JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **25**: 574–594 (2010)
Published online 13 April 2010 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/jae.1176

FORECAST EVALUATION OF SMALL NESTED MODEL SETS

KIRSTIN HUBRICH[a] AND KENNETH D. WEST[b]*
[a] *European Central Bank, Frankfurt, Germany*
[b] *Department of Economics, University of Wisconsin, Madison, WI, USA*

SUMMARY

We propose two new procedures for comparing the mean squared prediction error (MSPE) of a benchmark model to the MSPEs of a small set of alternative models that nest the benchmark. Our procedures compare the benchmark to all the alternative models simultaneously rather than sequentially, and do not require re-estimation of models as part of a bootstrap procedure. Both procedures adjust MSPE differences in accordance with Clark and West (2007); one procedure then examines the maximum t-statistic, while the other computes a chi-squared statistic. Our simulations examine the proposed procedures and two existing procedures that do not adjust the MSPE differences: a chi-squared statistic and White's (2000) reality check. In these simulations, the two statistics that adjust MSPE differences have the most accurate size, and the procedure that looks at the maximum t-statistic has the best power. We illustrate our procedures by comparing forecasts of different models for US inflation. Copyright © 2010 John Wiley & Sons, Ltd.

Received 31 March 2008; Revised 17 February 2009

Supporting information may be found in the online version of this article.

1. INTRODUCTION

Forecast evaluation frequently involves comparison of a small set of models, one of which is a null model nested in the alternative models. There are two broad classes of applications. In one class, applicable to studies of asset returns, the null model is a martingale difference, perhaps with drift (i.e., a random walk or random walk with drift for the asset price). Examples include Hong and Lee (2003), who study exchange rates, and Sarno *et al*. (2005), who study interest rates; each paper compares a random walk to a half dozen or so other models. In the second class of applications, the null model sometimes relies on stochastic predictors, typically via a univariate autoregression. Examples include Billmeier (2004), who compares a univariate autoregression (AR) to four other models that include four different measures of the output gap, and Hubrich (2005) and Hendry and Hubrich (2006, 2009), who compare univariate forecasts of aggregate inflation to a couple of other forecast models that add disaggregate components of inflation to the univariate model. This class of applications is important at policy institutions or for policy observers where it is of interest to compare forecasts from different models in a suite of models built to account for different aspects of the economy.

Our aim in this paper is to propose and evaluate procedures for performing inference about equality of mean squared prediction errors (MSPEs) in applications, such as these, that involve a small number of models. We do not have a precise definition of 'small' but, loosely, the idea is that the number of alternative models m is much less than the sample size T.

* Correspondence to: Kenneth D. West, Department of Economics, University of Wisconsin, 1180 Observatory Dr., Madison, WI 53706, USA. E-mail: kdwest@wisc.edu

There are at least two existing procedures. Both use an $m \times 1$ vector whose elements consist of the difference between the MSPE of the null model and the MSPE of one of the alternative models. To test the null of equality of MSPEs across the models, one approach is to conduct a chi-squared test that is the straightforward generalization of the Diebold and Mariano (1995) and West (1996) (DMW) statistic used to compare a pair of models. This chi-squared statistic was used in West *et al*. (1993) and West and Cho (1995). It is referenced in our paper as 'χ^2 that does not adjust MSPE differences' or 'χ^2 (unadj.)'; the reason for the qualification 'unadjusted' will become clear shortly. Under our null hypothesis, however, this statistic is flawed in terms of both size and power. In terms of size: under a reasonable set of technical assumptions, the statistic is unlikely to be well approximated by chi-squared, because the vector of MSPE differences is not centered at zero, even under the null. We explain this point in Section 2 below. In terms of power: as argued by Ashley *et al*. (1980), the alternative in question is one-sided. So even if the statistic is adjusted so as to be centered at zero under the null, a large chi-squared value can reflect extreme behavior in either tail of the underlying distribution, and thus this statistic potentially has poor power.

A second procedure, or perhaps we should say class of procedures, is to obtain critical values on the distribution of the vector of MSPE differences via simulation. One such possibility is White's (2000) reality check. While White (2000) proposed his procedure in the context of applications with many ($m \approx T$) rather than a small ($m \ll T$) number of nested models, the technique has also been applied to small sets of nested models (Hong and Lee, 2003). A possible problem is that White's procedure might not accurately account for dependence of predictions on estimated regression parameters (a key aspect of the computational appeal of White's procedure is that it does not require re-estimation of models during bootstrap repetitions). Under our null, this problem is relevant also to Hansen's (2005) test for superior predictive ability.[1] Alternatively, one could bootstrap in a fashion that includes re-estimation of models (e.g., Rapach and Wohar, 2006). Such a bootstrap has been found to work well (Clark and McCracken, 2006; Clark and West, 2007). Nevertheless, in our own applied work and, we presume, in the applied work of some others, it will at times be desirable to have procedures that do not require repeated re-estimation of models.

In this paper, we develop two closely related procedures for multi-model comparisons in which the alternative models nest a benchmark model. Key features are that we take estimation uncertainty into account, and that we use standard or easily computed critical values. We compare our proposed procedures to the unadjusted chi-squared and White's (2000) reality check via simulations.

Let model 0 denote the benchmark model, and number the alternative models 1 to m. Our main proposal involves two steps: (a) adjust the difference between the MSPE of the benchmark model and each of the alternative models as in Clark and West (2007). The result will be m 'MSPE-adjusted' t-statistics, one of which compares model 0 to model 1, the second of which compares model 0 to model 2, ..., and the last of which compares model 0 to model m. Next, (b) conduct inference on the largest of the m adjusted t-statistics via the distribution of the maximum of correlated normals. In our tables, this is called 'max t-stat (adj.)', where the qualifier 'adj.' signals use of adjusted MSPEs.[2]

[1] Hansen's (2005) test is related to that of White (2000) but has much better power because it standardizes the test statistic and reduces the influence of many 'bad' alternative models (with high MSPE) on the test statistic. Our reference to Hansen assumes that the researcher is using Hansen's procedure to test the null of equal predictive accuracy described here. Hansen's (2005) procedure is intended to test a different null hypothesis, one that involves inequalities rather than strict equality.

[2] We note that while this paper focuses on comparisons for a small number of models, this statistic might also be applied when comparing a large number of models. We leave this as a task for future research.

DOI: 10.1002/jae

In step (a), the adjustment of the MSPE differences is intended to center the vector at zero, under the null. Step (b) respects the one-sided nature of the alternative and is intended to lead to good power. When there are only $m = 2$ alternative models in addition to the benchmark model, as in some of the simulations presented below, critical values for this test vary with a single parameter, namely, the correlation between the two t-statistics. We include a table that presents critical values for 10% and 5% tests, for a crude grid of possible correlations. We supply detailed critical values for a fine grid of correlations in a not-for-publication appendix available from the journal webpage. When the number of alternatives is $m > 2$, critical values for this statistic are easily obtained by a simple procedure: (1) draw many times from an m-dimensional normal distribution whose variance–covariance matrix is set to the sample variance–covariance matrix of the MSPE-adjusted t-statistics; (2) use the quantiles of the maximum of the m correlated values.

Our second proposal is to compute a conventional $\chi^2(m)$ statistic from the $m \times 1$ vector of Clark and West's (2007) MSPE-adjusted values. Since this procedure uses the adjusted differences, we conjecture that it will be well sized. But since it uses both tails of the distribution, it is likely to have less power than does the procedure that considers the maximum of the individual t-statistics. This procedure is denoted 'χ^2 (adj.)' in our tables and is sometimes referenced in our text as 'χ^2 statistics based on the adjusted MSPE differences'.

In our simulations, we find the following, for one-step-ahead predictions: the max t-stat (adj.) statistic is slightly undersized; the χ^2 (adj.) statistic is slightly oversized. The χ^2 statistic used in West *et al*. (1993) and West and Cho (1995)—referenced as 'χ^2 (unadj.) in our tables, because it is computed from the usual rather than from adjusted MSPE differences—is somewhat, and for small sample sizes grossly, oversized; the reality check statistic is somewhat, and for small sample sizes grossly, undersized. In terms of power (not adjusted for size), as expected, max t-stat (adj.) has higher power than the χ^2 (adj.) statistic (although the differences are found not to be large); the χ^2 (adj.) statistic in turn has greater power than either the reality check or the χ^2 (unadj.) statistics (often substantially higher power, as it turns out).

We close our Introduction by noting that we do not attempt to explain or defend the use of out-of-sample analysis. As is usual in out-of-sample analyses, our null is one that could be tested by conventional in-sample tools, in our case by testing whether certain regression coefficients are zero. Out-of-sample analyses may or may not have power relative to in-sample analyses. See, for example, Inoue and Kilian (2004, 2006) and Hansen (2008) for theoretical comparisons of in- and out-of-sample analysis. Our aim is not to argue for out-of-sample analysis, but to supply tools to researchers who have concluded that out-of-sample analysis is informative for the application at hand.

Section 2 motivates our two new procedures. Section 3 gives a precise statement of the environment and the statistics we compute. While the statement is precise, the argument is informal: we do not prove any theorems, but instead refer the reader to other literature. Section 4 gives an overview of our simulation setup. Section 5 presents simulation results. Section 6 presents an empirical example. Section 7 concludes. An appendix, available from the journal webpage, contains some additional simulation results omitted from the paper to save space.

2. OVERVIEW AND INTUITION

We propose two tests to compare a parsimonious benchmark model to a set of $m > 1$ other models that nest the benchmark model. Both tests explicitly take estimation uncertainty into account. A

DOI: 10.1002/jae

key motivation for these two procedures is the following observation in Clark and West (2006, 2007), who developed a test to compare a parsimonious model to a single ($m = 1$) larger model that nests the parsimonious model: under the null that the additional variables in the larger model have coefficients that in population are zero, the more parsimonious model is expected to have a strictly smaller out-of-sample mean squared prediction error (MSPE). This is because the attempt to estimate coefficients whose population values are zero inflates the variance of the prediction error of the larger model.

Figure 1 illustrates the logic spelled out in detail in Clark and West (2006, 2007). The figure depicts some densities of the difference between the MSPE from the null model and the MSPE from an alternative, larger model, or, in self-evident notation, $\hat{\sigma}_0^2 - \hat{\sigma}_1^2$. The alternative model estimates coefficients whose population values are zero. The densities were obtained from 1000 simulations of the AR(1), $m = 2$, data-generating process (DGP) described in the simulations below. The top panel (Figure 1(A)) is one in which the number of predictions P used to construct MSPEs was held constant at 100; the number of observations R used in the rolling sample to compute predictions varied from 40 to 400. All the densities are centered below zero. This is because, on average, the null model has a strictly smaller sample MSPE than does the alternative model. As the regression sample size R increases, the densities shift towards zero. This is because

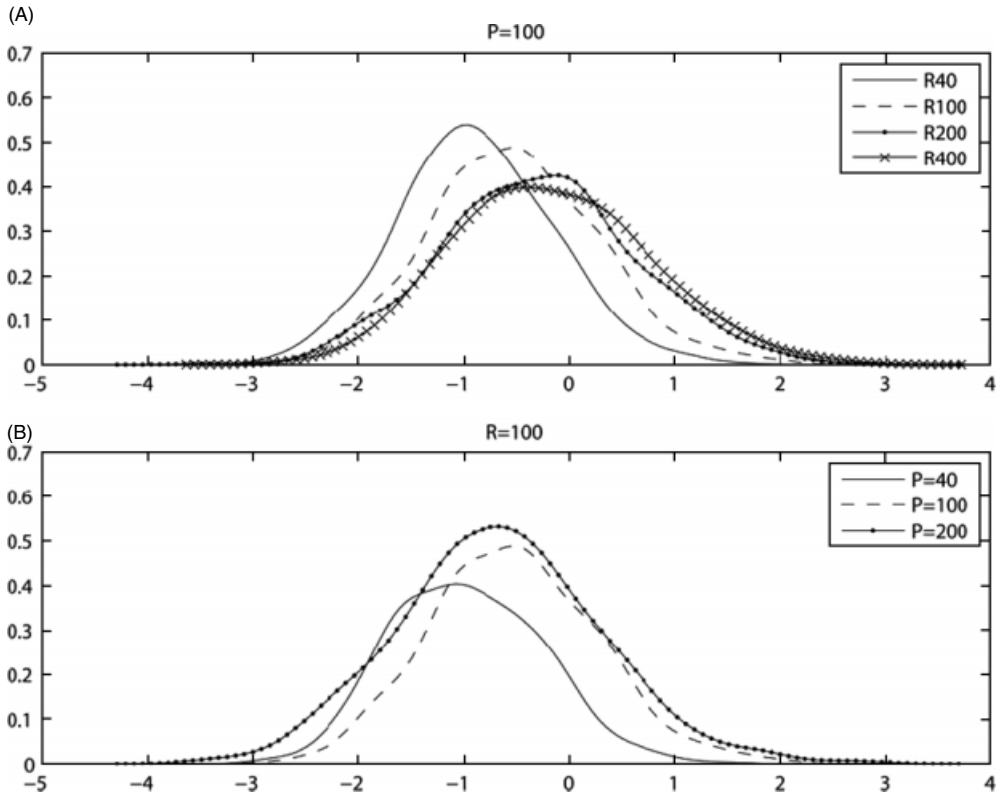

Figure 1. Density of MSPE differences under the null, DGP 1: (A) $P = 100$, R varying; (B) $R = 100$, P varying

DOI: 10.1002/jae

a larger sample typically delivers estimates of coefficients closer to their population values of zero. Hence the inflation of the MSPE in the alternative model diminishes as R increases.

The lower panel is one in which the regression sample size R is held fixed at 100, but the number of predictions P varies from 40 to 200. The difference in MSPEs stays centered at approximately the same value, but the distribution gets tighter and tighter around that value. This is because the law of large numbers causes the difference in MSPEs to pile up at the expected difference in MSPEs.

Clark and West (2006, 2007) proposed adjusting the difference in MSPEs between a pair of models to account for the inflation of the variance of the prediction error of the larger model. This adjustment centers the difference at zero and is intended to produce a test statistic with good size. We will give a precise description of this adjustment in the next section.

It may be shown that by centering the difference in MSPEs at zero, the adjustment transforms the difference in MSPEs into an encompassing statistic (Clark and West, 2007, p. 297). The two ways of writing the statistic—i.e., adjusted difference in MSPEs, or encompassing—are algebraically identical. We prefer the 'adjusted difference in MSPEs' way of writing the statistic because in our view this makes it easier to see how the statistic compares to the conventional DMW statistic for equal MSPE. Readers who prefer the encompassing interpretation should note that one of our contributions is to provide an encompassing test for small model sets, rather than a pairwise one as in previous literature.[3]

3. ECONOMETRIC PROCEDURE

We suppose that there are $m + 1$ models under consideration. Each of the models is to be used to predict a scalar y_t. For expositional clarity, we assume in this section that $m = 2$ and that the forecast horizon is one step ahead. (Generalization to arbitrary m is straightforward. Also, the procedures about to be described extend immediately to multistep forecasts using the direct method, though, as noted below, the theoretical justification for our procedure does not always extend.) Model 0 is a parsimonious benchmark model nested in alternative models 1 and 2. For example, model 0 might be a univariate autoregression in y_t, and models 1 and 2 bivariate and trivariate vector autoregressions in which the right-hand side variables include lags of y_t. Alternatively, model 1 might add lags of a second variable while model 2 instead adds lags of a third variable. Thus, while model 0 is nested in models 1 and 2, model 1 may or may not be nested in model 2 and model 2 may or may not be nested in model 1.

3.1. Mechanics

Write the null and two alternative models as

$$\begin{aligned} y_t &= X_{0t}'\beta_0^* + e_{0t} \\ y_t &= X_{1t}'\beta_1^* + e_{1t} \\ y_t &= X_{2t}'\beta_2^* + e_{2t} \end{aligned} \tag{1}$$

[3] Harvey and Newbold (2000) also propose an encompassing test for small nested model sets, but their approach seems oriented towards a different class of applications. They abstract from noise introduced by estimation of parameters used to make predictions, in both their analysis and simulations. West (2001) shows both analytically and via simulations that in the two-model version of Harvey and Newbold (2000) failure to adjust for such noise causes serious mis-sizing of Harvey and Newbold's proposed statistics. Hence in applications in which forecasts rely on estimated regression parameters, our approach is likely to have distinct advantages relative to Harvey and Newbold.

DOI: 10.1002/jae

By assumption, X_{0t} is a strict subset of X_{1t} and of X_{2t}. Our dating convention allows (indeed, presumes) that for each model X_{it} is observed prior to period t. For example, we might have $X_{0t} = (1\ \ y_{t-1})'$, $X_{1t} = (1\ \ y_{t-1}\ \ y_{t-2})'$, $X_{2t} = (1\ \ y_{t-1}\ \ z_{t-1})'$ for some z that is observed in period $t-1$ (or $X_{2t} = (1\ \ y_{t-1}\ \ y_{t-2}\ \ y_{t-3}\ \ y_{t-4})'$—again, models 1 and 2 may or may not be nested in one another). It is possible that $X_{0t} \equiv 0$, i.e., that the null model presumes that y_t is white noise. The β^*'s are understood to be linear projections, with e_{it} by construction orthogonal to X_{it}. The assumption of linearity is for expositional convenience; methods such as nonlinear least squares are allowed by our test procedures.

Under the null, the coefficients on the additional regressors in X_{1t} and X_{2t} are zero. (In the example just given, this means that the coefficients on y_{t-2} in X_{1t} and on z_{t-1} in X_{2t} are zero.) That is, under the null, $X'_{0t}\beta_0^* = X'_{1t}\beta_1^* = X'_{2t}\beta_2^*$ and $e_{0t} = e_{1t} = e_{2t}$. Under the alternative, at least one of the additional regressors in X_{1t} and/or X_{2t} has a nonzero coefficient. For $i = 0, 1, 2$, let $\sigma_i^2 \equiv Ee_{it}^2$ denote the population variance of the forecast error.[4] We have

$$H_0 : \sigma_0^2 - \sigma_1^2 = 0,\ \ \sigma_0^2 - \sigma_2^2 = 0;\ \ H_A : \max(\sigma_0^2 - \sigma_1^2,\ \ \sigma_0^2 - \sigma_2^2) > 0 \tag{2}$$

Note that the alternative is one-sided. This is in accordance with Ashley *et al*. (1980) and a long list of subsequent studies. If, indeed, one or more of the coefficients in β_1^* or β_2^* are nonzero, then σ_1^2 or σ_2^2 must be less than σ_0^2.

Define the following notation, putting aside for the moment details such as whether a rolling or recursive scheme is used to generate prediction errors:

(a) $\hat{\beta}_{it}$: an estimate of β_i^* computed using period t or earlier data, $i = 0, 1, 2$; (3)
(b) $\hat{y}_{it+1}$: the one-step-ahead forecast from model i ($i = 0, 1, 2$), $\hat{y}_{it+1} = X'_{it+1}\hat{\beta}_{it}$;
(c) $\hat{e}_{it+1}$: one-step-ahead prediction error from model i ($i = 0, 1, 2$), $\hat{e}_{it+1} \equiv y_{t+1} - \hat{y}_{it+1}$;
(d) P: the number of predictions and prediction errors;
(e) $\hat{\sigma}_i^2$: MSPE from model i ($i = 0, 1, 2$), $\hat{\sigma}_i^2 \equiv P^{-1}\sum_t \hat{e}_{it+1}^2$;
(f) $\hat{\sigma}_i^2$-adj.: Clark and West's (2007) adjusted MSPE for models $i = 1, 2$, $\hat{\sigma}_i^2$-adj. $= \hat{\sigma}_i^2 - P^{-1}\sum_t(\hat{y}_{0t+1} - \hat{y}_{it+1})^2$;
(g) $\hat{f}_{it+1} \equiv \hat{e}_{0t+1}^2 - \hat{e}_{it+1}^2 + (\hat{y}_{0t+1} - \hat{y}_{it+1})^2$ ($i = 1, 2$);
(h) $\overline{f}_i$: the adjusted difference in MSPEs between model i ($i = 1, 2$) and model 0, $\overline{f}_i = \hat{\sigma}_0^2 - (\hat{\sigma}_i^2$- adj.) $= P^{-1}\sum_t \hat{f}_{it+1}$;
(i) $\hat{v}_i$: an estimate of a long-run variance computed using autocovariances of $\hat{f}_{it+1}$ ($i = 1, 2$) (typically, for one-step-ahead predictions, $\hat{v}_i$ = sample variance of $\hat{f}_{it+1}$);
(j) $P^{1/2}\overline{f}_i/\sqrt{\hat{v}_i}$: for $i = 1, 2$, the *MSPE-adjusted* t-statistic.

Clark and West (2006, 2007) argue that for the purpose of comparing model 0 to model 1, one can compute the MSPE-adjusted t-statistic $P^{1/2}\overline{f}_1/\sqrt{\hat{v}_1}$ and use standard normal critical values, i.e., one can assume $P^{1/2}\overline{f}_1/\sqrt{\hat{v}_1} \sim_A N(0, 1)$; similarly, one can compare model 0 to model 2 via

[4] The assumption of constant second moments (i.e., the fact that σ is not subscripted by t) is for expositional convenience. We can accommodate moment drift at the expense of complications in notation.

DOI: 10.1002/jae

$P^{1/2}\overline{f}_2/\sqrt{\hat{v}_2} \sim_A N(0, 1)$. This motivates us to assume the following when we conduct inference:

$$P^{1/2}\begin{pmatrix} \dfrac{\overline{f}_1}{\sqrt{\hat{v}_1}} \\ \dfrac{\overline{f}_2}{\sqrt{\hat{v}_2}} \end{pmatrix} \sim_A N(0, \Omega), \Omega \equiv \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{4}$$

Our *first proposed test statistic* is as follows. Let $\hat{z}$ be the larger of the two t-statistics:

$$\hat{z} = \max\lfloor P^{1/2}\overline{f}_1/\sqrt{\hat{v}_1}, P^{1/2}\overline{f}_2/\sqrt{\hat{v}_2}\rfloor \equiv \max t\text{-stat (adj.)} \tag{5}$$

Consider a test at the α level of significance. Let $g_z(z)$ denote the density of the larger of two standard normal random variables with correlation ρ. Let $c_\alpha(\rho)$ be such that $\int_{-\infty}^{c_\alpha(\rho)} g_z(z)\mathrm{d}z = 1 - \alpha$. We propose rejecting the null in favor of the alternative if $\hat{z} > c_\alpha(\hat{\rho})$, where $\hat{\rho}$ is the sample correlation between the two t-statistics $\overline{f}_1/\sqrt{\hat{v}_1}$ and $\overline{f}_2/\sqrt{\hat{v}_2}$.

To use this result requires knowledge of the quantiles of $g_z(z)$. For the case of $m = 2$, the density is presented in Cain (1994) and Ker (2001). We use this density to solve numerically for the value of c such that $\int_{-\infty}^{c} g_z(z)\mathrm{d}z = 0.90$ or $\int_{-\infty}^{c} g_z(z)\mathrm{d}z = 0.95$. Table 1 shows the results. The entries for positive ρ may also be found in Gupta *et al.* (1973). The entries for $\rho = -1$, $\rho = 1$ and $\rho = 0$ are intuitive or familiar. Let z_1 and z_2 denote two standard normal variables. If $\rho = -1$, then $z_1 = -z_2$ and prob[max$(z_1, z_2) > c$] = prob[$z_1 > c$] + prob[$z_1 < -c$], so a 10% critical value is $c = 1.645$ (since prob[$z_1 > 1.645$] + prob[$z_1 < -1.645$] = 0.10). If $\rho = 1$, then $z_1 = z_2$ and prob[max$(z_1, z_2) > c$] = prob($z_1 > c$), so a 10% critical value is $c = 1.282$. If $\rho = 0$, familiar results on order statistics from independent observations tell us that the 10% critical value satisfies $\Phi(c)^2 = 0.90$, yielding the value of $c = 1.632$ given in the table. The critical values fall monotonically as ρ rises, initially with little change, but with an accelerating decline as ρ nears 1.

The second of our two new procedures computes a chi-squared statistic from the vector of adjusted MSPE differences. Define

$$\begin{aligned} &\text{(a) } \hat{f}_{t+1} \equiv (\hat{f}_{1t+1}, \hat{f}_{2t+1})' \equiv (\hat{e}_{0t+1}^2 - \hat{e}_{1t+1}^2 + (\hat{y}_{0t+1} - \hat{y}_{1t+1})^2, \hat{e}_{0t+1}^2 - \hat{e}_{2t+1}^2 \\ &\quad +(\hat{y}_{0t+1} - \hat{y}_{2t+1})^2)'; \\ &\text{(b) } \hat{V} \equiv P^{-1}\textstyle\sum_t(\hat{f}_{t+1} - \overline{f})(\hat{f}_{t+1} - \overline{f})'. \end{aligned} \tag{6}$$

Our *second proposed test statistic* is

$$\chi^2(\text{adj.}) \equiv P\overline{f}'\hat{V}^{-1}\overline{f} \tag{7}$$

Table I. Critical values for the maximum of two correlated standard normals

	ρ										
	1	0.8	0.6	0.4	0.2	0	−0.2	−0.4	−0.6	−0.8	−1
Size = 5%	1.645	1.846	1.900	1.929	1.946	1.955	1.959	1.960	1.960	1.960	1.960
Size = 10%	1.282	1.493	1.556	1.594	1.617	1.632	1.640	1.644	1.645	1.645	1.645

Note: Let z_1 and z_2 be standard normal variables, with correlation ρ. The table presents the 0.95 and 0.90 quantiles for the random variable: $z \equiv \max(z_1, z_2)$.

DOI: 10.1002/jae

We evaluate (7) using $\chi^2(2)$ critical values; the diagonal elements of $\hat{V}$ are $\hat{v}_1$ and $\hat{v}_2$, defined in (3(i)).

We use simulations to compare our two new procedures to two existing procedures: a χ^2 statistic that uses unadjusted MSPEs, and White's (2000) reality check.

- **χ^2 using unadjusted MSPEs.** Use '$\sim$' above a quantity to define MSPE differences that are not adjusted as in Clark and West (2007). This leads to the χ^2 statistic proposed in West and Cho (1995):

$$
\begin{aligned}
&\text{(a) } \tilde{f}_{it+1} : \hat{e}^2_{0t+1} - \hat{e}^2_{it+1} (i = 1, 2); \\
&\text{(b) } \overline{\tilde{f}}_i = \hat{\sigma}^2_0 - \hat{\sigma}^2_i = P^{-1} \textstyle\sum_t \tilde{f}_{it+1} \ (i = 1, 2); \\
&\text{(c) } \tilde{f}_{t+1} \equiv (\tilde{f}_{1t+1}, \tilde{f}_{2t+1})' \equiv (\hat{e}^2_{0t+1} - \hat{e}^2_{1t+1}, \hat{e}^2_{0t+1} - \hat{e}^2_{2t+1})'; \\
&\text{(d) } \overline{\tilde{f}} \equiv (\overline{\tilde{f}}_1, \overline{\tilde{f}}_2)' = P^{-1} \textstyle\sum_t \tilde{f}_{t+1}; \\
&\text{(e) } \tilde{V} \equiv P^{-1} \textstyle\sum_t (\tilde{f}_{t+1} - \overline{\tilde{f}})(\tilde{f}_{t+1} - \overline{\tilde{f}})'; \\
&\text{(f) } \chi^2 \text{ (unadj.)} \equiv P \overline{\tilde{f}}' \tilde{V}^{-1} \overline{\tilde{f}}.
\end{aligned}
\tag{8}
$$

We evaluate (8(f)) using $\chi^2(2)$ critical values. For clarity, we observe that the adjusted and unadjusted MSPE differences are related via: ith adjusted MSPE difference $= i$th unadjusted MSPE difference $+P^{-1} \sum_t (\hat{y}_{0t+1} - \hat{y}_{it+1})^2$.

- **White's (2000) reality check.** When comparing nested models, the null in White (2000) is the one considered here—equality of MSPEs. See the discussion at the end of Section 3.3.

3.2. Mechanics: More Complex Settings

For the case of $m = 2$ alternative models, a not-for-publication appendix available from the journal webpage extends the coarse grid supplied in Table 1 by providing critical values for steps in ρ of 0.01. More generally, for any m, one can compute the p-value of a 'max MSPE-adj. t-statistic' by a simple simulation. One computes m MSPE-adjusted t-statistics and constructs an $m \times m$ matrix $\hat{\Omega}$; here, the i, j element of $\hat{\Omega}$ is the sample correlation between $\hat{\sigma}^2_0 - (\hat{\sigma}^2_i$-adj.) and $\hat{\sigma}^2_0 - (\hat{\sigma}^2_j$-adj.). One then does a series of draws (say, 50,000 draws)[5] on an $N(0, \hat{\Omega})$ random vector and, for each draw, saves the largest of the m elements of that draw's random vector. The p-value for the sample maximum MSPE-adjusted statistic is computed from the distribution of maxima from the simulation.

The statistics defined in (7) and (8(f)) generalize immediately to an environment with $m > 2$.

All the statistics we consider also generalize immediately to multistep forecasts executed using the direct method. $\hat{V}$ (6(b)) and $\tilde{V}$ (8(e)) become estimates of a long-run variance; the diagonal elements of $\hat{V}$ are used in the denominator of the MSPE-adjusted statistics, and the off-diagonal correlations in $\hat{\Omega}$ are computed from the off-diagonal elements of the long-run variance estimate $\hat{V}$. With these changes, the formulas above are applicable.

[5] Gupta *et al*. (1973) provide critical values computed by numerical integration, in the special case when the m statistics are equicorrelated. We found that 50,000 draws were sufficient to match the Gupta *et al*. (1973) critical values to three decimal places.

DOI: 10.1002/jae

3.3. Theoretical Justification

As a formal matter, the max t-stat (adj.) and χ^2 (adj.) procedures require that $m \times 1$ vector $P^{1/2}\overline{f}$ be asymptotically normal with a variance that can be estimated in standard fashion. Under technical conditions such as in Giacomini and White (2006), it is straightforward to show that this holds when (a) the null model posits that y_t is a martingale difference (i.e., $X_{0t} \equiv 0$, $\hat{y}_{0t+1} \equiv 0$ for all t), and (b) rolling samples are used to generate the regression estimates.[6] Under the conditions just stated, asymptotic normality also follows for multistep prediction errors if predictions are made using the direct method.[7]

Alternatively, under the technical conditions of Clark and McCracken (2001), asymptotic normality follows if P is very small relative to the number of observations R used in the first regression sample used to estimate the β^*'s. The precise requirement is that $P/R \to 0$ as the total sample size grows. This result holds for both recursive and rolling samples, and does not require that the null model be a martingale difference. It does require one-step-ahead predictions. Extension to multistep predictions has been worked out only in special cases (Clark and McCracken, 2005).

The conditions of the previous two paragraphs do not by any means span the environment of applications that compare small sets of nested models. But the argument of Clark and West (2007) suggests that the quantiles of the right tail of the t-statistics described above will be approximately those of a standard normal in a wide range of environments. Hence the max t-stat (adj.) procedure should yield tests that are approximately accurately sized. In particular, using numerical methods, Clark and McCracken (2001) have tabulated critical values for the adjusted t-statistic, which they call 'enc-t'. These critical values assume that $P, R \to \infty$. The critical values depend on the limiting value of P/R, on the regression scheme (rolling vs. recursive) and on the number of extra regressors in the larger model (i.e., on the difference between the dimension of X_{1t} and X_{0t} or between X_{2t} and X_{0t}). But apart from a handful of exceptions, for all tabulated values of P/R and the number of extra regressors, the critical values obey the following inequalities: 0.90 quantile $\leq 1.282 \leq$ 0.95 quantile $\leq 1.645 \leq$ 0.99 quantile. For a standard normal, the 0.90 quantile is of course 1.282 and the 0.95 quantile is 1.645. Hence t-tests using standard normal critical values will be somewhat undersized. Our presumption is that the same will apply to the max t-stat (adj.) procedure.

Rationalization of χ^2 (adj.) requires that the quantiles of the left as well as the right tails of the MSPE-adj. statistics are approximately those of a standard normal. Tables of quantiles of the MSPE-adj. t-statistics published on Todd Clark's web page, and additional unpublished tables, indicate that, apart from a handful of cases,

$$0.02 < \text{prob. [square of } t \text{ statistic (adj.)} > 1.96^2] < 0.10 \tag{9}$$
$$0.06 < \text{prob. [square of } t\text{-statistic (adj.)} > 1.645^2] < 0.15$$

The handful of exceptions to the above inequalities would all be eliminated were we to slightly increase the 1.645^2 in the second line to 1.66^2. Hence, were we to apply our χ^2 (adj.) statistic to

[6] To prevent confusion, note that we reference the technical conditions and not the procedures in Giacomini and White (2006).

[7] Suppose the null model relies on estimated regression parameters to predict. Then under the conditions of this paragraph, and for either single or multistep predictions, the vector of adjusted MSPE differences will be asymptotically normal, but possibly not centered at zero. In this case we expect some mis-sizing; see the discussion around equation (4.4) in Clark and West (2007).

DOI: 10.1002/jae

an example with $m = 1$ (which we have not done), we expect the size of tests computed using the standard critical values for a $\chi^2(1)$ to be roughly right.

Under any of the conditions described above, χ^2 (unadj.), the statistic defined in (8(f)) will not be correctly sized. This is because of the miscentering depicted in Figure 1.[8]

We close this section with a brief comparison of our procedure with those proposed in White (2000) and Hansen (2005). We observe first that White's (2000) procedure relies on raw rather than adjusted differences in MSPEs. Hence, whether one follows a bootstrap procedure (as preferred by White and, we believe, others implementing the reality check), or a certain Monte Carlo reality check described by White (2000, p. 1103) (which requires a simulation similar to the one described in the first paragraph of Section 3.2), the resulting test statistic will be different from ours.

A second point is that White's null (in our notation) is $Ef(\beta^*) \leq 0$ (White, 2000, p. 1099). Here, $Ef(\beta^*)$ is the $m \times 1$ vector $(\sigma_0^2 - \sigma_1^2, \sigma_0^2 - \sigma_2^2, \ldots, \sigma_0^2 - \sigma_m^2)'$. White's procedure has been applied to both nested and non-nested model comparisons.[9] But in the context of nested models—which is the relevant context for our paper—the differences in population MSPEs is zero. Thus, in our context White's null simplifies to strict equality: $Ef(\beta^*) = 0$. We note also that in the related context of Hansen (2005) a central innovation relates to handling models whose predictive ability is strictly worse than model 0. But, once again, in our context (nested models and a null of strict equality) this innovation is not relevant for size.[10]

4. SIMULATION OVERVIEW

We completed simulations on two classes of DGP. To conserve space, we report results from only one. This DGP is motivated by the macroeconomic literature on forecasting inflation and generates the predictand via an AR(1) process. A second set of simulations is reported in the Appendix. It is motivated by the finance literature on forecasting changes in asset prices, and assumes the predictand is white noise.

4.1. AR(1) DGP

These DGPs used in our simulations are motivated by the use of disaggregate data to forecast an aggregate (Hubrich, 2005; Hendry and Hubrich, 2006, 2009) in the literature. There is an aggregate y_t that is the sum of several disaggregate series. We report results when 'several' is

[8] The χ^2 (unadjusted) statistic is the multivariate statistic proposed in Giacomini and White (2006). Under their null, the data-generating processes in our simulations will be capturing power rather than size. Our null concerns population parameters. Simulations for the Giacomini and White (2006) null require specifying processes in which finite sample biases in each model lead to equal finite sample performance, which is probably not the case for any of the values of R used in our simulations.

[9] An anonymous referee has pointed out to us that it appears that White's technical conditions rule out nested models. This is an important topic for future research. One path to analytical characterization of White's procedure in nested models is via generalization of McCracken (2007), who considers comparison of nested models when $m = 1$. Simulations of McCracken (2007) in Clark and McCracken (2001), Clark and West (2006) and in McCracken (2007) indicate that his asymptotic approximation can work well. This referee has also pointed out that although White's (2000) procedure has been used in applications with large m, White's technical conditions seem to hold m fixed as $T \to \infty$. Finally, this referee has suggested to us that Hansen (2005) provides an alternative approach to analytical characterization of White's procedure in nested models.

[10] On the subject of related literature: (1)We endorse Hansen's (2005, p. 366) critique of using Bonferroni bounds. (2) Molodtsova and Papell (2008) apply Hansen's (2005) test to a combination of MSPE adjusted according to Clark and West (2006, 2007). This is in the spirit of our proposed procedure.

DOI: 10.1002/jae

three and when 'several' is four. We present algebra here for the simpler case of 'three', with obvious generalization to a larger number of disaggregates. When the aggregate is the sum of three disaggregate components:

$$y_t = y_{1t} + y_{2t} + y_{3t} \tag{10}$$

We consider $m = 2$ alternative models in addition to the benchmark. The benchmark, model 0, is a univariate autoregression in the aggregate y_t; model i, $i = 1, 2$ (or $i = 1, \ldots, 4$ for $m = 4$) adds a lag of y_{it} as a right-hand-side variable:

$$\begin{aligned} y_t &= \text{const.} + \beta^*_{10} y_{t-1} + e_{0t} \equiv X^{'*}_{0t} \beta^*_0 + e_{0t} \\ y_t &= \text{const.} + \beta^*_{1i} y_{t-1} + \beta^*_{2i} y_{it-1} + e_{1t} \equiv X^{'*}_{it} \beta^*_i + e^*_{it}, \;\; i = 1, 2 \end{aligned} \tag{11}$$

We specify the DGPs in terms of the disaggregates. When $m = 2$, we assume that $(y_{1t}, y_{2t}, y_{3t})'$ follows a VAR of order 1 with 3×3 matrix of autoregressive parameters Φ, and zero mean i.i.d. normal disturbances $U_t \equiv (u_{1t} u_{2t} u_{3t})'$:

$$Y_t \equiv (y_{1t}, y_{2t}, y_{3t})' = \mu + \Phi Y_{t-1} + U_t, EU_t U_t' = I_3 \tag{12}$$

Throughout, the mean vector μ was set to $(1, 1, 1)'$.

When examining size properties, we ensure that the three models in (11) have equal MSPE by specifying Φ to be diagonal with common parameter ϕ on the diagonal.[11] That is, each disaggregate follows a univariate AR(1) with common parameter ϕ:

$$\begin{aligned} &y_{it} = 1 + \phi y_{it-1} + u_{it}, |\phi| < 1, \;\; i = 1, 2, 3 \\ &\Rightarrow y_t = 3 + \phi y_{t-1} + e_t, e_t = u_{1t} + u_{2t} + u_{3t}, Ee_t^2 = 3 \end{aligned} \tag{13}$$

As indicated in (13), it follows that y_t also follows an AR(1) with parameter ϕ. The baseline simulations set $\phi = 0.5$. This process is motivated by empirical applications involving aggregate inflation and its disaggregate components.

In (12), the aggregate will be Granger caused by one of the disaggregates once we depart from the specification (13). In simulations reported below, for evaluation of power for the case of $m = 2$, we set

$$\Phi = \begin{pmatrix} 0.5 & -0.6 & 0 \\ -0.4 & 0.3 & 0 \\ 0 & 0 & 0.5 \end{pmatrix} \tag{14}$$

In such a setting, the univariate process for the aggregate y_t is an ARMA(3, 2). The eigenvalues of Φ are 0.5, -0.1 and 0.9. Two components that depend on each other might be commodities and services inflation, while there is a third component (such as food or energy inflation) that shows less interdependence with the other two components on average.

For the size simulations of the five-model comparison ($m = 4$) we extend (13) to include $i = 1, \ldots, 4$ disaggregate components, again with $\phi = 0.5$. For the power simulations in the case

[11] Hendry and Hubrich (2009) show that slope misspecification and estimation uncertainty will dominate forecast accuracy comparisons of different methods to forecast an aggregate.

DOI: 10.1002/jae

of $m = 4$ we set

$$\Phi = \begin{pmatrix} 0.5 & -0.6 & 0 & 0 \\ -0.4 & 0.3 & 0 & 0 \\ 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0.3 \end{pmatrix} \tag{15}$$

The univariate process of the aggregate then follows an ARMA(4, 3) with eigenvalues 0.9, −0.1, 0.3, 0.3.

In all simulations reported here, predictions are made using the regressions given in (11). In simulations related to size, these regression specifications are correct in that the regression disturbance/forecast error is white noise. In simulations related to power, these specifications are incorrect in that the regression disturbance/forecast error is serially correlated.

4.2. Some Details

An overview of our procedures (additional detail is available in a not-for-publication appendix available from the journal webpage): our simulations rely on 1000 replications, with all shocks i.i.d. normal random variables. We report results from rolling samples, for nominal 0.10 tests. Results for recursive samples and for nominal 0.05 tests are reported in the additional appendix. These are qualitatively similar to the results reported here.

We report results for the following four statistics:

(a) Max t-stat (adj.). For $m = 2$ alternative models, we flag rejections by rounding to the nearest tenth and using the critical values such as those given in Table I (the complete set of critical values is in the not-for-publication appendix). For $m = 4$, we use the procedure described in Section 3.2, with critical values determined by 50,000 draws.
(b, c) χ^2 (adj.), χ^2 (unadj.). Rejections flagged using standard χ^2 critical values, e.g., 4.61 for 10% tests when $m = 2$.
(d) Reality check. The bootstrap reality check procedure described by White (2000) was followed, which means in particular that we use unadjusted differences in MSPEs. We used the stationary bootstrap of Politis and Romano (1994), with the parameter that White (2000) calls q set to 0.5. The number of bootstrap repetitions in each simulation sample was 1000.

We report results for 12 combinations of R (rolling regression size/size of smallest recursive sample) and P (number of predictions): $R = 40$, 100, 200 and 400, with each value of R paired with $P = 40$, 100 and 200. These sample sizes for DGP 1 reflect typical monthly and quarterly values for macro data.

5. SIMULATION RESULTS

Table II shows the results of the simulations, for one-step-ahead predictions.

We find that 'max t-stat (adj.)' is modestly undersized. Across the 24 entries in the table, the median value is 0.071, the maximum 0.100 and the minimum 0.043. This is consistent with the modest undersizing predicted by the Clark and McCracken (2005) univariate asymptotics and the simulation results in Clark and West (2007). There does not appear to be a consistent pattern for which entries are closer rather than farther from 0.10. In particular, we do not find

DOI: 10.1002/jae

Table II. Empirical size of nominal 0.10 tests, one-step-ahead predictions

P		$m = 2$				$m = 4$			
		$R = 40$	$R = 100$	$R = 200$	$R = 400$	$R = 40$	$R = 100$	$R = 200$	$R = 400$
40	Max t-stat (adj.)	0.081	0.082	0.085	0.084	0.065	0.072	0.085	0.076
	χ^2 (adj.)	0.119	0.138	0.134	0.109	0.148	0.161	0.185	0.162
	χ^2 (unadj.)	0.157	0.134	0.137	0.116	0.191	0.175	0.187	0.177
	Reality check	0.019	0.039	0.066	0.072	0.013	0.038	0.063	0.068
100	Max t-stat (adj.)	0.073	0.058	0.080	0.065	0.069	0.075	0.063	0.064
	χ^2 (adj.)	0.112	0.109	0.125	0.129	0.114	0.113	0.112	0.133
	χ^2 (unadj.)	0.241	0.147	0.147	0.137	0.299	0.162	0.135	0.134
	Reality check	0.001	0.011	0.036	0.047	0.000	0.017	0.031	0.056
200	Max t-stat (adj.)	0.100	0.069	0.060	0.043	0.084	0.055	0.057	0.062
	χ^2 (adj.)	0.134	0.114	0.098	0.101	0.117	0.091	0.120	0.144
	χ^2 (unadj.)	0.416	0.200	0.122	0.127	0.505	0.210	0.158	0.168
	Reality check	0.000	0.005	0.018	0.024	0.000	0.001	0.011	0.035

Notes:
1. The mean squared prediction error (MSPE) from a null model is compared to MSPEs from m other models. The data are generated according to an AR(1). The alternative models add a lag of a single other variable. The exact form and parameters for the DGPs are described in Section 3 of the paper. In each simulation, one-step-ahead forecasts of y_{t+1} are formed from each of the $m + 1$ models, using least squares regressions.
2. The number of simulations is 1000. R is the size of the rolling regression sample. P is the number of out-of-sample predictions.
3. The qualifier '(adj.)' means that the statistic is computed using MSPE differences adjusted as recommended in Clark and West (2007) and defined in equation (3h); '(unadj.)' means that the usual equation (9b) MSPE difference is used.
4. 'Max t-stat' is the largest of the m Clark and West (2007) MSPE-adjusted t-statistics, and is defined in equation (5). For $m = 2$, Table II reports the fraction of simulations in which each test statistic was greater than the critical value obtained by (a) rounding the sample correlation between the two MSPE-adjusted t-statistics to the nearest 0.1, and (b) using critical values obtained from numerically integrating the density given in (6). Critical values for all other 'max t-stat' entries were obtained by doing the following in each simulation: (a) drawing 50,000 (DGP 1) or 5000 (DGP 2) times from an m-dimensional normal distribution whose variance–covariance matrix was set to the sample variance–covariance matrix of that simulation's MSPE-adjusted t-statistics; (2) using the quantiles of the maximum of the m correlated values.
5. The χ^2 statistics are computed in standard fashion from the $m \times 1$ vector of differences in MSPEs or adjusted difference in MSPEs, see (8f) and (7). For the reality check, White's (2000) bootstrap procedure was used, with 1000 bootstrap repetitions per simulation sample.

consistent improvements as P increases for given R (as is suggested by Clark and West's, 2007, asymptotics) or when P/R is very small (as is suggested by the different Clark and McCracken, 2001, asymptotics). A bit of good news is that size is about the same for larger as for smaller values of m.

While not reported in the table, it is worth mentioning that we also analyzed the maximum t-statistic computed from unadjusted MSPE differences. As in Clark and West (2006, 2007) for $m = 1$, this statistic was quite undersized under the null, and displayed very poor power under the alternative. For example, for $m = 2$, $P = 400$ size, the entries for the four values of R were: 0.0, 0.004, 0.012, 0.015 (as compared to the Table II values of 0.1, 0.069, 0.060, 0.043 for the max t-stat (adj.)).

χ^2 using the adjusted difference in MSPEs (i.e., 'χ^2 (adj.)') is modestly oversized. The median of the 24 values for χ^2 (adj.) is 0.120. Size improves as P gets larger, except for very small R.

χ^2 using the unadjusted difference in MSPEs (i.e., 'χ^2 (unadj.)'): this statistic is clearly oversized in most entries and grossly oversized for $R = 40$. The extreme example of the latter is a size of

DOI: 10.1002/jae

0.505 for $R = 40$, $P = 200$, $m = 4$. For larger values of P and R, χ^2 (unadj.) is slightly oversized relative to χ^2 (adj.).[12] This is consistent with Figure 1A and 1B: for rolling samples, the mis-sizing is worse for smaller R holding P fixed, and it is worse for smaller P holding R fixed. The median of the 24 values for χ^2 (unadj.) is 0.160; all 24 entries are above 0.100.

The reality check is grossly undersized. The median size is 0.022; three of the entries are identically zero, indicating that not a single one of the simulation samples led to a rejection; the largest size (for $R = 400$, $P = 40$, $m = 2$) is 0.072. Performance shows a distinct pattern of improvement as P/R falls (i.e., moving from right to left in a given row in the table, or bottom to top in a given column). The fact that the reality check works better for small values of P/R is consistent with the technical conditions in White (2000), which include the requirement that $P/R \to 0$ at a certain rate. That the reality check is undersized is consistent with the simulations in Hansen (2005) and Clark and McCracken (2006).

Table III shows results for power. This is raw and not size-adjusted power (though we note that under the $P/R \to 0$ asymptotics referenced in Section 3.3 all our statistics are correctly sized asymptotically). As expected, max t-stat (adj.) has greater power than does χ^2 (adj.). Since max t-stat (adj.) is undersized and χ^2 (adj.) is oversized (see Table II), the discrepancy in power would be greater had we reported size-adjusted power. For smaller P or R, the reality check has considerably less power than do the other two statistics. For example, for $R = 40$, power for the reality check is in each case less than half that for max t-stat (adj.) and χ^2 (adj.). Poor power for the reality check was also found in Hansen (2005). Power for χ^2 (unadj.) lies between that for χ^2 (adj.) and the reality check.

We conclude that from the perspective of size, max t-stat (adj.) and χ^2 (adj.) are about comparable, with one being slightly undersized and the other slightly oversized. From the point of view of power, max t-stat (adj.) is slightly preferable.

Table III. Power

P		$m = 2$				$m = 4$			
		$R = 40$	$R = 100$	$R = 200$	$R = 400$	$R = 40$	$R = 100$	$R = 200$	$R = 400$
40	Max t-stat (adj.)	0.648	0.767	0.809	0.832	0.422	0.502	0.559	0.567
	χ^2 (adj.)	0.584	0.651	0.703	0.708	0.394	0.474	0.513	0.530
	χ^2 (unadj.)	0.177	0.252	0.301	0.298	0.198	0.244	0.254	0.280
	Reality check	0.230	0.408	0.478	0.522	0.140	0.256	0.342	0.364
100	Max t-stat (adj.)	0.885	0.983	0.987	0.991	0.672	0.831	0.856	0.876
	χ^2 (adj.)	0.851	0.954	0.966	0.971	0.603	0.781	0.803	0.841
	χ^2 (unadj.)	0.268	0.430	0.519	0.564	0.244	0.297	0.359	0.437
	Reality check	0.314	0.658	0.753	0.766	0.171	0.425	0.532	0.578
200	Max t-stat (adj.)	0.989	0.997	0.999	1.000	0.891	0.962	0.981	0.989
	χ^2 (adj.)	0.986	0.998	0.997	1.000	0.851	0.953	0.984	0.988
	χ^2 (unadj.)	0.465	0.743	0.790	0.814	0.424	0.484	0.554	0.604
	Reality check	0.483	0.900	0.933	0.944	0.269	0.664	0.766	0.799

Notes:
1. See notes to Table II.
2. The VAR parameters of the DGPs are as follows: $m = 2$, equation (14); $m = 4$, equation (15).

[12] In a first round of simulations, we tried computing a robust HAC covariance matrix, as recommended by Giacomini and White (2006). The behavior of χ^2 (unadj.) was similar to what is reported in the table.

DOI: 10.1002/jae

6. FORECASTING AGGREGATE US INFLATION

In this section, we analyze empirically different methods of forecast accuracy evaluation for a small set of models that nest the benchmark, including the tests we proposed. The series we forecast is aggregate inflation, as measured by the US CPI. We present two empirical applications. In the first, we investigate whether including disaggregate inflation components in the aggregate model does improve over forecasting the aggregate only using past aggregate information. The second application also includes models with activity variables.

We focus on one-step-ahead forecasts. We compare forecast accuracy of the different models based on the test procedures we propose and those previously suggested in the literature. We relate the findings to our simulation results. We choose our forecast evaluation periods (pre-1984 and post-1984) to allow us to evaluate not only the predictive content of disaggregate information and/or macroeconomic variables, but also whether there is a difference in the predictive content of disaggregates and macroeconomic variables for aggregate inflation in a low- and a high-inflation regime.

The remainder of the section is organized as follows. Section 6.1 describes the data, while Section 6.2 describes the forecast methods employed. That section also presents details on the transformations used for building the forecast models and for forecast evaluation. Finally, the results of the pseudo out-of-sample forecast experiment based on a rolling estimation sample are discussed.

6.1. Data

The data employed in this study include all items in the US consumer price index as well as its breakdown into four subcomponents: food (p^f), commodities less food and energy commodities (p^c), energy (p^e) and services less energy services prices (p^s). We employ monthly, seasonally adjusted CPI data (source: Bureau of Labor Statistics). In our second application we also employ industrial production and unemployment as predictors. These are two activity variables that are available on a monthly frequency.

We consider a sample period for inflation from 1960(1) to 2004(12), where earlier data from 1959(1) onwards are used for the transformation of the price level. As observed by other authors before, there has been a substantial change in the mean and the volatility of aggregate inflation (e.g., Stock and Watson, 2007), as well as in the disaggregates (see Hendry and Hubrich, 2009) between the two samples. Aggregate as well as component inflation all exhibit high and volatile inflation until the beginning or mid 1980s and lower, more stable inflation rates afterwards. In Table IV we show the substantial reduction in the mean for the disaggregate component inflation

Table IV. US descriptive statistics, year-on-year CPI inflation

	All items	Energy	Commodities	Food	Services
1960–1983					
Mean	4.86	5.91	3.80	4.75	5.81
SD	3.41	8.17	2.89	4.11	3.40
1984–2004					
Mean	2.99	2.28	1.43	2.93	3.91
SD	1.06	8.26	1.65	1.26	0.99

DOI: 10.1002/jae

from the first to the second sample. The mean inflation rate has been reduced from 3.8–5.9% to 1.4–3.9%, while the standard deviation is reduced from between 2.9% and 8.2% to a range of 1.0–8.3%. Thus, also the standard deviation has been reduced substantially, except for energy prices.

We consider two different forecast evaluation periods: 1970(1)–1983(12) and 1984(1)–2004(12). The date 1984 for splitting the sample coincides with estimates of the beginning of the great moderation and is in line with what is chosen in Atkeson and Ohanian (2001) and Stock and Watson (2007). We use the same split sample for comparability of our results to those studies in terms of aggregate inflation forecasts. We use rolling estimation samples.[13]

The out-of-sample forecast evaluation period includes therefore 14 and 21 years for forecast evaluation, respectively. Hendry and Hubrich (2007) report mixed results for simple ADF unit root tests for aggregate and disaggregate inflation, for different samples. For the purpose of illustrating the application of our proposed test procedures, we present empirical results for the level of inflation.

6.2. Forecast Methods and Test Results

Model selection and estimation are carried out for each rolling sample. The models selected are based on the AIC criterion due to the overall favorable forecast accuracy for US inflation (see Stock and Watson, 2007; Hendry and Hubrich, 2009). The forecast evaluation results presented are based on models formulated in first differences of price levels, and forecast accuracy is evaluated based on year-on-year price change forecasts. Hendry and Hubrich (2009) find that formulating the model in terms of month-on-month inflation improves forecast accuracy over formulating the model in year-on-year differences directly.

The 1-month-ahead forecast is based on the following model:

$$\pi^a_{t+1} = \text{const.} + \alpha_1 \pi^a_t + \sum\nolimits_{i=1}^{n} \alpha_{i2} \pi^i_t + e_{t+1} \tag{16}$$

where aggregate inflation π^a_t as growth in prices $(P^a_t - P^a_{t-1})/P^a_{t-1}$ and π^i_t (also specified as growth of prices) are the i subcomponents of inflation (or other macroeconomic variables in the second application) included in the forecasting model.[14] The forecast evaluation is based on a transformation of the resulting forecasts to year-on-year inflation $(\hat{P}^a_{t+1} - P^a_{t+1-12})/P^a_{t+1-12}$. We estimate VARs, but since we only present 1-month-ahead forecasts in the following, (16) presents only the equation for the aggregate from the VAR.

In each of our two empirical applications, we applied our proposed test procedures, i.e., the test based on the maximum of correlated normal random variables (max t-stat (adj.)) and the adjusted chi-squared statistic, using here and throughout a 10% level of significance. We also present the respective critical value for the different tests in each of the applications. We compare the results for the pairwise model comparison based on the t-statistic, again adjusted in line with Clark and West (2006, 2007), with the other tests that compare all the models simultaneously. Additional test results displayed in the table include the unadjusted chi-squared statistic and White's reality check that we have also analyzed in the simulation study. We also present the absolute root mean squared prediction error (RMSPE) for the $\text{AR}_{(p)}$ model and the relative RMSPE for the alternative models.

[13] Results with recursive samples are similar and are omitted to save space.
[14] To prevent confusion, we note that π^a_t plays the role of the variable called y_t in previous sections.

DOI: 10.1002/jae

6.2.1 *Five-Model Comparison: Disaggregate Predictors*

The first application presents a five-model comparison. We compare the $AR_{(p)}$ benchmark model against four different VAR models, where each of the alternative models includes a different disaggregate predictor in addition to lagged aggregate inflation: $VAR^{a,f}_{(p)}$, $VAR^{a,e}_{(p)}$, $VAR^{a,c}_{(p)}$, $VAR^{a,s}_{(p)}$. The setup of the simulations for $m = 4$ is motivated by this application. The different models therefore include disaggregate regressors with very different properties. Energy and food inflation are much more volatile and difficult to forecast than commodities and services inflation (see Table IV). The four alternative models in this example are not nested within one another; only the benchmark model is nested in both alternative models.

The results of this application are presented in Table V. As mentioned, in this example the four alternative models include disaggregate inflation rates as predictors that have quite different properties. When we carry out a pairwise model forecast evaluation using the adjusted t-statistic, we find a rejection of equal forecast accuracy of the benchmark AR model and two alternative models with commodities and services inflation for the high-inflation period. If we compare the

Table V. Tests of equal forecast accuracy, US year-on-year inflation

Method	RMSPE (altern)/RMSPE (bench)	t-stat adj.	Max t-stat adj.	χ^2 adj.	χ^2 unadj.	Reality check
1970–1983						
$AR_{(AIC)}$ (bench)	0.307					
Test AR						
vs. $VAR^{a,f}_{(AIC)}$	1.039	0.666				
vs. $VAR^{a,e}_{(AIC)}$	1.029	0.891				
vs. $VAR^{a,c}_{(AIC)}$	1.016	1.743*				
vs. $VAR^{a,s}_{(AIC)}$	0.986	2.311*				
vs. 4 models			2.311*	7.743	7.207	0.032
Critical value		1.282	1.902	7.78	7.78	0.118
1984–2004						
$AR_{(AIC)}$ (bench)	0.187					
Test AR						
vs. $VAR^{a,f}_{(AIC)}$	0.999	1.860*				
vs. $VAR^{a,e}_{(AIC)}$	1.097	−0.027				
vs. $VAR^{a,c}_{(AIC)}$	1.048	0.290				
vs. $VAR^{a,s}_{(AIC)}$	1.027	−0.463				
vs. 4 models			1.860	3.905	11.926*	0.0007
Critical value		1.282	1.919	7.78	7.78	0.059

Note: Forecast evaluation for 1-month-ahead forecasts; actual RMSPE (non-annualized) for $AR_{(AIC)}$ benchmark model in percentage points, for other models RMSPE relative to AR (RMSPE (altern)/RMSPE (bench)); rolling estimation window; rolling estimation samples 1960(1) to 1970(1), ..., 1983(12) (i.e., $R = 120$ and $P = 168$) and 1960(1) to 1984(1), ..., 2004(12), (i.e., $R = 288$ and $P = 252$); maximum number of lags: $p = 13$; subscripts indicate model selection procedure: AIC, Akaike criterion; superscripts indicate model: $VAR^{a,f}_{(AIC)}$, VAR with lags of aggregate and food inflation, $VAR^{a,e}_{(AIC)}$, VAR with lags of aggregate and energy inflation; $VAR^{a,c}_{(AIC)}$, VAR with lags of aggregate and commodities inflation; $VAR^{a,s}_{(AIC)}$, VAR with aggregate and services inflation; model specification in terms of month-on-month inflation; forecast evaluation for year-on-year inflation; estimated correlation between $f_i = \hat{e}_0 - \hat{e}_i$, for comparing model $i = 1, \ldots, 4$ to the benchmark; critical value of respective test statistic (simulated for max t-stat adj.).

* Significance on a 10% nominal significance level.

DOI: 10.1002/jae

five models simultaneously using the appropriate simulated critical value for correlated normals, we still find that the null of equal forecast accuracy is rejected.[15] The χ^2 (adj.) statistic does not reject but the statistic is close to the critical value. The χ^2 (unadj.) statistic and the reality check do not reject, perhaps because of low power.

Notably, for the low and stable inflation period—where it is usually difficult to improve over a simple AR model—the tests for the pairwise model comparisons presented in the lower panel of Table V indicate predictive content of food inflation for aggregate inflation, but no predictive content of the other disaggregate components. However, the result for the maximum t-test based on the higher critical value simulated for this test statistic based on the maximum of correlated normals does not reject equal forecast accuracy of all models. Here we get a different test result once we take into account that we are comparing five models simultaneously. The only rejection we get is for the χ^2 (unadj.) test—a test that we find to be oversized in our simulations, also for this kind of sample size. Therefore, our results are overall in line with previous findings that during the recent period of low and relatively stable inflation it is difficult for any model to outperform simple benchmark models such as the autoregressive model.

6.2.2 Five-Model Comparison: Disaggregate and Other Macroeconomic Predictors

In the second empirical application, we consider two models with disaggregate predictors, i.e., with services and commodity inflation, and two models that include other macroeconomic variables, and compare those models to the benchmark. One of those four models is a Phillips curve type model, including the change in unemployment as a predictor (in this context the change in unemployment provided lower RMSPE than the level of unemployment). The other is a model with output growth capturing economic activity (see Orphanides and van Norden, 2005, who suggest that using output growth instead of an output gap measure might be useful for forecasting in real time).

In this empirical application, reported in Table VI, all pairwise model comparisons reject equal forecast accuracy for the sample 1970–1983. Furthermore, also our proposed test procedures for the five-model comparison both reject. Only the unadjusted chi-squared statistic and the reality check do not reject, which is likely due to the very low power of those procedures. Overall, we conclude that for this sample period at least one of the alternative models, in particular the one with the largest test statistic, i.e., including unemployment changes, has higher forecast accuracy than the benchmark.

For the sample period 1984–2004 we do find predictive content of unemployment changes for aggregate inflation from the pairwise model comparison, but not for the five-model comparison. From all tests applied for the five-model comparison, again only the unadjusted chi-squared test rejects. We have greater confidence in the results of the other tests on the basis of our simulation results. Therefore, we conclude that equal forecast accuracy of those five models is not rejected, indicating—in line with previous literature—no predictive content of those disaggregates and macroeconomic variables employed here over the information contained in lags of aggregate inflation. This is due to a lack of variability of aggregate inflation to be explained and a lack of predictive content of most explanatory variables.

[15] Consider a three-model ($m = 2$) comparison. We see from Table I that, in that case, if the correlation between the two forecast error differentials of the alternative models and the benchmark is relatively small, the critical value is clearly higher than the one for the adjusted t-statistic for pairwise model comparison. If the correlation is high, then the critical value might not be much higher than the adjusted t-statistic. The Appendix includes simulation results for $m = 2$.

DOI: 10.1002/jae

Table VI. Tests of equal forecast accuracy, US year-on-year inflation

Method	RMSPE (altern)/RMSPE (bench)	t-stat adj.	Max t-stat adj.	χ^2 adj.	χ^2 unadj.	Reality check
1970–1983						
$AR_{(AIC)}$ (bench)	0.307					
Test AR						
vs. $VAR^{a,y}_{(AIC)}$	0.987	2.013*				
vs. $VAR^{a,u}_{(AIC)}$	0.974	3.439*				
vs. $VAR^{a,c}_{(AIC)}$	1.016	1.743*				
vs. $VAR^{a,s}_{(AIC)}$	0.986	2.311*				
vs. 4 models			3.439*	21.762*	2.432	0.061
Critical value		1.282	1.917	7.78	7.78	0.146
1984–2004						
$AR_{(AIC)}$ (bench)	0.187					
Test AR						
vs. $VAR^{a,y}_{(AIC)}$	1.046	−0.047				
vs. $VAR^{a,u}_{(AIC)}$	1.024	1.867*				
vs. $VAR^{a,c}_{(AIC)}$	1.048	0.290				
vs. $VAR^{a,s}_{(AIC)}$	1.027	−0.463				
vs. 4 models			1.867	4.605	12.680*	0.026
Critical value		1.282	1.934	7.78	7.78	0.046

Note: Forecast evaluation for 1-month-ahead forecasts; actual RMSPE (non-annualized) for $AR_{(AIC)}$ benchmark model in percentage points, for other models RMSPE relative to AR (RMSPE (altern)/RMSPE (bench)); rolling estimation window; rolling estimation samples 1960(1) to 1970(1), ..., 1983(12) (i.e. $R = 120$ and $P = 168$) and 1960(1) to 1984(1), ..., 2004(12), (i.e., $R = 288$ and $P = 252$); maximum number of lags: $p = 13$; Subscripts indicate model selection procedure: AIC, Akaike criterion; superscripts indicate model: $VAR^{a,y}_{(AIC)}$, VAR with lags of aggregate inflation and output growth; $VAR^{a,u}_{(AIC)}$, VAR with lags of aggregate inflation and change in unemployment; $VAR^{a,c}_{(AIC)}$, VAR with lags of aggregate and commodities inflation; $VAR^{a,s}_{(AIC)}$, VAR with aggregate and services inflation; model specification in terms of month-on-month inflation; forecast evaluation for year-on-year inflation; estimated correlation between $f_i = \hat{e}_0 - \hat{e}_i$, for comparing model $i = 1, \ldots, 4$ to the benchmark; critical value of respective test statistic (simulated for max t-stat adj.).
* Significance on a 10% nominal significance level.

To conclude, these applications demonstrate that one might draw wrong conclusions on the basis of pairwise model forecast evaluation tests. This is particularly the case if the correlations between the forecast error differentials *vis-à-vis* the benchmark are quite low and the critical value of the maximum t-test is therefore high. Also, this might occur in times of low inflation where the differences in terms of forecast accuracy of alternative models in comparison to the benchmark are rather small.

7. CONCLUSIONS

We have proposed and evaluated two procedures to compare a benchmark model against a small number of alternative models that nest the benchmark. These two procedures, which explicitly account for estimation error in parameters used to make predictions, are easily executed, and do not require bootstrap procedures. Using simulations, we evaluated our procedures and two existing procedures. Our procedures had distinctly better size and power than did the existing procedures.

DOI: 10.1002/jae

On balance, we recommend the procedure that we call 'max t-stat (adj.)'. Our empirical application demonstrates that one might draw wrong conclusions on the basis of pairwise model forecast evaluation tests, and it is therefore important to apply the appropriate statistic for comparing model sets.

We have focused our analysis and discussion on applications with a small number of competing models. But one of our two statistics—the max t-stat (adj.) statistic—might also be applicable to environments in which the number of competing models m is of the same order of magnitude as the sample size. Investigation of that possibility is a priority for future research. A second priority is development of procedures for environments that contain a mixture of nested and non-nested models.

ACKNOWLEDGEMENTS

We thank Roberto Duncan, Eleonora Granziera and Maria Zucca for research assistance, Michael McCracken for supplying the unpublished tables of quantiles referenced in Section 3, and Raffaella Giacomini, Dobrislav Dobrev, participants at the 5th ECB Workshop on Forecasting Techniques 2007, the North American Winter Meeting of the Econometric Society 2009, the Federal Reserve System's Macroeconomics Conference 2009 and seminars at the European University Institute, the Board of Governors of the Federal Reserve System, Carleton University, George Washington University, the Federal Reserve Bank of New York and the Federal Reserve Bank of Philadelphia, two anonymous referees and Tim Bollerslev (the editor) for helpful comments. West thanks the National Science Foundation for financial support. The views expressed here are not necessarily those of the European Central Bank.

REFERENCES

Ashley R, Granger CWJ, Schmalensee R. 1980. Advertising and aggregate consumption: an analysis of causality. *Econometrica* **48**: 1149–1168.

Atkeson A, Ohanian LE. 2001. Are Phillips curves useful for forecasting inflation? *Federal Reserve Bank of Minneapolis Quarterly Review* **25**(1): 2–11.

Billmeier A. 2004. *Ghostbusting: which output gap measure really works?*. IMF Working paper WP/04/146.

Cain M. 1994. The moment generating function of the minimum of bivariate normal random variables. *American Statistician* **48**(2): 124–125.

Clark TE, McCracken MW. 2001. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* **105**: 85–110.

Clark TE, McCracken MW. 2005. Evaluating direct multistep forecasts. *Econometric Reviews* **24**: 369–404.

Clark TE, McCracken MW. 2006. *Reality checks and nested forecast model comparisons*. Manuscript, Board of Governors of the Federal Reserve.

Clark TE, West KD. 2006. Using out-of-sample mean squared prediction errors to test the martingale difference hypothesis. *Journal of Econometrics* **135**(1–2): 155–186.

Clark TE, West KD. 2007. Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics* **138**(1): 291–311.

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

Giacomini R, White H. 2006. Tests of conditional predictive ability. *Econometrica* **74**: 1545–1578.

Gupta SS, Nagel K, Panchapakesan S. 1973. On the order statistics from equally correlated normal random variables. *Biometrika* **60**: 403–413.

Hansen PR. 2005. A test for superior predictive ability. *Journal of Business and Economic Statistics* **23**: 365–380.

DOI: 10.1002/jae

Hansen PR. 2008. *In-sample fit and out-of-sample fit: their joint distribution and its implications for model selection*. Manuscript, Stanford University.
Harvey D, Newbold P. 2000. Tests for multiple forecast encompassing. *Journal of Applied Econometrics* **15**: 471–482.
Hendry DF, Hubrich K. 2006. *Forecasting aggregates by disaggregates*. European Central Bank Working Paper 589.
Hendry DF, Hubrich K. 2009. *Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate*. *Journal of Business and Economic Statistics*, accepted for publication, DOI:10.1198/jbes.2009.07112.
Hong Y, Lee TH. 2003. Inference on predictability of foreign exchange rates via generalized spectrum and nonlinear time series models. *Review of Economics and Statistics* **85**: 1048–1062.
Hubrich K. 2005. Forecasting euro area inflation: does aggregating forecasts by HICP component improve forecast accuracy? *International Journal of Forecasting* **21**: 119–136.
Ker A. 2001. On the maximum of bivariate normal variables. *Extremes* **4**(2): 185–186.
Inoue A, Kilian L. 2004. In-sample or out-of-sample tests of predictability: which one should we use? *Econometric Reviews* **23**(4): 371–402.
Inoue A, Kilian L. 2006. On the selection of forecasting models. *Journal of Econometrics* **130**(2): 273–306.
Molodtsova T, Papell DH. 2008. "Out-of-Sample Exchange Rate Predictability with Taylor Rule Fundamentals." Manuscript, University of Houston.
McCracken MW. 2007. Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics* **140**: 719–752.
Orphanides A, Norden S van. 2005. The reliability of inflation forecast based on output gap estimates in real time. *Journal of Money, Credit, and Banking* **37**: 583–600.
Politis DN, Romano JP. 1994. The stationary bootstrap. *Journal of the American Statistical Association* **89**: 1303–1313.
Rapach DE, Wohar ME. 2006. In-sample vs. out-of-sample tests of stock return predictability in the context of data mining. *Journal of Empirical Finance* **13**(2): 231–247.
Sarno L, Thornton DL, Valente G. 2005. Federal funds rate prediction. *Journal of Money, Credit and Banking* **37**: 449–472.
Stock JH, Watson MW. 2007. Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking* **39**(s1): 3–33.
West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* **64**: 1067–1084.
West KD. 2001. Tests of forecast encompassing when forecasts depend on estimated regression parameters. *Journal of Business and Economic Statistics* **19**: 29–33.
West KD, Cho D. 1995. The predictive ability of several models of exchange rate volatility. *Journal of Econometrics* **69**: 367–391.
West KD, Edison HJ, Cho D. 1993. A utility based comparison of some models of exchange rate volatility. *Journal of International Economics* **35**: 23–46.
White H. 2000. A reality check for data snooping. *Econometrica* **68**: 1097–1126.