Serial Correlation and the Combination of Forecasts

Francis X. Diebold

Division of Research and Statistics, Board of Governors of the Federal Reserve System, Washington, DC 20551

It is shown that regression-based methods of forecast combination lead to serially correlated combined prediction errors. The form of the serial correlation is characterized, and specification, estimation, and prediction are treated. A fully optimal combined predictor, which exploits the serial correlation, is developed and compared with existing regression-based methods in a numerical example, leading to decreases in mean squared prediction error.

KEY WORDS: Prediction; Pooling; Time-series model; Regression.

1. INTRODUCTION

The theory of forecast combination, which deals with optimal ways of combining different forecasts of the same future variable, was developed by Bates and Granger (1969) and extended by many others, including Granger and Ramanathan (1984). Availability of more than one forecast is the rule rather than the exception in many applied forecasting situations. For reasons that will shortly become clear, I refer to the Bates–Granger approach as the *variance–covariance* method and the Granger–Ramanathan approach as the *regression-based* method. The regression-based approach makes all of the standard results for the linear model immediately applicable to forecast combination, but to date little attention has been paid to the properties of residuals from combining regressions. [One exception is Diebold and Pauly (1987), who explored the effects of structural change of unknown form on the combining weights and developed weighted and time-varying coefficient methods of forecast combination.] Although it is true that ordinary least squares (OLS) regression-based combined forecasts have smaller within-sample mean squared prediction error than variance–covariance combinations, if the errors from the combining regression are nonwhite we should be able to exploit that fact to obtain better combined predictors. In fact, Granger and Ramanathan (1984) noted in passing that "Even if a pair of forecasts is combined, and each has white noise forecast errors, there is no reason to suppose that the combination will have white noise errors" (p. 203), and they suggested that the area is ripe for future research. In this article, the results of such an investigation are presented.

The possible presence of serial correlation in the combined prediction errors is an important issue for at least two reasons. First, if serial correlation is present, OLS estimates of the combining weights are inefficient and their associated standard deviation estimates are incon-sistent. Second, if serial correlation is present, then the combined forecast is no longer the best unbiased linear combination. Rather, best linear unbiased prediction requires prediction of the disturbance process as well (e.g., Goldberger 1964).

To fix ideas and notation, Section 2 begins with a brief literature review. In Section 3, the conditions under which serial correlation will arise in combining regressions are determined, and it is shown that the serial correlation problem is likely to be quite common. In Section 4, the form of the serial correlation is characterized. In Section 5, estimation and prediction with serially correlated combining regressions are treated, and a convenient (and optimal) approximation to the true serial correlation structure is developed. Section 6 contains a numerical example that illustrates the results, and Section 7 concludes the article.

2. THE THEORY OF COMBINING

2.1 The Variance–Covariance Method

Suppose that we have two competing unbiased forecasts of y_t made at time $t - 1$, $_{t-1}f_t^1$, and $_{t-1}f_t^2$, and suppose that we restrict ourselves to combined forecasts of the form

$$_{t-1}f_t^c = \phi \, _{t-1}f_t^1 + (1 - \phi) \, _{t-1}f_t^2, \qquad \phi \in R. \quad (1)$$

(The general case of K-step-ahead forecasts poses no additional problems.) Note that the combining weights, although summing to unity, are not necessarily convex. It is easily verified that the one-step-ahead combined prediction error satisfies the same equality:

$$e_t^c = \phi e_t^1 + (1 - \phi)e_t^2. \quad (2)$$

Thus,

$$\text{var } e_t^c = \phi^2 \sigma_1^2 + (1 - \phi)^2 \sigma_2^2 + 2\phi(1 - \phi)\sigma_{12}, \quad (3)$$

where σ_1^2 and σ_2^2 are the one-step-ahead prediction error variances of f^1 and f^2, respectively, and σ_{12} is their covariance.

Minimizing $\mathrm{var}(e_t^c)$ with respect to ϕ, the optimal combining weight is given by

$$\phi^* = (\sigma_2^2 - \sigma_{12})/(\sigma_1^2 + \sigma_2^2 - 2\sigma_{12}); \qquad (4)$$

hence the name "variance–covariance method." By inserting (4) into (3), we have

$$\sigma_{\min}^2 = \sigma_1^2\sigma_2^2(1 - \rho^2)/(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2), \qquad (5)$$

which is the minimum attainable combined prediction error variance. Note that $\sigma_{\min}^2 < \min(\sigma_1^2, \sigma_2^2)$, unless $\rho = \sigma_1/\sigma_2$ or $\rho = \sigma_2/\sigma_1$, in which case $\sigma_{\min}^2 = \min(\sigma_1^2, \sigma_2^2)$. Thus we have nothing to lose by combining and much to gain.

Maintaining the restriction of unbiasedness of the primary forecasts but allowing for an arbitrary finite number (m) of them, Reid (1969) and Granger and Newbold (1974) showed that the optimal combining weight vector ϕ^* is given by

$$\phi^*_{(m \times 1)} = (\Omega_i^{-1})/(i'\Omega^{-1}i), \qquad (6)$$

where Ω is the variance–covariance matrix of one-step-ahead forecast errors and i is a conformable column vector of ones. In practice, I estimate ϕ^* by replacing Ω with an estimate $\hat{\Omega}$, where $\hat{\Omega}_{ij} = 1/T \sum e_{it}e_{jt}$, T is the sample size, and the summation runs from $t = 1, \dots, T$, Thus, for example, in the two-variable case we have

$$\hat{\phi}^* = (\hat{\sigma}_2^2 - \hat{\sigma}_{12})/(\hat{\sigma}_1^2 + \hat{\sigma}_2^2 - 2\hat{\sigma}_{12}), \qquad (7)$$

where $\hat{\sigma}_1^2 = 1/T \sum e_{1t}^2$, $\hat{\sigma}_2^2 = 1/T \sum e_{2t}^2$, and $\hat{\sigma}_{12} = 1/T \sum e_{1t}e_{2t}$.

2.2 The Regression Method

Granger and Ramanathan (1984) showed that the preceding forecast combination theory has a regression interpretation, because ϕ^* is the least squares parameter vector estimate in the constrained regression

$$y_t = \sum_{i=1}^{m} \beta_i{}_{t-1}f_t^i + \varepsilon_t, \qquad (8)$$

where $\sum_{i=1}^{m} \beta_i = 1$. In the case of two primary forecasts, for example, the restriction may be conveniently imposed by writing

$$(y_t - {}_{t-1}f_t^2) = \phi({}_{t-1}f_t^1 - {}_{t-1}f_t^2) + \varepsilon_t^c, \qquad (9)$$

where $\varepsilon_t^c \equiv (y_t - {}_{t-1}f_t^2) - \phi({}_{t-1}f_t^1 - {}_{t-1}f_t^2)$. It is important to note that this is a homogeneous regression.

Estimation of ϕ by OLS then yields a result numerically identical to ϕ^* in the preceding. I will refer to the calculation of the optimal weights in this fashion as the *regression method*. The important point, however, is that the relaxation of a number of implicit restrictions (no intercept, slope coefficients sum to 1) that have been imposed so far leads to a combined within-sample mean squared prediction error even lower than the one resulting from the variance–covariance method. The

variance–covariance weight leads to an unbiased combined forecast *if* the component forecasts are unbiased, but the imposition of the constraint leads to a higher sum of within-sample squared prediction errors than would otherwise occur.

Failure to impose the $\sum \beta_i = 1$ constraint in the regression method (with no intercept) leads to a combined forecast that is biased unless (a) $E_{t-1}f_t^i = y_t$ ($i = 1, \dots, m$) and (b) $\sum \hat{\beta}_i = 1$, which is highly unlikely. It is well known, however, that under quadratic loss there is nothing *necessarily* undesirable about a biased forecast, and, as mentioned, the sum of within-sample squared prediction errors will be lower than if the constraint had been imposed. In addition, any bias that may be present in the component forecasts may be eliminated by including an intercept in the combining regression; the resulting combined forecast will be unbiased and have a smaller sum of within-sample squared prediction errors than the forecast obtained by any other combining method.

The unrestricted regression method amounts to the inclusion of one more forecast among those to be combined—the unconditional mean. It is therefore required that the variable being forecast be stationary; otherwise, an appropriate transformation should be performed prior to analysis. Inclusion of an intercept then is much more than a "bias correction"; an unrestricted regression-based combination may offer substantial benefits relative to a restricted combination even for unbiased forecasts. On the other hand, noisy economic data tend to produce volatile combining weights, so the incorporation of some prior information may provide valuable robustness (see Clemen and Winkler 1986; Kang 1986).

3. SERIAL CORRELATION IN THE COMBINING REGRESSION

Consider the usual combining regression:

$$y_t = \beta_0 + \sum_{i=1}^{m} \beta_i{}_{t-1}f_t^i + \varepsilon_t^c. \qquad (10)$$

The combined forecast is given by

$${}_{t-1}f_t^c = \hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i{}_{t-1}f_t^i, \qquad (11)$$

and the combined prediction error is

$$({}_{t-1}f_t^c - y_t) = \hat{\beta}_0 + y_t\left(\sum_{i=1}^{m} \hat{\beta}_i - 1\right)$$

$$+ \sum_{i=1}^{m} \hat{\beta}_i({}_{t-1}f_t^i - y_t)$$

$$= \hat{\beta}_0 + y_t\left(\sum_{i=1}^{m} \hat{\beta}_i - 1\right) + \sum_{i=1}^{m} \hat{\beta}_i\varepsilon_t^i$$

$$= \hat{\beta}_0 + y_t\left(\sum_{i=1}^{m} \hat{\beta}_i - 1\right) + v_t, \qquad (12)$$

where $v_t \equiv \Sigma_{i=1}^m \hat{\beta}_i \varepsilon_t^i$ and $\varepsilon_t^i \equiv {}_{t-1}f_t^i - y_t$. Although the presence and importance of serial correlation in the combined prediction errors may be approached in a number of ways, perhaps the most instructive is characterization of the conditions under which the combined prediction errors are not serially correlated. It is easily seen from (12) that, in general, if $\Sigma_{i=1}^m \hat{\beta}_i \neq 1$, then the combined prediction errors are serially uncorrelated only if (a) y is serially uncorrelated and (b) ε_t^i is serially uncorrelated ($i = 1, \ldots, m$).

The highly stringent nature of these conditions is at once apparent. Similarly, it is easy to show that for the case in which $\Sigma \hat{\beta}_i = 1$ the combined prediction errors are serially uncorrelated only if ε_i is serially uncorrelated ($i = 1, \ldots, m$).

By manipulating (12), we can easily obtain a general combined prediction error decomposition. Multiplying by -1, we have

$$(y_t - {}_{t-1}f_t^c)$$
$$= -\hat{\beta}_0 - y_t \left(\sum_{i=1}^m \hat{\beta}_i - 1 \right) + \sum_{i=1}^m \hat{\beta}_i \varepsilon_i, \quad (13)$$

where $\varepsilon_i \equiv (y_t - {}_{t-1}f_t^i)$. Granger and Ramanathan (1984) showed that the combined predictor is unbiased, so

$$E(y_t - {}_{t-1}f_t^c)$$
$$= -\hat{\beta}_0 - \left(\sum_{i=1}^m \hat{\beta}_i - 1 \right) \mu + \sum_{i=1}^m \hat{\beta}_i b_i = 0, \quad (14)$$

where b_i is the mean of the ith prediction error, $E(y - f^i)$, and $\mu = E(y)$. Thus,

$$\hat{\beta}_0 = \mu \left(1 - \sum_{i=1}^m \hat{\beta}_i \right) + \sum_{i=1}^m \hat{\beta}_i b_i. \quad (15)$$

Substituting this into (13), we obtain

$$(y_t - {}_{t-1}f_t^c)$$
$$= \left(1 - \sum_{i=1}^m \hat{\beta}_i \right) (y_t - \mu) + \sum_{i=1}^m \hat{\beta}_i (\varepsilon_i - b_i). \quad (16)$$

This shows that the combined prediction error may be decomposed into a linear combination, with weights summing to unity, of the prediction errors of the unconditional mean forecast and the bias-corrected primary forecast errors.

These results suggest that serial correlation is likely to play a much more important role in unrestricted regression-based forecast combination than in variance–covariance combination (equivalent to regression combination with no intercept and subject to the constraint that the weights sum to unity) or constrained regression combination in which an intercept is included but the other weights are still constrained to sum to unity. To see this, consider either of the latter two approaches to forecast combination. The combined prediction errors will be serially correlated if one or more of the ε_i ($i = 1, \ldots, m$) is serially correlated. This

may be viewed as rather unlikely, however, given that a common criterion of model adequacy is absence (or near absence) of serial correlation in the prediction errors. Thus it is likely that the combined prediction errors will display little, if any, serial correlation. On the other hand, the unrestricted regression-based combined prediction errors will be serially correlated (with probability 1) if (a) y is serially correlated or (b) ε_t ($i = 1, \ldots, m$) is serially correlated. Although (b) is again likely to be of small importance, (a) is likely to be of great importance. In fact, if y is not serially correlated, then it is (linearly) unpredictable, and there is no point in trying to forecast it. Note, however, that the nature and extent of the serial correlation induced in ($f^c - y$) (as well as our ability to detect it) depends critically on $\Sigma_{i=1}^m \hat{\beta}_i$ as well as the variance–covariance structure of the primary forecast errors.

The intuition behind the appearance of serial correlation in the combined prediction errors is easily seen by considering a simple special case, which will also appear later in a simulation example. Suppose the primary forecasts f^1 and f^2 have zero-mean prediction errors ε^1 and ε^2, and suppose that these errors are both contemporaneously and serially uncorrelated and normally distributed with identical variances. Then it is easily seen that the unconstrained regression-based combination will be such that (at least asymptotically) $\hat{\beta}_1 = \hat{\beta}_2$, but that we do not necessarily have $\hat{\beta}_1 + \hat{\beta}_2 = 1$. By unbiasedness of the combined forecast and primary forecasts, then $\hat{\beta}_0 = (1 - \hat{\beta}_1 - \hat{\beta}_2)\mu$. Thus, the more weight given to the primary forecasts, the less given to μ, and vice versa. It should also be noted that the "split" of the weights between f^1, f^2, and μ is determined by σ^2, the common prediction error variance of the two forecasts. If σ^2 is very large relative to the innovation variance of y, then a great deal of weight will be placed on μ, with correspondingly little weight placed on f^1 and f^2. The opposite is true for small σ^2, in which case we would expect $\hat{\beta}_1$ and $\hat{\beta}_2$ to be close to $\frac{1}{2}$, with $\hat{\beta}_0$ close to 0.

Consider the extreme case of $\hat{\beta}_1 = \hat{\beta}_2 = 0$ and $\hat{\beta}_0 = \mu$. It is immediately clear that this combined forecast $f^c = \mu$ has serially correlated errors, because the autoregressive moving average (ARMA) representation of y is defined by $(y_t - \mu) = ARMA(p, q)$. But $(y_t - \mu)$ is exactly the "combined" prediction error! In more standard situations, in which all weights are nonzero so that some weight is given to f^1 and f^2 in addition to μ, serial correlation is still induced in the combined prediction errors (although to a lesser extent) by virtue of the fact that some weight, however small, is given to μ.

By making use of my earlier result (16), we can get an informative decomposition of the combined prediction error:

$$(y - f^c) = \hat{\beta}_1 \varepsilon_1 + \hat{\beta}_2 \varepsilon_2 + (1 - \hat{\beta}_1 - \hat{\beta}_2)(y - \mu).$$
$$(17)$$

Thus the combined prediction error is a linear combination of the prediction error from the unconditional mean and the prediction errors of the primary forecasts. It is a weighted sum of the ARMA process $(y_t - \mu)$ and the white noise processes ε^i and, therefore, an ARMA process as well.

4. ON THE NATURE OF THE SERIAL CORRELATION

Because we may write the combined prediction error as

$$(1 - \sum \hat{\beta}_i)(y - \mu) + \sum \hat{\beta}_i(\varepsilon_i - b_i), \quad (18)$$

we may view it as the sum of $(m + 1)$ possibly non-independent zero-mean time series processes. If $(y - \mu)$ and $(\varepsilon_i - b_i)$, $i = 1, \ldots, m$, follow ARMA(p, q) processes (in which p, q, or both could be 0), then the combined prediction error will generally follow an ARMA(p, q) process, where both p and q are nonzero.

Specifically, it is easily seen that unless both $(y - \mu)$ and all of the $(\varepsilon_i - b_i)$, $i = 1, \ldots, m$, follow (possibly degenerate) finite moving averages, the combined prediction error will follow an ARMA(p, q) process with both $p, q > 0$. This follows from the aggregation theorem of Granger and Morris (1976), who showed that

$$\sum_{j=1}^{N} \text{ARMA}(p_j, q_j) = \text{ARMA}(x, y), \quad (19)$$

where $x \leq \sum_{j=1}^{N} p_j$ and $y \leq \max(x - p_j + q_j, j = 1, \ldots, N)$.

The inequalities arise because of the possible presence of common factors. For example, if

$$y_{1t} = \rho_1 y_{1,t-1} + \varepsilon_t^1, \qquad y_{2t} = \rho_2 y_{2,t-1} + \varepsilon_t^2,$$

$$|\rho_1| < 1, \quad |\rho_2| < 1, \quad \rho_1 \neq \rho_2,$$

$$\varepsilon_t^1 \sim (0, \sigma_1^2), \quad \varepsilon_t^2 \sim (0, \sigma_2^2), \quad (20)$$

then $x_t \equiv (y_{1t} + y_{2t})$ is an ARMA(2, 1) process:

$$(1 - \phi_1 L - \phi_2 L^2)x_t = v_t + \theta v_{t-1},$$

$$v_t \sim (0, \sigma_v^2), \quad (21)$$

where ϕ_1, ϕ_2, θ, and σ_v^2 depend on ρ_1, ρ_2, σ_1^2, and σ_2^2, and L is the lag operator. If, however, we have a common factor $(\rho_1 = \rho_2 = \rho)$, then $x_t = \text{AR}(1)$ because

$$(y_{1t} + y_{2t}) = \rho(y_{1,t-1} + y_{2,t-1}) + \xi_t, \quad (22)$$

where $\xi_t \equiv (\varepsilon_t^1 + \varepsilon_t^2) \sim (0, \sigma_1^2 + \sigma_2^2)$. It is unlikely that a common factor will occur, however, so in general the results hold with equality.

The aggregation theorem is valid for non-independent as well as independent component processes. The critical point is that the combined prediction error is the sum of $(m + 1)$ processes, which by the Granger–Morris theorem will in general be a nondegenerate ARMA(p, q) process if any of those $(m + 1)$ processes

is either AR or ARMA. One situation that is likely to arise in practice is

$$y \sim \text{ARMA}(p, q), \qquad \varepsilon^i \sim (0, \sigma_i^2).$$

Then

$$(f^c - y) = y\left(\sum_{i=1}^{m} \hat{\beta}_i - 1\right) + \sum_{i=1}^{m} \hat{\beta}_i \varepsilon^i \quad (23)$$

is the sum of an ARMA(p, q) process and white noise, which is ARMA$(p, \max(p, q))$. Our ability to detect this serial correlation and exploit it for predictive purposes critically on the variance of y relative to the variance of the white noise, as well as the nature of the serial correlation in y.

5. SPECIFICATION, ESTIMATION, AND PREDICTION

It will be convenient to assume first that a generalized combining model,

$$y = F\beta + \varepsilon, \qquad \varepsilon \sim \text{ARMA}(p, q), \quad (24)$$

has been properly specified and estimated and consider the prediction problem. Specification and estimation will be treated subsequently.

Rewrite the preceding model as

$$y_t = \beta_0 + \sum_{i=1}^{m} \beta_{i\,t-1} f_t^i + \varepsilon_t, \quad (24')$$

where $\varepsilon_t = [\Theta(L)/\Phi(L)]v_t$, $\Theta(L) = (1 - \theta_1 L - \cdots - \theta_q L^q)$, $\Phi(L) = (1 - \phi_1 L - \cdots - \phi_p L^p)$, and the innovation sequence $\{v_t\}$ is white noise. Thus,

$$y_t = \left(\beta_0 + \sum_{i=1}^{m} \beta_{i\,t-1} f_t^i\right) + [\Theta(L)/\Phi(L)]v_t, \quad (25)$$

so

$$\Phi(L)y_t = \Phi(L)\left(\beta_0 + \sum_{i=1}^{m} \beta_{i\,t-1} f_t^i\right) + \Theta(L)v_t, \quad (26)$$

which can be rewritten as

$$y_t = \left(\beta_0 + \sum_{i=1}^{m} \beta_{i\,t-1} f_t^i\right) + \Phi^*(L)\varepsilon_t + \Theta(L)v_t, \quad (27)$$

where $\Phi^*(L) = (\phi_1 L + \cdots + \phi_p L^p)$. Leading this equation by one period and expanding, we obtain

$$y_{t+1} = \beta_0 + \sum \beta_{i\,t} f_{t+1}^i + \phi_1 \varepsilon_t + \cdots + \phi_p \varepsilon_{t-p+1}$$

$$+ v_{t+1} - \theta_1 v_t - \cdots - \theta_q v_{t-q+1}, \quad (28)$$

which, when projected on the information set available at time t, yields the optimal combined predictor with estimated parameters:

$$_t f_{t+1}^c = \hat{\beta}_0 + \sum \hat{\beta}_{i\,t} f_{t+1}^i + \hat{\phi}_1 \hat{\varepsilon}_t + \cdots + \hat{\phi}_p \hat{\varepsilon}_{t-p+1}$$

$$- \hat{\theta}_1 \hat{v}_t - \cdots - \hat{\theta}_q \hat{v}_{t-q+1}. \quad (29)$$

This expression makes clear that [unless $\Theta(L) = 1$] the combined forecast arising from a combining regression with serially correlated residuals is dependent on all past residuals. We may approximate the optimal predictor, however, by setting $v_0 = 0$ (i.e., its unconditional mean) and then solving recursively. For example, if the distrubance in (24) follows an MA(1) process,

$$\varepsilon_t = v_t - \theta v_{t-1}, \tag{30}$$

then estimation yields $\hat{\beta}_i$ $(i = 0, \ldots, m)$ and $\hat{\theta}$. We set $\hat{v}_0 = 0$ and then obtain $\hat{v}_t = \hat{\varepsilon}_t + \hat{\theta}\hat{v}_{t-1}$ over the sample. Then

$$_t f^c_{t+1} = \hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i \,_t f^i_{t+1} - \hat{\theta}\hat{v}_t. \tag{31}$$

On the other hand, if the disturbance is a pure AR(p) [corresponding to the case of $\Theta(L) = 1$, as mentioned previously], then the component of the one-step-ahead combined forecast arising from prediction of the disturbance will depend on only the p most recent residuals. Consider, for example, an AR(1) disturbance:

$$y_t = \beta_0 + \sum_{i=1}^{m} \beta_i \,_{t-1} f^i_t + \varepsilon_t, \qquad \varepsilon_t = \rho\varepsilon_{t-1} + v_t. \tag{32}$$

In that case the optimal combined forecast is given by

$$_t f^c_{t+1} = \hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i \,_t f^i_{t+1} + \hat{\rho}\hat{\varepsilon}_t. \tag{33}$$

Now let us consider the specification problem. One obvious, although cumbersome, approach is to make use of the earlier prediction error decomposition. Because the OLS composite prediction errors are a weighted sum of the $(m + 1)$ primary prediction errors (including that from the unconditional mean), I could use standard methods to determine the stochastic structure of the primary prediction errors and then deduce the appropriate form of the composite prediction error. Such an approach, in addition to being tedious, is likely to be suboptimal in the sense that, in the presence of common (or nearly common) factors, it will lead to overparameterized models.

A more convenient approach is the use of consistent model specification procedures, such as the Schwarz (1978) information criterion (SIC). I select the disturbance process for which

$$\text{SIC} = \ln \hat{\sigma}^2_{\text{ML}} + \ln T/T(k) \tag{34}$$

is minimized, where T is the sample size, k is the number of estimated parameters, and $\hat{\sigma}^2_{\text{ML}}$ is the maximum likelihood (ML) estimate of the innovation variance.

Another approach to the specification of the disturbance term, convenient for both estimation and prediction, is approximation as a finite-order autoregression. Let the stationary invertible process followed by ε^c be given by

$$\psi(L)\varepsilon^c_t = \Theta(L)\eta_t. \tag{35}$$

Then the (generally infinite) AR representation is given by

$$\Theta^{-1}(L)\psi(L)\varepsilon^c_t = \eta_t \tag{36}$$

or $\sum_{i=0}^{\infty} b_i \varepsilon^c_{t-i} = \eta_t$. We can approximate this infinite autoregression to any desired degree of accuracy by a finite (pth order) autoregression. In addition, even very low-order autoregressions, such as AR(2), are capable of capturing a wide variety of stochastic structures, in the sense that the range of possible shapes of the associated spectral density is very broad (see Granger 1966). As shown previously, prediction in combining equations with AR(p) disturbances is also a straightforward task, because only the p most recent observations contain information relevant for prediction. This is in marked contrast to predictions in models with MA or ARMA disturbances, in which complicated transformations or approximate recursions are needed for prediction.

The problem of optimal order selection for the autoregressive approximation is easily solved by applying a result of Shibata (1980), who showed that if the Akaike (1974) information criterion (AIC) is used to fit an AR model to a process that in fact does not have a finite AR representation, then the chosen AR model asymptotically attains the minimum mean squared prediction error in the class of AR models. The AIC is given by

$$\text{AIC} = \ln \hat{\sigma}^2_{\text{ML}} + 2k/T. \tag{37}$$

6. AN EXAMPLE

To illustrate the results, two forecasts were generated for the AR(1) process:

$$(y_t - 20) = .9(y_{t-1} - 20) + \varepsilon_t,$$

$$\varepsilon_t \overset{\text{iid}}{\sim} N(0, 1), \quad t = 1, \ldots, 100, \tag{38}$$

as $f^1_{it} = y_t + \mu_{it}$ $(i = 1, 2)$ and $f^2_{it} = y_t + v_{it}$ $(i = 1, 2)$. The exercise was repeated two times, with

$$\mu_{1t} \sim \text{NID}(0, 6), \qquad v_{1t} \sim \text{NID}(0, 6)$$

$$\mu_{2t} \sim \text{NID}(0, 3), \qquad v_{2t} \sim \text{NID}(0, 3). \tag{39}$$

The successively smaller forecast-error variances associated with cases 1 and 2 should lead to progressively smaller intercepts in the combining equation, with more weight attached to the primary forecasts. Thus, as we move from case 1 to case 2, we would expect the sum of the combining weights to approach unity, with a corresponding decrease in serial correlation in the combined residuals. As indicated by earlier results then, I would expect large increases in combined forecasting performance for the first data set because of modeling of the serial correlation, with less forecasting enhancement in data set 2. The results, contained in Table 1, illustrate these points. [In each case, the combining regression was run over the first 80 observations, with the remaining 20 used as a "holdout sample" for mean squared prediction error (MSPE) comparison.]

Table 1. Ordinary Least Squares and Autoregressive Combining Results

	Data set 1		Data set 2	
	OLS	AR	OLS	AR
$\hat{\beta}_0$	9.16	14.95	5.66	11.07
	(7.21)	(12.92)	(4.91)	(8.90)
$\hat{\beta}_1$	.24	.14	.32	.23
	(4.92)	(3.64)	(6.00)	(5.03)
$\hat{\beta}_2$	.30	.13	.40	.22
	(5.55)	(3.48)	(6.95)	(4.89)
$\hat{\rho}$		.72		.64
		(8.96)		(7.30)
DW	1.08	1.89	1.30	1.95
MSPE(f^1)	6.44		3.22	
MSPE(f^2)	7.50		3.75	
MSPE(f^c)	5.71	1.56	2.98	1.45
MSPE(f^R)	3.38		1.69	
MSPE(f^{AV})	3.27		1.63	
z(1, 2)	−.03		−.03	
z(AR, 1)	−3.86		−2.81	
z(AR, 2)	−3.95		−2.89	

NOTE: t statistics are in parentheses; f^1 is the primary forecast 1, f^2 is the primary forecast 2, f^c is the unrestricted combined forecast (OLS or AR), f^R is the OLS restricted combined forecast, f^{AV} is the simple average combined forecast, and DW represents the Durbin–Watson statistic.

Consider first the OLS regressions. All coefficients are highly significant, and, as expected, the weight given to the unconditional mean is smaller (and the weights on the primary forecasts larger) for data set 2. Furthermore, the OLS combined forecast outperforms the best primary forecast in terms of MSPE in both data sets. Note also the very low Durbin–Watson statistics and that the DW for data set 1 is substantially smaller than that of data set 2.

As discussed earlier, both the SIC and AIC were used to identify an appropriate ARMA(p, q) model for the disturbance, where both p and q were allowed to range from 0 to 4. A simple model with AR(1) disturbances was unambiguously selected. The estimation results for the serially correlated combining regressions are also contained in Table 1. As before, all coefficients are highly significant, and the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ are smaller with a corresponding increase in $\hat{\beta}_0$. The first-order serial correlation coefficients are highly significant for both data sets, and, as expected, $\rho_1 > \rho_2$. Dramatic increases in forecasting performance are obtained when this serial correlation is taken into account.

For comparative purposes, the results for the restricted OLS combined forecast (f^R) and the simple arithmetic average combined forecast (f^{AV}) are also reported in Table 1. It should be noted that we have "stacked the deck" in favor of f^{AV} and f^R, since both primary forecasts are unbiased and the optimal combining weights should in fact be equal, because of equal primary forecast error variances within each data set. In spite of this, the autoregressive combined forecast maintains clear superiority in each data set. f^{AV} and f^R do beat the simple unrestricted OLS combined forecast in this example, however, illustrating the gains that can be made by imposing (valid) prior information.

We can formally test the null of equality of MSPE's for different forecasting methods, as follows (see Granger and Newbold 1977): Because

$$E(\varepsilon^1 + \varepsilon^2)(\varepsilon^1 - \varepsilon^2) = \sigma_1^2 - \sigma_2^2, \qquad (40)$$

we know that the expectation on the left is 0 iff the difference of variances is also 0. But the zero left-side expectation is equivalent to a zero correlation between $(\varepsilon^1 + \varepsilon^2)$ and $(\varepsilon^1 - \varepsilon^2)$. Thus the usual sample correlation coefficient may be used to test the null hypothesis of equal prediction error variances. Since $\frac{1}{2} \ln[(1 + \hat{\rho})/(1 - \hat{\rho})]$ is approximately normal with mean $\frac{1}{2} \ln[(1 + \rho)/(1 - \rho)]$ and variance $1/(T - 3)$, the test statistic

$$z = \sqrt{T - 3}/2 \ln[(1 + \hat{\rho})(1 - \rho)/(1 - \hat{\rho})(1 + \rho)] \qquad (41)$$

is approximately $N(0, 1)$, where T is the sample size used for predictive comparison.

These test statistics, denoted by $z(i, j)$, where i and j are the forecasts being compared, are also given in Table 1. Although (as expected) we cannot reject the null that $MSPE_1 = MSPE_2$, we reject the null of equal MSPE between either of the primary forecasts and the combined forecasts from the regression with AR(1) disturbances. The decreases in MSPE caused by modeling the serial correlation in the combining regression are clear.

7. CONCLUSIONS AND DIRECTIONS FOR FUTURE RESEARCH

It has been shown that regression-based combined forecast errors are likely to be serially correlated and that explicit modeling of the serial correlation can lead to improved combined predictors. In work in progress, the results of this article are applied to forecasting a variety of economic time series. Because proper evaluation can be achieved only by examining a large cross-section of representative time series, the task is quite complex and appropriately reserved for a future article.

Furthermore, the results indicate that attention must be paid to the method by which the time series are made stationary, since inappropriate differencing leads to a unit root in the moving-average lag operator polynomial of the combined prediction error, rendering it noninvertible. Methods of discriminating between trended and integrated processes in the context of forecast combination are therefore being developed and applied.

Finally, and more generally, the regression-based development of the forecast-combining problem leads to the direct applicability of techniques such as time-varying parameter models, robust estimation methods, intrinsically nonlinear combinations, and nonparametric prediction interval calculation (e.g., by the bootstrap). All of these directions are currently being explored.

ACKNOWLEDGMENTS

REFERENCES

Akaike, H. (1974), "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19, 716–723.

Bates, J. M., and Granger, C. W. J. (1969), "The Combination of Forecasts," *Operations Research Quarterly*, 20, 451–468.

Clemen, R. T., and Winkler, R. L. (1986), "Combining Economic Forecasts," *Journal of Business and Economic Statistics*, 4, 39–46.

Diebold, F. X., and Pauly, P. (1987), "Structural Change and the Combination of Forecasts," *Journal of Forecasting*, 6, 21–40.

Goldberger, A. S. (1964), *Economic Theory*, New York: John Wiley.

Granger, C. W. J. (1966), "The Typical Spectral Shape of an Economic Variable," *Econometrica*, 34, 150–161.

Granger, C. W. J., and Morris, M. J. (1976), "Time Series Modelling and Control," *Journal of the Royal Statistical Society*, Ser. A, 139, 246–257.

Granger, C. W. J., and Newbold, P. (1974), "Experience With Forecasting Univariate Time Series and the Combination of Forecasts," *Journal of the Royal Statistical Society*, Ser. A, 137, 131–164.

—— (1977), *Forecasting Economic Time Series*, New York: Academic Press.

Granger, C. W. J., and Ramanathan, R. (1984), "Improved Methods of Combining Forecasts," *Journal of Forecasting*, 3, 197–204.

Kang, H. (1986), "Unstable Weights in the Combination of Forecasts," *Management Science*, 32, 683–695.

Reid, D. J. (1969), "A Comparative Study of Time Series Prediction Techniques on Economic Data," unpublished Ph.D. thesis, University of Nottingham, Dept. of Mathematics.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.

Shibata, R. (1980), "Asymptotically Efficient Estimates of the Order of the Model for Estimating Parameters of a Linear Process," *The Annals of Statistics*, 8, 147–164.