Tests for Multiple Forecast Encompassing
Author(s): David Harvey and Paul Newbold
Source: *Journal of Applied Econometrics*, Vol. 15, No. 5 (Sep. - Oct., 2000), pp. 471-482
Published by: Wiley
Stable URL: http://www.jstor.org/stable/2678604
Accessed: 04-11-2016 15:27 UTC

JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **15**: 471–482 (2000)

TESTS FOR MULTIPLE FORECAST ENCOMPASSING

DAVID HARVEY[a] AND PAUL NEWBOLD[b]*
[a]*Department of Economics, Loughborough University, Loughborough, Leicestershire, LE11 3TU, UK*
[b]*School of Economics, University of Nottingham, University Park, Nottingham, NG7 2RD, UK*

SUMMARY

In the evaluation of economic forecasts, it is frequently the case that comparisons are made between a number of competing predictors. A natural question to ask in such contexts is whether one forecast encompasses its competitors, in the sense that they contain no useful information not present in the superior forecast. We develop tests for this notion of multiple forecast encompassing which are robust to properties expected in the forecast errors, and apply the tests to forecasts of UK growth and inflation. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

An important aspect of forecast evaluation is the comparison of competing forecasts. One of the many techniques employed to perform such comparison is that of testing for forecast encompassing, the motivation for which stems from the combination of forecasts literature (see, *inter alia*, Bates and Granger, 1969; Clemen, 1989; Granger, 1989). If a composite predictor formed as a weighted average of two individual forecasts is considered, then one forecast is said to encompass (or be conditionally efficient with respect to) the other if the inferior forecast's optimal weight in the composite predictor is zero. This definition follows Granger and Newbold (1973, 1986), Chong and Hendry (1986) and Clements and Hendry (1993). The inferior forecast then contains no useful information not present in the encompassing forecast, at least in the sense of linear combination. Harvey *et al.* (1998) investigated the issue of testing for forecast encompassing when two forecasts of the same quantity are available. The natural regression-based test using records of past forecast errors was found to lack robustness to forecast error non-normality,[1] and a number of modified tests were proposed and examined.

This paper generalizes the forecast encompassing approach to situations where, as is frequently the case, comparisons of a forecast with more than one competitor are required. Let $(f_{1t}, \ldots, f_{Kt})$ be K competing forecasts, taken to be unbiased or bias-corrected, of the actual quantity A_t. Assume for now that the forecasts are made one-step-ahead, with non-autocorrelated errors. Consider testing the null hypothesis that one forecast, f_1, encompasses its competitors. The joint testing procedure begins with a composite predictor

* Correspondence to: Paul Newbold, School of Economics, University of Nottingham, University Park, Nottingham, NG7 2RD, UK; e-mail: Paul.Newbold@Nottingham.ac.uk

[1] In work in progress, studying predictions of US macroeconomic variables reported in the 'Survey of Professional Forecasters', we have found strong evidence that the forecast error series are leptokurtic, to the extent that if these series were generated by Student's t distributions, the appropriate degrees of freedom would be in the range 5–7. Moderate evidence of forecast error skewness was also found.

Received 14 August 1998
Revised 14 December 1999

$$f_{ct} = (1-\lambda_1-\lambda_2-\cdots-\lambda_{K-1})f_{1t} + \lambda_1 f_{2t} + \lambda_2 f_{3t} + \cdots + \lambda_{K-1} f_{Kt} \quad 0 \le \lambda_i \le 1 \qquad (1)$$

which can alternatively be written as

$$e_{1t} = \lambda_1(e_{1t}-e_{2t}) + \lambda_2(e_{1t}-e_{3t}) + \cdots + \lambda_{K-1}(e_{1t}-e_{Kt}) + \varepsilon_t \quad 0 \le \lambda_i \le 1 \qquad (2)$$

where $e_{it} = A_t - f_{it}$ and ε_t is the error of the combined forecast. The null hypothesis that f_1 encompasses $f_2, \ldots, f_K$ is

$$\mathrm{H}_0 : \lambda_1 = \lambda_2 = \cdots = \lambda_{K-1} = 0 \qquad (3)$$

We refer to this concept as multiple forecast encompassing.

The natural regression-based test for multiple forecast encompassing is an F-test of the joint significance of the parameters in equation (2). Since this is a generalization of the two-forecast regression-based test studied by Harvey *et al.* (1998), we might expect similar problems to be displayed in the F-test if the forecast errors depart from normality. Formal analysis of the F-test's null distribution under forecast error non-normality reveals a critical lack of robustness. Writing the regression (2) in general form gives

$$y_t = X_t'\beta + \varepsilon_t \quad y = X\beta + \varepsilon \qquad (4)$$

where $y_t = e_{1t}$, $\beta=[\lambda_1\ \lambda_2\ \ldots\ \lambda_{K-1}]'$ and $X_t = [(e_{1t}-e_{2t})(e_{1t}-e_{3t})\ldots(e_{1t}-e_{Kt})]'$. Denoting the least squares estimator of β by $\hat{\beta}$, and assuming that $(e_{1t}, \ldots, e_{Kt})$ is an independent identically distributed sequence, theorem 5·3 of White (1984, p. 109) shows that, under the null hypothesis, given certain standard conditions,

$$D^{-1/2}n^{1/2}(\hat{\beta}-\beta) \xrightarrow{d} \mathrm{N}(0, I_{K-1}); \quad n(\hat{\beta}-\beta)'D^{-1}(\hat{\beta}-\beta) \xrightarrow{d} \chi^2_{K-1} \qquad (5)$$

where $D = M^{-1}QM^{-1}$, $M = E(X_tX_t')$ and $Q = V(n^{-1/2}X'\varepsilon)$. The F-test statistic is

$$F = n(K-1)^{-1}\hat{\beta}'\hat{D}^{-1}\hat{\beta} \qquad (6)$$

where $\hat{D} = \hat{M}^{-1}\hat{Q}\hat{M}^{-1}$, $\hat{M} = n^{-1}X'X$, $\hat{Q} = s^2n^{-1}X'X$ and $s^2 = (n-K+1)^{-1}\Sigma\hat{\varepsilon}_t^2$. Admitting the possibility of non-normal forecast errors allows the regression errors to be conditionally heteroscedastic, i.e. $E(\varepsilon_t^2 \mid X_t) = g(X_t)$. It is then possible to show that

$$\hat{D} \xrightarrow{p} HD \quad H = \{E[(X_tX_t')g(X_t)]\}^{-1}E(\varepsilon_t^2)E(X_tX_t') \qquad (7)$$

In the particular case of normal forecast errors, there is no conditional heteroscedasticity and $H = I_{K-1}$. The test statistic (6) is then correctly sized in the limit, with $(K-1)F \xrightarrow{d} \chi^2_{K-1}$. However, $H \neq I_{K-1}$ in general. For example, when the forecast errors are generated by the multivariate Student's t-distribution with v degrees of freedom (Dunnett and Sobel, 1954), it can be shown that $H = (v-4)(v-2)^{-1}I_{K-1}$, so that $(K-1)F \xrightarrow{d} (v-4)^{-1}(v-2)\chi^2_{K-1}$. For the case of K=3 and v=6, this result implies that nominal 5%-level and 10%-level F-tests have true asymptotic sizes 22·4% and 31·6% respectively.

Given the problem of lack of robustness to non-normality in the standard test, we consider three modified tests which are robust to conditional heteroscedasticity in the regression errors. These tests also allow for forecast error autocorrelation, permitting comparison of forecasts made at horizons greater than one. Section 2 introduces two regression-based modified tests, and Section 3 proposes a multiple encompassing test based on the modified Diebold–Mariano approach recommended by Harvey *et al.* (1998).[2] Section 4 reports results from a Monte Carlo simulation study, examining the tests' empirical size behaviour and comparing their relative powers. In Section 5 we test for multiple forecast encompassing among the historical forecasts of UK growth and inflation produced by the London Business School, Her Majesty's Treasury and the National Institute of Economic and Social Research.

2. MODIFIED REGRESSION-BASED TESTS

The non-normality problems associated with the standard regression-based test for multiple forecast encompassing can be shown to result from inconsistent estimation of the quantity Q implicit in (5), induced by conditional heteroscedasticity in the regression errors. The obvious modification, therefore, is to employ the heteroscedasticity-robust estimator of White (1980). In addition, we also make use of another robust estimator which is consistent under the null, but not the alternative hypothesis, as proposed in the two-forecast case by Harvey *et al.* (1998). Further, we follow the proposition of Diebold and Mariano (1995) and Harvey *et al.* (1998) that, for h-steps-ahead forecasts, a rectangular kernel with bandwidth $(h-1)$ should be adopted to account for forecast error autocorrelation. The multivariate test statistics which utilise these robust estimators are then

$$F_\alpha = (K-1)^{-1} n\hat{\beta}'\hat{D}_\alpha^{-1}\hat{\beta} \quad \alpha = 1, 2 \tag{8}$$

where $\hat{D}_\alpha = \hat{M}^{-1}\hat{Q}_\alpha\hat{M}^{-1}$, $\hat{M} = n^{-1}X'X$ and $\hat{Q}_\alpha$ have (i, j)th elements

$$\hat{q}_{\alpha,ij} = n^{-1}\left[\sum_{t=1}^{n} x_{it}x_{jt}u_{\alpha t}^2 + \sum_{m=1}^{h-1}\sum_{t=m+1}^{n} x_{it}x_{j,t-m}u_{\alpha t}u_{\alpha,t-m} + \sum_{m=1}^{h-1}\sum_{t=m+1}^{n} x_{i,t-m}x_{jt}u_{\alpha t}u_{\alpha,t-m}\right] \tag{9}$$

with $u_{1t} = \hat{\varepsilon}_t$ and $u_{2t} = y_t$. The modified estimators $\hat{Q}_\alpha$ are consistent for Q (under the null only in the case of $\hat{Q}_2$), and the statistics $(K-1)F_\alpha$ have null asymptotic $\chi^2{}_{K-1}$ distributions. In finite samples, although the statistics F_α no longer precisely follow $F_{K-1,n-K+1}$ distributions under the null, it is reasonable in practice to compare them with critical values from this distribution. The test statistics can also be written in slightly simpler forms:

$$F_\alpha = (K-1)^{-1} y'X\Phi_\alpha^{-1}X'y \tag{10}$$

where Φ_α have (i, j)th elements $\phi_{\alpha ij} = n\hat{q}_{\alpha,ij}$.

[2] In Harvey *et al* (1988), Spearman's rank correlation test was also analysed, because of its robustness to non-normality. However, this approach is not readily adapted to prediction several steps ahead when forecast errors will be autocorrelated. A further nonparametric approach might be based on the tests of Pesaran and Timmerman (1992, 1994), for the prediction of change or more generally of categorized data. Again, extension of these tests to multi-step prediction would be valuable.

3. MODIFIED DIEBOLD–MARIANO-TYPE TEST

Diebold and Mariano (1995) proposed an asymptotic test for the equality of two forecasts' accuracy, according to some measure of economic loss, such as the mean squared forecast error. This statistic was modified by Harvey *et al.* (1997) who derived a finite sample correction and proposed comparison of the statistic with critical values from the Student's t_{n-1} distribution. This approach can be applied to testing for forecast encompassing, as shown by Harvey *et al.* (1998) for the two-forecast case. These authors found the modified Diebold–Mariano-type test to display good size and power properties, and recommended its use in practical applications. It is consequently worth exploring whether a reliable version of the test exists for multiple forecast encompassing.

The two-forecast Diebold–Mariano approach requires definition of the null hypothesis in the form $\mathrm{H}_0 : E(d_t) = 0$. The null hypothesis for multiple forecast encompassing (3), in terms of the regression (4), is $\mathrm{H}_0{:}\beta{=}0$, or

$$\mathrm{H}_0 : [E(X_t X_t')]^{-1} E(X_t y_t) = 0 \tag{11}$$

Clearly, equation (11) is true if and only if

$$\mathrm{H}_0 : E(\Delta_t) = 0; \quad \Delta_t = [d_{1t} d_{2t} \dots d_{K-1,t}]'; \quad d_{it} = e_{1t}(e_{1t} - e_{i+1,t}) \tag{12}$$

The problem is now reduced to testing for the zero mean of a vector of random variables, so the multivariate analogue of the Diebold–Mariano statistic takes the form of Hotelling's (1931) generalized T^2-statistic (see, for example, Anderson, 1958)

$$MS^* = (K-1)^{-1}(n-1)^{-1}(n-K+1)\bar{d}'\hat{V}^{-1}\bar{d} \tag{13}$$

where $\bar{d} = [\bar{d}_1 \bar{d}_2 \dots \bar{d}_{K-1}]'$, $\bar{d}_i = n^{-1}\Sigma d_{it}$ and $\hat{V}$ is the sample covariance matrix. Our construction of $\hat{V}$ assumes $(h-1)$-dependency and a corresponding rectangular kernel as with the modified regression-based tests. The finite sample modification due to Harvey *et al.* (1997) applies directly to the sample variance (diagonal) terms of $\hat{V}$; it is also straightforward to show that the same correction factor is appropriate when estimating the covariance terms, so $\hat{V}$ has (i, j)th element

$$\begin{aligned}\hat{v}_{ij} =& n^{-1}[n+1-2h+n^{-1}h(h-1)]^{-1} \\ &\times \left[\sum_{t=1}^{n}(d_{it}-\bar{d}_i)(d_{jt}-\bar{d}_j) + \sum_{m=1}^{h-1}\sum_{t=m+1}^{n}(d_{it}-\bar{d}_i)(d_{j,t-m}-\bar{d}_j) + \sum_{m=1}^{h-1}\sum_{t=m+1}^{n}(d_{i,t-m}-\bar{d}_i)(d_{jt}-\bar{d}_j)\right]\end{aligned} \tag{14}$$

Hotelling's T^2-statistic (statistic (13) with an unmodified covariance estimator taking no account of autocorrelation) has, in the limit, as a result of the multivariate central limit theorem, a $(K-1)^{-1}\chi^2_{K-1}$ distribution. Although the finite sample distributional result is not exact, we maintain the use of $F_{K-1,n-K+1}$ critical values for statistic (13) in application.

We now note that the test is related to the second modified regression test of the previous section, F_2. First, it can be shown that F_2 is identical to $n(n-1)(n-K+1)^{-1}[n+1-2h+n^{-1}h(h-1)]^{-1}MS^*$ with the $(d_{it}-\bar{d}_i)$ in equation (14) replaced by d_{it}. Second, for $h{=}1$, after a little algebra,

$$MS^* = [n - (K-1)F_2]^{-1}(n - K + 1)F_2 \tag{15}$$

This result does not hold for $h > 1$. However, a more general monotonic relationship does exist for large n. For any forecast horizon, h, using the approximation $\sum_{t=m+1}^{n-m} d_{it} \approx \sum_{t=1}^{n} d_{it}$, the following relationship obtains asymptotically:

$$MS^* \approx \{n - [2h - 1 + n^{-1}h(h-1)](K-1)F_2\}^{-1}[n + 1 - 2h + n^{-1}h(h-1)](n-1)^{-1}(n - K + 1)F_2 \tag{16}$$

Thus, to the extent that F_2 is justified through its foundation on regression (2), MS^* is equally motivated.

4. MONTE CARLO SIMULATION

4.1. Empirical Sizes

In order to assess the finite sample behaviour of the multiple forecast encompassing tests, we conducted a simulation study of their empirical sizes. For simplicity, we ran simulation experiments for $K=3$, i.e. where one forecast encompasses two rival predictors under the null. Normalizing on $V(e_{1t}) = C(e_{1t}, e_{2t}) = C(e_{1t}, e_{3t}) = 1$ without loss of generality, the forecast errors have, under the null hypothesis, covariance matrix

$$V(ee') = \begin{bmatrix} 1 & 1 & 1 \\ 1 & v_2 & c_{23} \\ 1 & c_{23} & v_3 \end{bmatrix} \quad v_2, v_3 > 1 \quad 0 < c_{23} < (v_2 v_3)^{1/2} \tag{17}$$

It can be shown for all the test statistics considered that the null distributions are invariant to the arbitrary choice of the parameters v_2, v_3, c_{23}, provided they satisfy the conditions in equation (17). Forecast errors were generated from the multivariate normal distribution and the multivariate Student's t-distribution with five and six degrees of freedom. Empirical sizes were calculated for nominal 5%-level and 10%-level tests; simulations here and throughout this paper were performed using 10,000 replications, and all calculations were programmed in GAUSS.

The results for one-step-ahead prediction are given in Table I. (An unusually large sample size, $n=10{,}000$, was included to check asymptotic behaviour.) First, the simulations confirm the theoretical result that the F-test is incorrectly sized in the limit under forecast error non-normality. The over-sizing for this test is quite severe, increases with sample size, and, although convergence is slow, F is still somewhat over-sized in small and moderate samples. The three modified tests, F_1, F_2, MS^* all have the correct size for the largest sample simulation for non-normal, as well as normal, errors. The fundamental lack of robustness in the standard F-test is thus overcome asymptotically, although the size improvements vary a lot in finite samples. The two modified regression-based tests exhibit significant size distortions in small and moderate samples, with F_1 over-sized F_2, under-sized. The fact that these distortions persist for normal and Student's t errors until the sample size is very large provides little motivation for adopting these testing procedures in practice, given the relatively short series that one generally has available. However, the modified Diebold–Mariano-type test, MS^*, has more attractive properties. This test is also subject to some under-sizing in the smallest samples, but is robust and provides a broadly reliable alternative to the regression-based tests. Clearly, while the size advantage of this test over the F-test may not be compelling in small samples, that advantage increases dramatically with increasing number of sample observations.

We also investigated the size properties of our preferred test MS^* when using forecasts made at horizons greater than one. In fact we ran simulations for h=2, 4, 6, 8, and for purposes of illustration, analysis was restricted to normal errors. Table II shows the results, where for h=2, forecast errors e_{it} were generated from MA(1) processes $e_{it} = \varepsilon_{it} + \theta\varepsilon_{i,t-1}$, while for larger h these errors were generated as white noise. The empirical sizes are not as reliable as for one-step-ahead prediction, particularly for small samples and long horizons; this feature was observed, to a lesser extent, by Harvey *et al.* (1998) for the two-forecast test. For moderate samples and short horizons

Table I. Empirical sizes of nominal 5%-level and 10%-level tests for K=3, h=1

		5%-level			10%-level		
n	Test	N errors	t_6 errors	t_5 errors	N errors	t_6 errors	t_5 errors
8	F	4·7	9·2	10·8	9·6	16·8	18·2
	F_1	22·1	30·0	31·6	31·8	39·8	41·0
	F_2	0·0	0·0	0·0	0·0	0·0	0·0
	MS^*	2·2	1·7	1·5	6·3	5·0	4·6
16	F	4·8	11·6	14·1	9·4	19·4	22·0
	F_1	14·5	21·6	23·5	22·1	30·1	32·4
	F_2	0·4	0·2	0·3	3·3	2·7	2·6
	MS^*	3·0	2·4	2·3	7·8	6·9	6·5
32	F	4·9	14·7	18·0	9·8	22·5	26·3
	F_1	10·0	15·8	17·9	16·3	23·9	25·3
	F_2	2·5	1·6	1·4	7·0	6·0	5·6
	MS^*	4·2	3·1	2·8	9·1	8·1	7·9
64	F	5·2	16·6	20·0	10·1	24·3	28·9
	F_1	7·8	11·9	12·7	13·3	18·7	20·0
	F_2	4·0	2·6	2·4	8·9	7·5	7·3
	MS^*	5·0	3·5	3·3	9·7	8·5	8·3
128	F	5·0	18·2	22·6	10·1	26·5	31·4
	F_1	6·2	9·0	9·6	11·8	15·5	16·4
	F_2	4·2	3·5	3·2	9·3	8·8	8·1
	MS^*	4·6	4·0	3·5	9·9	9·3	8·5
256	F	4·8	18·7	24·2	10·1	26·9	33·3
	F_1	5·5	7·6	8·1	11·1	13·1	14·4
	F_2	4·5	4·0	3·8	9·9	9·1	8·6
	MS^*	4·7	4·2	4·0	10·1	9·4	8·8
512	F	5·3	20·2	26·2	10·0	29·1	35·2
	F_1	5·5	6·4	6·8	10·4	12·1	12·3
	F_2	5·0	4·5	3·9	9·9	9·5	8·9
	MS^*	5·1	4·6	4·0	10·0	9·7	9·0
10,000	F	4·9	22·5	32·0	9·8	31·8	41·7
	F_1	5·0	5·2	4·9	9·9	10·3	9·9
	F_2	4·9	5·1	4·6	9·9	10·0	9·3
	MS^*	5·0	5·1	4·6	9·9	10·0	9·3

the test is viable, but caution must be taken to avoid strong inference in practical applications where medium to long forecast horizons are being evaluated.

4.2. Power Comparisons

In the second part of our simulation study we estimated size-adjusted powers of each test for the case $K=3$. We considered one-step-ahead forecasts with errors drawn from the multivariate normal and Student's t_6-distributions. Tests were conducted at the 5% level. Since from equation (15) MS^* is a monotonic function of F_2, these two tests have identical size-adjusted powers.

It is possible to show that, under the alternative hypothesis, the four test statistics' distributions depend only on the population R^2 of the encompassing regression (2). In the case $K=3$, the encompassing regression is, in the notation of (4),

$$y_t = \lambda_1 x_{1t} + \lambda_2 x_{2t} + \varepsilon_t \tag{18}$$

We can assume $V(\varepsilon_t) = 1$ without loss of generality. Consider now the orthogonal pair

$$w_{1t} = \lambda_1 x_{1t} + \lambda_2 x_{2t} \quad V(w_{1t}) = \sigma^2_{w1} = \lambda_1^2 V(x_{1t}) + \lambda_2^2 V(x_{2t}) + 2\lambda_1\lambda_2 C(x_{1t}, x_{2t}) \tag{19}$$

$$w_{2t} = \alpha_1 x_{1t} + \alpha_2 x_{2t} \quad V(w_{2t}) = \sigma^2_{w2} = \alpha_1^2 V(x_{1t}) + \alpha_2^2 V(x_{2t}) + 2\alpha_1\alpha_2 C(x_{1t}, x_{2t}) \tag{20}$$

Table II. Empirical sizes of nominal 5%-level and 10%-level MS^* tests for $K=3$ (normal errors)

	5%-level			10%-level		
n	$h=2$, $\theta=0$	$h=2$, $\theta=0·5$	$h=2$, $\theta=0·9$	$h=2$, $\theta=0$	$h=2$, $\theta=0·5$	$h=2$, $\theta=0·9$
8	9·6	10·0	10·2	14·2	15·3	15·7
16	12·4	10·6	9·6	18·2	17·0	15·8
32	8·9	7·5	6·9	15·2	13·1	12·9
64	6·8	5·7	5·5	12·3	11·2	11·2
128	5·7	5·6	5·6	11·2	10·8	10·8
256	5·1	5·2	5·1	10·2	10·4	10·2
512	5·1	4·9	4·7	10·3	10·2	10·0

	5%-level			10%-level		
n	$h=4$	$h=6$	$h=8$	$h=4$	$h=6$	$h=8$
8	4·9	1·8	–	7·2	2·5	–
16	11·6	7·4	5·2	15·3	10·1	6·9
32	15·1	12·7	10·5	21·0	16·8	14·2
64	12·8	14·1	13·7	18·5	19·4	18·5
128	8·2	10·2	11·8	13·5	15·7	17·5
256	6·5	7·5	8·7	11·5	13·1	14·4
512	5·9	6·5	7·0	11·0	11·6	12·3

Notes: In the upper panel of the table, autocorrelated forecast errors were generated with MA(1) structure and moving average parameter θ. In the lower panel of the table, white noise forecast errors were generated.

where α_1, α_2 are defined such that $C(w_{1t}, w_{2t}) = 0$. Further consider the standardized pair

$$z_{1t} = \sigma_{w1}^{-1} w_{1t} \quad z_{2t} = \sigma_{w2}^{-1} w_{2t} \tag{21}$$

Then z_{1t}, z_{2t} are uncorrelated with mean zero and variance one, and we can consider a transformation of regression (18)

$$y_t = \delta_1 z_{1t} + \delta_2 z_{2t} + \varepsilon_t \tag{22}$$

where the true parameters are $\delta_1 = \sigma_{w1}$, $\delta_2 = 0$. Now the test statistics F, F_1, F_2, MS^* are invariant to the transformation from X to Z; the statistics will consequently be identical irrespective of whether regression (18) or (22) is used. Clearly, the only parameter that the test statistics derived from (22) can depend on is σ_{w1}. It is also true that

$$\sigma_{w1}^2 = (1 - R_p^2)^{-1} R_p^2 \tag{23}$$

where R^2_p is the population R^2 of the original encompassing regression (18).

Given that the tests' power depends solely on R^2_p, we simulated the alternative hypothesis using (22) and appropriate values of R^2_p for each sample size to achieve sensible size-adjusted powers. Table III reports the results of these simulations. In large samples, there is little to choose between the size-adjusted powers of the four tests. However, for moderately sized samples, the F-test has (in some cases dramatically) greater size-adjusted power, and F_1, while not as powerful as F, displays significantly superior power performance to F_2 and MS^*. The differences are more marked for normal forecast errors than t_6 errors, and for the same R^2_p, the tests have, as might be expected, higher power for normal errors. Given its power advantages, it is unfortunate that the F-test is not robust to non-normality, and that in the cases where the most natural White-type modified test, F_1, has greater power, its use is precluded by severe over-sizing. Since F_2 and MS^* have identical size-adjusted power, we suggest MS^* should be preferred due to its more reliable finite sample size behaviour. However, our preference for MS^* over F is clearly far more pronounced in large samples than in small. Indeed, the simulation evidence of this section suggests that, unless one is very concerned about the possibility of non-normality, it may be preferable to use the F-test in very small sample sizes, in spite of its undesirable asymptotic properties.

5. ANALYSIS OF UK FORECASTS

In this section we present an empirical application of testing for multiple forecast encompassing. The data are forecasts of UK growth and inflation produced by the London Business School (LBS), Her Majesty's Treasury (HMT) and the National Institute of Economic and Social Research (NIESR), as originally constructed and discussed by Pepper (1998, Chapter 8). The forecasts are end-of-year predictions of growth and inflation for the years 1978–1995 (18 observations), made at two different horizons.

A degree of variation exists in the measures of the variables being forecast, both across institutions and through time. Growth is consistently measured as the percentage change over a calendar year of GDP at factor cost, but a mixture of output, expenditure and average bases are employed in the estimation of this variable. The specific inflation measures used are the retail price index (with and without mortgage interest payments) and the consumer price index; the measures also vary between calendar year and fourth quarter comparisons. The precise variable measures used for each institution and time period are detailed in Pepper (1998, Appendix 8·3). In order to achieve consistency when calculating forecast errors, the actuals with which the forecasts are compared match the particular definition of the corresponding forecast, and are taken from each institution's own data (with the exception of the most recent HMT data which are taken from *Economic Trends*). Clearly some concern exists as to the legitimacy of comparison when such variations in the data are manifest; since our analysis is based on forecast *errors*, and concern is taken to compare the forecasts with the appropriate actuals, we postulate that any biases introduced by such discrepancies will be small. Further, the correlations between the series of actuals across institutions are very high — at least 0·95 (see Mills and Pepper, 1999, Table 2) — implying a large degree of conformity.

With regard to the forecast horizons, the LBS forecasts are made ten and fourteen months ahead, HMT nine and thirteen months ahead, and NIESR ten and thirteen months ahead. Although the forecasters differ in their months of prediction, we follow Mills and Pepper (1999) in the assumption that the forecasts produced at nine and ten months ahead, and thirteen and fourteen months ahead, are sufficiently close to be compared. Further, also following these authors, we deem the nine/ten months ahead and thirteen/fourteen months ahead predictions to be one-step-ahead ($h=1$) and two-steps-ahead ($h=2$) forecasts respectively.

Mills and Pepper (1999) provide an extensive analysis of the LBS, HMT and NIESR forecasting records using this data set, including evaluation of forecasts at other horizons not considered here since data are not available for all three institutions in these cases. These authors examine the performance of each set of forecasts in terms of unbiasedness, efficiency and

Table III. Estimated size-adjusted powers of 5%-level tests for $K=3$, $h=1$

n	Population R^2	Test	N errors	t_6 errors
8	0·66	F	71·1	59·9
		F_1	54·9	40·5
		F_2, MS^*	23·4	21·7
16	0·40	F	71·6	55·2
		F_1	61·8	45·4
		F_2, MS^*	44·6	37·3
32	0·22	F	71·1	49·5
		F_1	64·4	47·1
		F_2, MS^*	58·6	46·0
64	0·12	F	71·9	49·1
		F_1	67·6	49·1
		F_2, MS^*	65·4	52·2
128	0·06	F	71·8	44·0
		F_1	70·7	48·4
		F_2, MS^*	69·9	50·9
256	0·031	F	71·7	44·8
		F_1	70·6	49·2
		F_2, MS^*	70·5	51·7
512	0·016	F	72·3	43·4
		F_1	72·5	48·7
		F_2, MS^*	72·3	51·0

accuracy, and investigate evidence for the notions that the forecasts might have a tendency to be incorrect in the same direction, underestimate actual changes and poorly predict major business cycle turns. In addition, Mills and Pepper perform tests for the equality of prediction mean squared errors among the sets of forecasts, and conduct pairwise forecast encompassing tests for the cases where the equal accuracy null hypothesis is rejected. Their test results fail to reject the null hypotheses that the LBS and HMT one-step-ahead inflation forecasts respectively encompass the corresponding NIESR forecasts, and reject the reverse hypotheses that the NIESR forecasts encompass those of LBS and HMT respectively.

We extend the work of Mills and Pepper by testing for encompassing jointly using the tests discussed in the previous sections. We account for potentially biased forecasts by de-meaning the forecast error series used in the encompassing test statistics. For a given variable and horizon, one forecast must be selected as the numeraire in the tests (f_1 in the notation of this paper), the interpretation being that this forecast encompasses the other two. In order to allow for all possible results without pre-selection we execute tests with each forecast used as the numeraire in turn.

Table IV reports the results of tests for multiple forecast encompassing. Test statistics (corrected for forecast error bias) and associated probability values are provided for our recommended test, MS^*, and the standard regression-based F-test. The F-test requires modification in order to achieve robustness to forecast error autocorrelation, which would be expected for the two-steps-ahead predictions. This is done by allowing the regression errors in the encompassing regression (2) to follow a moving average process of order one, i.e. $\varepsilon_t = v_t - \theta v_{t-1}$ where v_t is white noise. Given that the sample of forecast errors is quite short, we do not report results for the modified regression-based tests F_1 and F_2, which are severely mis-sized in such cases.

As would be expected from the simulation study findings in the previous section, the F-test and MS^* test yield different results, the F-test almost invariably having the lower probability value. The small sample available precludes reliable testing for forecast error normality, thus it is unclear whether this greater tendency towards rejection is a consequence of the test's superior power in small samples, or of over-size induced by the test's inherent lack of robustness. However, in the majority of cases, the degree of difference between the tests is quite small, and only leads to strongly conflicting inferences for the growth two-steps-ahead forecasts.

Table IV. Tests for forecast encompassing

			Numeraire		
Variable	h	Test	LBS	HMT	NIESR
Growth	1	F	3·26 [0·065]	2·64 [0·102]	7·35 [0·005]
		MS^*	3·15 [0·070]	1·30 [0·299]	4·15 [0·035]
	2	F	8·11 [0·004]	10·66 [0·001]	11·24 [0·001]
		MS^*	2·91 [0·084]	0·64 [0·543]	1·98 [0·170]
Inflation	1	F	6·01 [0·011]	6·16 [0·010]	18·59 [0·000]
		MS^*	2·66 [0·101]	2·97 [0·080]	4·25 [0·033]
	2	F	0·65 [0·536]	10·60 [0·001]	4·17 [0·035]
		MS^*	0·28 [0·761]	3·99 [0·039]	7·66 [0·005]

Notes: Probability values are given in brackets. LBS denotes London Business School, HMT denotes Her Majesty's Treasury, and NIESR denotes National Institute of Economic and Social Research.

For the one-step-ahead growth forecasts, we fail to reject the null that the HMT forecasts encompass, or cannot be improved by combination with, the corresponding LBS and NIESR predictions at the 10% significance level. In contrast, the hypotheses that the LBS and NIESR forecasts encompass their respective rivals are rejected at this significance level, implying that combination of the LBS (NIESR) predictions with those of HMT and/or NIESR (LBS) would lead to an improvement in forecast performance.

The test results for the two-steps-ahead growth forecasts give conflicting inferences: the F-test yields strong rejections of each of the null hypotheses, implying no institution's predictions contain all the useful information present in the three sets of forecasts considered. When MS^* is employed, rejection at the 10% level is likewise obtained when LBS is the numeraire, but for HMT and NIESR the inferences are the opposite of the F-test's, with no strong evidence against the respective nulls. Which set of results is to be believed is unclear, as discussed above, although our preference is for inference from the more reliable MS^* test.

For the one-step-ahead inflation forecasts, the null hypotheses that the HMT and NIESR forecasts encompass their rivals are rejected at the 10%-level by both testing methodologies. The F-test further rejects the LBS encompassing null, while for MS^*, rejection of this hypothesis is obtained at the 10·1% significance level. Although this is marginally outside conventional significance levels, evidence does appear to exist (especially when bearing in mind the relatively low power of the MS^* test) that the LBS one-step-ahead inflation predictions, along with those of HMT and NIESR, can be improved by combination.

Finally, for the two-steps-ahead inflation forecasts, we fail to reject the null of LBS forecast encompassing at conventional significance levels, but strongly reject the HMT and NIESR encompassing hypotheses. The evidence therefore suggests that the HMT and NIESR predictions, but not those of LBS, would benefit from forecast combination in this case.

Failure to reject the null hypothesis of forecast encompassing among multiple forecasts does not necessarily imply that the numeraire forecast is substantially superior and dominant with respect to its competitors. While this is one legitimate possibility, failure to reject may also arise when the forecasts are very similar; in this case the encompassing forecast cannot be improved by combination as the rival predictions amount to very nearly the same forecast. Relatively large sampling variability, test under-size and low power may also drive non-rejections, as may a mixture of the dominance and similarity effects described above. The strong inference in encompassing test results is when rejection of the null occurs, the implication then being that the numeraire forecast can be improved through linear combination with the other forecasts.

6. CONCLUSION

It is often desirable to evaluate a forecast by testing for forecast encompassing against a number of competing predictors. In this paper we have generalized the forecast encompassing approach to accommodate this possibility. Further, we have demonstrated that, like its two-forecast counterpart, the natural regression-based F-test is not robust to conditional heteroscedasticity in the regression errors, a feature of forecast error non-normality. Three modified tests that are robust have been proposed, with all the tests' size and power properties being examined by way of Monte Carlo simulation. An empirical exercise in testing for multiple forecast encompassing has been undertaken using UK forecast data.

The results of the simulations lead us to recommend the modified Diebold–Mariano-type test, *MS**, as the preferred test for multiple forecast encompassing, at least in moderately large samples, due to its good size and reasonable power properties. Practitioners should not, however, forget its limitations when using small samples; in particular, a degree of under-sizing for one-step-ahead forecasts, over-sizing for multi-step-ahead evaluation, and low power. We regard the test's relatively low power in small samples as the price paid for robustness and reliable empirical size. Finally, the test could also be further modified along the lines of Harvey *et al.* (1999) to achieve improved size if autoregressive conditional heteroscedasticity is expected to be a feature of the forecast errors.

ACKNOWLEDGEMENTS

We are grateful to Terry Mills for the provision of the data used in the empirical example, and to Hashem Pesaran and three referees for helpful comments on an earlier draft.

REFERENCES

Anderson TW. 1958. *An Introduction to Multivariate Statistical Analysis*. Wiley: New York.

Bates JM, Granger CWJ. 1969. The combination of forecasts. *Operational Research Quarterly* **20**: 451–468.

Chong YY, Hendry DF. 1986. Econometric evaluation of linear macroeconomic models. *Review of Economic Studies* **53**: 671–690.

Clemen RT. 1989. Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* **5**: 559–583.

Clements MP, Hendry DF. 1993. On the limitations of comparing mean square forecast errors. *Journal of Forecasting* **12**: 617–637.

Diebold FX, Mariano RS. 1995. Comparing predictive accuracy. *Journal of Business and Economic Statistics* **13**: 253–263.

Dunnett CW, Sobel M. 1954. A bivariate generalisation of Student's *t*-distribution, with tables for certain special cases. *Biometrika* **41**: 153–169.

Granger CWJ. 1989. Combining forecasts — twenty years later. *Journal of Forecasting* **8**: 167–173.

Granger CWJ, Newbold P. 1973. Some comments on the evaluation of economic forecasts. *Applied Economics* **5**: 35–47.

Granger CWJ, Newbold P. 1986. *Forecasting Economic Time Series*, 2nd edn. Academic Press: Orlando, FL.

Harvey DI, Leybourne SJ, Newbold P. 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* **13**: 281–291.

Harvey DI, Leybourne SJ, Newbold P. 1998. Tests for forecast encompassing. *Journal of Business and Economic Statistics* **16**: 254–259.

Harvey DI, Leybourne SJ, Newbold P. 1999. Forecast evaluation tests in the presence of ARCH. *Journal of Forecasting* **18**: 435–445.

Hotelling H. 1931. The generalization of Student's ratio. *Annals of Mathematical Statistics* **2**: 360–378.

Mills TC, Pepper GT. 1999. Assessing the forecasters: an analysis of the forecasting records of the Treasury, the LBS and the National Institute. *International Journal of Forecasting* **15**: 249–259.

Pepper GT. 1998. *Inside Thatcher's Monetarist Revolution*. Macmillan: Basingstoke.

Pesaran MH, Timmermann AG. 1992. A simple non-parametric test of predictive performance. *Journal of Business and Economic Statistics* **10**: 461–465.

Pesaran MH, Timmermann AG. 1994. A generalization of the non-parametric Henriksson–Merton test of market timing. *Economics Letters* **44**: 1–7.

White H. 1980. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**: 817–838.

White H. 1984. *Asymptotic Theory for Econometricians*. Academic Press: Orlando, FL.