model is appropriate for integrated variables and, with this in mind, Hallman and Kamstra proposed a new combining form.

Writing the ith one step-ahead forecast at time t-1 as $f^i_{t\text{-}1,1}$, Hallman and Kamstra pointed out that if $(y_t - f^i_{t\text{-}1,1})$ and $(y_{t\text{-}1} - y_t)$ are both $I(0)$, then so is $(f^i_{t\text{-}1,1} - y_{t\text{-}1})$. In view of this, they developed a model for the forecast difference of y_t in terms of the new variables $(f^i_{t\text{-}1,1} - y_{t\text{-}1})$. This can be interpreted as explaining the change in y_t as a linear combination of forecasts of the change. This is equivalent to a model for y_t in terms of the forecasts and a lag $y_{t\text{-}1}$ of the series itself, with coefficients constrained to sum to one and the presence of a constant term. Since all the variables in the differenced formulation are $I(0)$, the t-values of the regressors, including the constant, can be used to decide whether to retain all the forecasts in the combination. Coulson and Robins (1993) also concluded that, rather than a forecast of the level, the forecast of the change should be made, based on a linear combination of forecasts of the change.

Cointegration is an issue if unrestricted regression is used for combining. We advise two possible courses of action when using unrestricted regression to combine forecasts of an integrated variable:

(1) Adopt the traditional time series approach and model forecasts of the corresponding stationary change.

(2) Follow an econometric (error-correction) approach supported by the theory of cointegration and include the corresponding lag of the series being forecast in the combining regression model.

6.3. Practical guidelines from a serial correlation perspective

The practical implication of Diebold's (1988) theoretical work is that using unrestricted least squares regression to combine forecasts will lead to serially correlated errors. His recommendation was to restrict the coefficients to sum to one. More recently, de Menezes and

Bunn (1993, 1998) carried out simulation and case-study analyses to investigate the issue of serial correlation. In broad terms, this work showed that combined forecast errors replicate some of the serial correlation from the individual forecast error processes, although some of the serial dependence can be diluted through combining, and that alternative combining forecast methods behave differently under serial correlation.

Based on the conclusions of these three studies, using a serial correlation criterion, we arrive at the following practical guidelines for combining forecasts:

(1) In cases of small sample sizes, use the simple average of the forecasts.

(2) Over larger samples, use an optimal combination, where independence is assumed if cross-correlation is small ($\rho < 0.5$), otherwise estimate the optimal weights restricted to the interval [0,1], or, a regression-based approach restricting the weights on the forecasts to sum to one.

(3) Check the combined forecast errors for serial correlation. If a pattern is found:

- Model the errors whenever the pattern is simple and does not involve a unit root.
- If there is a unit root in the errors, then revise the modelling procedure so as to consider the appropriate structure of the series (e.g. model the corresponding stationary change, restrict the combining regression or include the relevant lag in the model).
- If the serial correlation pattern is complex, then it might indicate another case of misspecification, the forecasting procedure as a whole should therefore be revised.

6.4 An illustrative example: percentage change in UK inflation

In business planning models, such as capital budgeting, a synthesis of several sources of economic forecasts are often required. Here, the data are one-year-ahead monthly forecasts of the annual percentage change in UK inflation from July 1983 to December 1988 from two forecasters London Business School and Goldman Sachs (de Menezes,1993). When we examined the individual forecast errors, we found that the correlation coefficient was reasonably stable and near zero; the distribution was normal, but serial correlation was an issue (ARIMA with a quarterly seasonal component and AR(1) structures).

We used the first five combining methods that were described in section 2 and estimated the combining weights based on the previous year (12 observations). The results are summarised in Table 4, where we note that all combinations have led to smaller forecast error variances than the larger of the individual variances. As the sample size is not small and the correlation coefficient (though not constant) is near zero, an adaptive method with independence assumption led to the smallest error variance, which is indeed smaller than both originals. Serial correlation was present in all combinations, but worse in those methods that attempted to estimate the correlation coefficient. Thus, confirming our expectations based on the proposed guidelines. Given the persistence of serial correlation in the combinations, the next stage in the analysis would be to model the error process.

******************* TABLE 4 APPROXIMATELY HERE ***********************

7. Summary and guidelines for combining forecasts

In the light of previous conclusions, a set of guidelines on the use of different combining rules emerge. Given stable statistical conditions and relatively well-behaved data, the comparative performances of the methods, in terms of accuracy, was found to depend on the error variance ratios of the forecasts, the correlation between forecast errors and the sample of past data used for estimation. However, with respect to either skewness or serial correlation, the individual forecast error variances became less relevant.

******************** TABLE 5 APPROXIMATELY HERE ***********************

Table 5 summarises the guidelines, which constitute a valuable decision support tool for combining forecasts. They aid on issues such as the choice of the initial combining rule, the continuous assessment of the appropriateness of the rule, and model-switching whenever the basic statistical conditions change. Furthermore, they emphasise the need for frequent diagnostic-tracking in order to enable the decision maker to exercise their judgement or preference.

Results from a large case-study on inflation forecasts (de Menezes and Bunn, 1993) have also shown the effectiveness of these guidelines. They confirm that, in practice, the forecaster is faced with different sources of error misspecification and, unfortunately, when eliminating or filtering one source, stronger evidence of other sources may emerge. Thus, combining forecasts becomes clearly a multi-attribute decision problem, where a significant amount of judgement on the value of each attribute may be required from the modeller, who in situations of structural change and instability will also need to decide whether and when to revert to simpler methods.

8. Conclusions and further implications

We have interpreted the problem of combining forecasts as a multi-criteria decision process. We have examined three criteria, although other dimensions can certainly be added to the analysis. The forecaster must decide upon which specification features to trade-off, and thus model-switching schemes will result from changes in the basic statistical conditions. Hence, the implementation of the proposed guidelines requires constant diagnostic-tracking and model-modeller interaction.

An approach that may be considered for model-switching is one of automatically generating various combinations and then selecting the best according to a set of diagnostics, for which priority rules would be given by the decision maker. The model would be updated with the arrival of individual forecasts or changes in priority rules. Following the history of individual forecast errors, a smaller set of combinations could be defined and generated. In practice, this implementation is not trivial. Real data is often not well behaved, making it sometimes difficult to define explicit criteria and constraints. Besides, the available diagnostics are usually poor and require considerable judgement. Therefore, provided some expertise from the modeller, a framework like Reeves and Lawrence's (1982, 1991), combining given multiple objectives, can become attractive. However, if one restricts one's choices to a subset of the more robust methods on grounds of simplicity and pragmatism, then Collopy and Armstrong's (1992) rule-based forecasting may become a viable alternative.

On the other hand, the complexity of such procedures should also be considered. Thus, an alternative and more effective approach consists of placing more emphasis upon diagnostic-tracking as a means for model review and reformulation, rather than "on-line" model-switching. In fact, diagnostic-tracking would in principle provide considerable feedback to the different modelling stages (e.g. forecast and model selection, correcting for serial correlation and bias). When there are indications that forecast errors are not subject to serial correlation and outliers, a predictive distribution can then be derived via a bootstrap, leading to measures of the uncertainty in the combined forecast. Nonetheless, there are still strong limitations concerning the choice of diagnostics available, since most statistical tests rely on either asymptotical or large-scale simulation results. In a combining forecast context, one often deals with small and time-variant forecasting histories, and thus judgement and expertise is often required.

We believe that this paper provides useful practical guidelines for those who are interested in combining forecasts, but who are also concerned about the structure of the errors that result from the combination. We observed that different combining rules react distinctly to misspecification. We discussed how the quality of the individual forecasts in the model have substantial implications for the overall forecasting performance, and thus we showed the importance of a rigorous analysis of the individual forecast errors prior to combining. Indeed, in order to be able to exercise preference and judgement while using multiple forecasts, decision makers should be attentive to the different individual forecasting patterns. They should be aware that, as individual forecasts can be subject to badly behaved errors, it would be a mistake to assume equivalent forecasting behaviour and produce a simple average (consensus) forecast without previously examining the individual forecasting patterns.

Acknowledgements

The authors would like to acknowledge the helpful comments of two anonymous referees.

References

Aksu, C., and Gunter, S.I. (1992), "An empirical analysis of the accuracy of SA, OLS, ERLS and NRLS combination forecasts", *International Journal of Forecasting* 8, 27-43.

Bates, J., and Granger, C. (1969), "The combination of forecasts", *Operations Research Quarterly* 20, 451-468.

Bera, A.K., and Jarque, C.M. (1980), "An efficient large sample test for normality of observations and regression residuals", Working Paper, The Australian National University, Canberra.

Bessler, D.A., and Brandt, J.A. (1981), "Forecasting livestock prices with individual and composite methods", *Applied Economics* 13, 513-522.

Bunn, D.W. (1975), "A Bayesian approach to the linear combination of forecasts", *Operational Research Quarterly* 26, 325-329.

Bunn, D.W. (1984), *Applied Decision Analysis*, McGraw-Hill, New York.

Bunn, D.W. (1985), "Statistical efficiency on the linear combination of forecasts", *International Journal of Forecasting* 1, 151-163.

Bunn, D.W. (1987), "Expert-use of forecasts: bootstrapping and linear models" in G. Wright and P Ayton (eds.), *Judgemental Forecasting*, Wiley, Chichester, 229-241.

Bunn, D.W. (1989), "Forecasting with more than one model", *Journal of Forecasting* 8, 161-166.

Bunn, D.W. (1996), "Non-traditional methods of forecasting", *European Journal of Operational Research* 92, 528-536.

Chatfield, C. (1988), "The future of time-series forecasting", *International Journal of Forecasting* 4, 411-419.

Clemen, R.T. (1986), "Linear constraints and the efficiency of combined forecasts", *Journal of Forecasting* 5, 31-38.

Clemen, R.T. (1989), "Combining forecasts: A review and annotated bibliography", *International Journal of Forecasting* 5, 559-583.

Clemen, R.T., and Winkler, R.L. (1987), "Calibrating and combining precipitation probability forecasts", in: R. Viertl (ed.), *Probability and Bayesian Statistics*, Plenum, New York, 97-110.

Clemen, R.T., and Winkler, R.L. (1993), "Aggregating point estimates: A flexible modeling approach", *Management Science* 39, 501-515.

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

Cooper, D., and Chapman, C. (1987), *Risk Analysis for Large Capital Projects*, Wiley, Chichester.

Coulson, N.E., and Robins, R.P. (1993), "Forecast combination in a dynamic setting", *Journal of Forecasting* 12, 63-67.

D'Agostino, R.B., Belanger, A., and D'Agostino Jr., R.B. (1990), "A suggestion for powerful and informative tests of normality", *The American Statistician*, 44, 316-321.

Deutsch, M., Granger, C.W.J. and Teräsvirta, T. (1994), "The combination of forecasts using changing weights", *International Journal of Forecasting* 10, 47-57.

Diebold, F.X. (1988), "Serial correlation and the combination of forecasts", *Journal of Business and Economic Statistics* 6, 105-111.

Diebold, F.X., and Pauly, P. (1987), "Structural change and the combination of forecasts", *Journal of Forecasting* 6, 21-40.

Donaldson, R.G., and Kamstra, M. (1996), "Forecast combining with neural networks", *Journal of Forecasting* 15, 49-61.

Engle, R.F., and Granger, C.W.J. (1987), "Cointegration and error correction: Representation, estimation and testing", *Econometrica* 55, 251-276.

Flores, B.E., and White, E.M. (1989), "Subjective vs. objective combining forecasts: An experiment", *Journal of Forecasting* 8, 331-341.

Gardner, E.S. Jr. (1983), "The trade-offs in choosing a time-series method", Commentaries on the M-Competition", *Journal of Forecasting* 2, 263-266.

Granger, C.W.J., and Ramanathan, R. (1984), "Improved methods of forecasting", *Journal of Forecasting* 3, 197-204.

Granger, C.W.J., White, H., and Kamstra, M. (1989), "Interval forecasting: An analysis based upon ARCH-quantile estimators", *Journal of Econometrics* 40, 87-96.

Guerard, J.B. (1987), "Linear constraints, robust-weighting and efficient composite modeling", *Journal of Forecasting* 6, 193-199.

Gunter, S.I. (1990), "Theoretical justification of the efficiency of simple average combinations", Working Paper, Temple University, Philadelphia, PA.

Gunter, S.I. (1992), "Nonnegativity restricted least squares combinations", *International Journal of Forecasting* 8, 45-59.

Hallman, J., and Kamstra, M. (1989), "Combining algorithms based on robust estimation techniques and co-integrating restrictions", *Journal of Forecasting* 8, 189-198.

Holden, K., and Peel, D.A. (1986), "An empirical investigation of combinations of economic forecasts", *Journal of Forecasting* 5, 229-242.

Holden, K., and Peel, D.A. (1989), "Unbiasedness, efficiency and the combination of economic forecasts", *Journal of Forecasting* 8, 175-188.

Holden, K., Peel, D.A., and Thomson, J.L. (1990), *Economic Forecasting: An Introduction*, Cambridge University Press, Cambridge.

Holly, S., and Tebbutt, S. (1993), "Composite forecasts, non-stationarity and the role of survey information", *Journal of Forecasting* 12, 291-300.

Holmen, J.S. (1987), "A note on the value of combining short-term earnings forecasts", *International Journal of Forecasting* 3, 239-243.

Kang, H. (1986), "Unstable weights in the combination of forecasts", *Management Science*, 32, 683-695.

LeSage, J.P., and Magura, M. (1992), "A mixture model approach to combining forecasts", *Journal of Business and Economic Statistics* 10, 445-452.

Levy, H. (1992), "Stochastic dominance and expected utility: Survey and analysis", *Management Science*, 38, 555-593.

Lobo, G.J. (1991), "Alternative methods of combining security analysts' and statistical forecasts of annual corporate earnings", *International Journal of Forecasting* 7, 57-63.

Makridakis, S., Anderson, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and 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, K., and Simmons, L. (1993), "The M-2 competition: A real-time judgementally based forecasting study", *International Journal of Forecasting* 9, 5-22.

MacDonald, R., and Marsh, I.W. (1994), "Combining exchange rate forecasts: What is the optimal consensus measure?", *Journal of Forecasting* 13, 313-33.

de Menezes, L.M. (1993), *Diagnostic-Tracking and Model Management in Combining Forecasts*, Unpublished Doctoral Thesis, London Business School, University of London.

de Menezes, L.M., and Bunn, D.W. (1993), "Diagnostic tracking and model specification in combined forecasts of U.K. inflation', *Journal of Forecasting* 12, 559-572.

de Menezes, L.M., and Bunn, D.W. (1998), "The persistence of specification problems in the distribution of combined forecast errors", forthcoming in *International Journal of Forecasting*.

Miller, C.M., Clemen, R., and Winkler, R.L. (1992), "The effect of nonstationarity on combined forecasts", *International Journal of Forecasting* 7, 515-529.

Mills, T.C., and Stephenson, M.J. (1985), "Forecasting contemporaneous aggregates and the combination of forecasts: The case of the UK monetary aggregates", *Journal of Forecasting* 4, 273-281.

Newbold, P. (1995), *Statistics for Business and Economics*, 4th edition, Prentice-Hall, New Jersey.

Newbold, P., and Granger, C.W.J. (1974), "Experience with forecasting univariate time series and the combination of forecasts (with discussion)", *Journal of the Royal Statistical Society, Series A* 137, 131-149.

Palm, F.C., and Zellner, A. (1992), "To combine or not to combine? Issues of combining forecasts', *Journal of Forecasting* 11, 687-701.

Ramsey, P.P., and Ramsey, P.H. (1990), "Simple tests of normality in small samples", *Journal of Quality and Technology*, **22**, 299-309.

Reeves, G.R., and Lawrence, K.D. (1982), "Combining multiple forecasts given multiple objectives", *Journal of Forecasting* 1, 271-279.

Reeves, G.R., and Lawrence, K.D. (1991), "Combining forecasts given different types of objectives", *European Journal of Operations Research* 51, 65-72.

Schmittlein, D.C., Kim, J., and Morrison, D.G. (1990), "Combining forecasts: Operational adjustments to theoretically optimal rules", *Management Science* 36, 1044-1056.

Schnaars, S.P. (1986a), "An evaluation of rules for selecting an extrapolation model on yearly sales forecasts", *Interfaces* 16, 100-107.

Schnaars, S.P. (1986b), "A comparison of extrapolation models on yearly sales forecasts", *International Journal of Forecasting* 2, 71-85.

Taylor, J.W., and Bunn, D.W. (1996), "Improving the accuracy of prediction intervals for combinations of forecasts: A simulation study", unpublished paper presented at the 15th International Symposium on Forecasting in Istanbul, Turkey, 23-26 June.

Taylor, J.W., and Bunn, D.W. (1998), "Combining forecast quantiles using quantile regression: Investigating the derived weights, estimator bias and imposing constraints", *Journal of Applied Statistics*, forthcoming.

Winkler, R.L., and Makridakis, S. (1983), "The combination of forecasts", *Journal of the Royal Statistical Society, Series A* 146, 150-157.

Subsamples	AREA & HEUR	AREA & DFS	HEUR & DFS
1-20	0.56	0.36	0.34
1-40	0.68	0.60	0.65
41-80	0.80	0.76	0.81
141-160	0.33	0.39	0.61
201-220	0.75	0.25	0.47

Table 1. Estimated Correlation Coefficients Between Errors from Individual Electric Load Forecasts

	AREA	DFS	HEUR
Mean	151.60	24.61	77.60
STD	480.47	465.34	464.66
MSE	253,830	217,143	221,945
Skewness	-0.20	-0.76	-0.15
Kurtosis	3.98	4.85	3.62
Bowman Shenton	11.48**	31.26**	4.86

** normality rejected at a 5% significance level

Table 2. Errors from Individual Electric Load Forecasts (246 observations)

	AREA, HEUR & DFS	AREA & HEUR	AREA & DFS	HEUR & DFS
Mean	84.64	114.65	88.11	51.15
STD	408.23	431.32	424.67	417.63
MSE	173,819	199,177	188,108	177,069
Skewness	-0.38	-0.23	-0.44	-0.41
Kurtosis	4.34	4.02	4.37	4.04
Bowman Shenton	24.33**	12.83**	27.16**	17.98**

** normality rejected for all combinations (5% level)

Table 3. Errors from Simple Average Combinations of Electric Load Forecasts (246 observations)

	Individual Forecasts Errors		Combined Forecasts Errors				
	LBS	GS	AVG	OUT	OPT	OPT^{I}	OPT^{R}
Mean	0.244	-0.169	0.038	0.119	0.208	0.175	0.163
Variance	0.261	1.241	0.412	0.272	0.328	0.246	0.278
Serial Correlation Structure	ARIMA(1,0,0) $\times(0,1,1)_4$	AR(1)	AR(1)	AR(1)	ARMA(1,2)	AR(1)	ARMA(4,3)

LBS - London Business School
GS - Goldman Sachs
AVG - Simple Average
OUT - Outperformance
OPT - Optimal
OPT^{I} - Optimal with independence assumption
OPT^{R} - Optimal with weights restricted to lie between 0 and 1

Table 4. Errors from Combinations of Forecasts of the Percentage Change in UK Inflation (54 observations)

CRITERIA		SMALL SAMPLES	MEDIUM SAMPLES	LARGE SAMPLES
VARIANCE	$\sigma_i \neq \sigma_j \quad \rho_{ij} < 0.5$	OUT	OPT^{I}	OPT/REG
	$\sigma_i \approx \sigma_j$	AVG	AVG	AVG
	$\sigma_i \neq \sigma_j \quad \rho_{ij} \geq 0.5$	OUT	OPT^{R}/REG	OPT^{R}/REG
SKEWNESS	$\rho_{ij} < 0.5$	OUT	OUT/OPT^{I}	OPT
	$\rho_{ij} \geq 0.5$	OUT	OPT^{R}	OPT^{R}
SERIAL CORRELATION	$\rho_{ij} < 0.5$	AVG	OUT^{I}/REG^{R}	OPT^{I}/REG^{R}
	$\rho_{ij} \geq 0.5$	AVG	OPT^{R}/REG^{R}	OPT^{R}/REG^{R}

AVG - Simple Average
OUT - Outperformance
OPT - Optimal
OPT^{I} - Optimal with independence assumption
OPT^{R} - Optimal with weights restricted to lie between 0 and 1
REG - Regression
REG^{R} - Regression with coefficients restricted to sum to 1

Table 5. Practical Guidelines for Combining Forecasts