Bagging or Combining (or Both)? An Analysis Based on Forecasting U.S. Employment Growth

[Forthcoming, *Econometric Reviews*]

Short title: Bagging or Combining (or Both)?

David E. Rapach*
Department of Economics
Saint Louis University
E-mail: rapachde@slu.edu

Jack K. Strauss
Department of Economics
Saint Louis University
E-mail: strausjk@slu.edu

September 10, 2007 (Revised)

Abstract

Forecasting a macroeconomic variable is challenging in an environment with many potential predictors whose predictive ability can vary over time. We compare two approaches to forecasting U.S. employment growth in this type of environment. The first approach applies bootstrap aggregating (bagging) to a general-to-specific procedure based on a general dynamic linear regression model with 30 potential predictors. The second approach considers several methods for combining forecasts from 30 individual autoregressive distributed lag (ARDL) models, where each individual ARDL model contains a potential predictor. We analyze bagging and combination forecasts at multiple horizons over four different out-of-sample periods using a mean square forecast error (MSFE) criterion and forecast encompassing tests. We find that bagging forecasts often deliver the lowest MSFE. Interestingly, we also find that incorporating information from both bagging and combination forecasts based on principal components often leads to further gains in forecast accuracy.

Keywords Bagging; Combination forecasts; Employment; Forecast encompassing; Principal components

JEL Classification C22, C52, C53, E24

*Corresponding author: Department of Economics, Saint Louis University, 3674 Lindell Boulevard, St. Louis, MO 63108-3397; E-mail: rapachde@slu.edu; Phone: 314-977-3601; Fax: 314-977-1478. The authors thank two anonymous referees for very helpful comments. The authors also gratefully acknowledge financial support from the Simon Center for Regional Forecasting at Saint Louis University. The results reported in this paper were generated using GAUSS 6.1. The GAUSS program files are available at http://pages.slu.edu/faculty/rapachde/Research.htm.

1 Introduction

Forecasting a macroeconomic variable is especially challenging when there are a plethora of potential predictors whose predictive ability can vary markedly over time. In an application in this type of environment—forecasting U.S. inflation—Inoue and Kilian (2007) apply the bootstrap aggregating (bagging) methodology developed by Breiman (1996) to dynamic linear regression models. They consider a situation where a researcher begins with a general model for inflation that contains a large number of potential predictors. A pre-testing, or general-to-specific, procedure then removes predictors that are "unimportant" according to a decision rule, and a forecast is formed using a re-estimated model that only includes the "important" variables. Inoue and Kilian (2007) recognize that bagging provides a way of generating more reliable inflation forecasts when the decision rule is unstable. Intuitively, as emphasized by Breiman (1996), bootstrapping provides new learning sets for the decision rule—instead of relying on a single historical realization—which can reduce its instability and thereby improve predictive performance. Inoue and Kilian (2007) find that bagging generally improves U.S. inflation forecast accuracy, suggesting that bagging offers a promising tool for generating reliable forecasts for macroeconomic variables in situations involving pre-testing with many potential predictors.

The application of bagging to forecasting a macroeconomic time series with many potential predictors is limited to Inoue and Kilan (2007).[1] A more sizable strand of the recent literature points to the usefulness of combining methods as a means of forecasting inflation and output growth in the presence of a large number of potential predictors. This research is perhaps best exemplified by the recent work of Stock and Watson (1999, 2003, 2004). They find that combination forecasts formed by taking a weighted average of the individual forecasts generated by autoregressive distributed lag (ARDL) models, each based on a potential predictor, provide an effective way of dealing with the substantial model uncertainty and structural instability that can plague individual forecasting models based on a single predictor or small set of predictors. Stock and Watson (1999, 2003, 2004) emphasize the consistent ability of simple combination forecasts, such as the mean or median of individual ARDL model forecasts, to improve upon autoregressive (AR) benchmark model forecasts, despite the inability of individual ARDL models to consistently outperform an AR benchmark model in terms of a mean square forecast error (MSFE) metric. In addition to simple averaging, other combining methods, such as those based on a small number of principal components extracted from the individual forecasts (Figlewski, 1983; Chan, Stock, and Watson,

[1] Lee and Yang (2006) use bagging techniques to develop binary and quantile forecasts of financial variables.

1999; Stock and Watson, 2004) or clusters of individual forecasts (Aiolfi and Timmermann, 2006), appear to perform well in forecasting macroeconomic variables.[2]

Both bagging and forecast combining procedures offer ways of generating forecasts in realistic environments where there are a wealth of potential predictors and the predictive power of individual variables changes over time. In light of this, it is natural to ask which approach, bagging or combining, produces superior forecasts. In the present paper, we address this in the context of forecasting U.S. employment growth. More specifically, we compare forecasts of U.S. employment growth generated by the Inoue and Kilian (2007) bagging-augmented general-to-specific procedure to combination forecasts formed from individual ARDL model forecasts. Our focus on U.S. employment growth is motivated by the fact that it is among the most closely watched economic statistics by both the popular and financial press and analysts of all stripes.[3] Rapach and Strauss (2006) find that certain forecast combining methods are able to consistently and substantially improve upon an AR benchmark model in terms of MSFE when forecasting U.S. employment growth over various out-of-sample periods. We evaluate bagging and combination forecasts in terms of MSFE and the Harvey, Leybourne, and Newbold (1998) test of forecast encompassing. The forecast encompassing tests allow us to investigate whether the bagging forecasts contain information useful for forecasting beyond that contained in the combination forecasts (and vice versa).

We consider forecasts of U.S. employment growth based on 30 potentially relevant predictors, including the Conference Board's ten leading indicators, various labor market statistics, variables related to manufacturing orders and output, and a number of financial and price variables. We generate recursive simulated out-of-sample forecasts of U.S. employment growth using the Inoue and Kilian (2007) bagging-augmented general-to-specific procedure based on a general model that includes all 30 potential predictors. We also compute recursive simulated out-of-sample forecasts for 30 individual ARDL models, where each individual ARDL model includes one of the potential predictors. We then form combination forecasts based on the individual ARDL model forecasts using several combining methods, including simple methods (mean, median, trimmed mean), as well as methods based on discount MSFE (Stock and Watson, 2004), clusters, and principal components. Rapach and Strauss (2006) find that these combining methods are able to consistently outperform

[2] See Timmermann (2006) for a comprehensive review of forecast combining methods.

[3] For example, employment growth—in the context of the so-called "jobless" recovery from the most recent U.S. recession—received considerable attention during the 2004 presidential election, arguably more than any other economic variable. Employment growth is also viewed as a key indicator of labor market activity and crucial in Federal Reserve formulations of interest rate policy, and hence forecasts of employment growth are important in ruminations about Fed policy actions.

an AR benchmark model with respect to forecasting U.S. employment growth.[4] In order to explore robustness, we consider four different out-of-sample periods, including a 1995:04–2005:03 (1985:04–2005:03) out-of-sample period covering the last decade (two decades) of our full sample and out-of-sample periods centered on the last two NBER-dated U.S. recession troughs in March of 1991 and November of 2001.

Previewing our results, we find that both bagging and combination forecasts based on a small number of principal components extracted from the the individual ARDL models are able to generate sizable decreases in MSFE relative to an AR benchmark model. Bagging forecasts produce average declines in MSFE of 11%, 30%, 37%, and 7% at horizons of 1, 3, 6, and 12 months, respectively, relative to an AR benchmark model over the four out-of-sample periods. Combination forecasts based on principal components lead to corresponding reductions in MSFE of 12%, 26%, 24%, and 18%. Combination forecasts based on simple averaging, discount MSFE, and clusters yield smaller reductions in MSFE on average, between 3%–8%. Interestingly, forecast encompassing tests show that there are a number of instances where the bagging and combination forecasts fail to encompass each other, indicating that both sets of forecasts contain information useful for forecasting U.S. employment growth not contained in the other. Overall, our results, as well as those in Inoue and Kilian (2007), indicate that bagging warrants further attention in the macroeconomic forecasting literature. Our results also point to potential gains from using bagging and combination forecasts in conjunction.

The rest of the paper is organized as follows: Section 2 describes the econometric methodology, Section 3 reports the empirical results, and Section 4 concludes.

2 Econometric Methodology

2.1 Construction of the Bagging Forecasts

Define $\Delta y_t = y_t - y_{t-1}$, where y_t is the log-level of U.S. employment at time (month) t. In addition, define $y^h_{t+h} = (1/h)\sum_{j=1}^{h} \Delta y_{t+j}$, so that y^h_{t+h} is the (approximate) monthly growth rate of U.S. employment from time t to $t+h$, where h is the forecast horizon. Let $x_{i,t}$ denote one of n potential predictors of U.S. employment growth (so that $i = 1, ..., n$). We consider 30 potential predictors of U.S. employment growth ($n = 30$) in our analysis.

[4]Note that Inoue and Kilian (2007) also compare bagging forecasts of U.S. inflation to combination forecasts computed from individual models. However, they do not consider methods for combining individual model forecasts based on discount MSFE, clusters, or principal components, and these are the combining methods that perform the best with respect to forecasting U.S. employment growth in Rapach and Strauss (2006).

We compute bagging forecasts of U.S. employment growth at horizon h using the bagging-augmented pre-testing procedure of Inoue and Kilian (2007). The procedure begins with the general model:

$$y_{t+h}^{h} = \mu + \sum_{j=0}^{q_1-1} \theta_j \Delta y_{t-j} + \sum_{i=1}^{n} \delta_i x_{i,t} + \xi_{t+h}^{h}, \tag{1}$$

where ξ_{t+h}^{h} is an error term characterized by autocorrelation of degree $h-1$. Suppose we are interested in forming a forecast of y_{t+h}^{h} at time t. The pre-testing procedure involves estimating (1) via OLS using data from the start of the available sample through time t and computing the t-statistics corresponding to each of the potential predictors, where q_1 is selected using the SIC.[5] The $x_{i,t}$ variables with t-statistics less than 1.645 in absolute value are dropped from (1), and the model is re-estimated.[6] The forecast of y_{t+h}^{h} is generated by plugging the included $x_{i,t}$ values along with the Δy_{t-j} $(j = 0, ..., q_1 - 1)$ values into the re-estimated version of (1) and setting the error term equal to its expected value of zero.

Bagging can be implemented for the pre-testing procedure via a moving-block bootstrap. More specifically, a large number (B) of pseudo-samples of size t for the left-hand-side and right-hand-side variables in (1) are generated by randomly drawing blocks of size m (with replacement) from the observations of these variables available from the beginning of the sample through time t. For each pseudo-sample, we estimate (1) using the pseudo-data and OLS, the pre-testing procedure determines the predictors to include in the forecasting model, the model is re-estimated using the pseudo-data, and a forecast of y_{t+h}^{h} is formed by plugging the actual included $x_{i,t}$ values and Δy_{t-j} $(j = 0, ..., q_1 - 1)$ values into the re-estimated version of the forecasting model (and again setting the error term equal to its expected value of zero). The bagging model forecast corresponds to the average of the B forecasts for the bootstrapped pseudo-samples.[7]

Dividing the complete available sample of T observations for Δy_t and $x_{i,t}$ $(i = 1, ..., n)$ into an in-sample portion comprised of the first R observations and an out-of-sample period comprised of the last P observations, we can form a series of $P - (h - 1)$ recursive simulated out-of-sample forecasts using the bagging procedure.[8] We denote this series by $\{\hat{y}_{BA,t+h|t}^{h}\}_{t=R}^{T-h}$.

[5]The t-statistics for the OLS estimates of δ_i in (1) are computed using Newey and West (1987) heteroskedasticity and autocorrelation consistent (HAC) standard errors based on a lag truncation of $h-1$.

[6]Inoue and Kilian (2007) consider a range of critical values. We obtain similar results using other conventional critical values such as 1.96.

[7]Following Inoue and Kilian (2007), we use $m = h$ and $B = 100$.

[8]"Recursive" indicates that the forecasts are generated using an expanding estimation window. The out-of-sample forecasts are "simulated"—as opposed to "real-time"—because they are based on revised data and not on the data actually available at the time of forecast formation. Real-time forecasting exercises are not feasible in the present paper, as data vintages are not readily available for all of the variables we consider in our forecasting exercise. We follow much of the macroeconomic forecasting literature, including Stock and Watson (1999, 2003, 2004), in analyzing

2.2 Construction of the ARDL Model Forecasts

Each ARDL model takes the form:

$$y_{t+h}^h = \alpha + \sum_{j=0}^{q_1-1} \beta_j \Delta y_{t-j} + \sum_{j=0}^{q_2-1} \gamma_j x_{i,t-j} + \epsilon_{t+h}^h, \tag{2}$$

where ϵ_{t+h}^h is an error term. We construct recursive simulated out-of-sample forecasts for y_{t+h}^h at time t for a given predictor $x_{i,t}$ (denoted by $\hat{y}_{i,t+h|t}^h$) using (2). More specifically, $\hat{y}_{i,t+h|t}^h$ is computed by plugging Δy_{t-j} $(j = 0, ..., q_1 - 1)$ and $x_{i,t-j}$ $(j = 1, ..., q_2 - 1)$ into (2) with the parameters set equal to their OLS estimates based on data available from the start of the sample through period t and ϵ_{t+h}^h set equal to its expected value of zero. The lag lengths in (2) are selected using the SIC, data through period t, a minimum lag length of 0 for q_1 (1 for q_2 to ensure that $x_{i,t}$ appears in (2)), and a maximum lag length of 6 for q_1 and q_2.[9] By continuing in this manner through the end of the out-of-sample period, we generate a series of $P - (h - 1)$ recursive simulated out-of-sample forecasts for the ARDL model containing the predictor $x_{i,t}$, $\{\hat{y}_{i,t+h|t}^h\}_{t=R}^{T-h}$. Note that we select q_1 and q_2 anew when computing each out-of-sample forecast, so that the lag lengths for the ARDL forecasting equations can vary through time.

We also compute recursive simulated out-of-sample forecasts for an AR model, (2) with the restriction $\gamma_j = 0$ $(j = 0, ..., q_2 - 1)$ imposed. The AR model is a common benchmark model when forecasting time-series variables. The AR model forecasts are computed in a manner similar to that described above for the ARDL model forecasts, and the lag length (q_1) for the AR model is selected using the SIC and a minimum (maximum) lag length of 0 (6).

2.3 Forecast Combining Methods

We consider a number of forecast combining methods that perform well relative to an AR benchmark model in forecasting U.S. employment growth in Rapach and Strauss (2006). Some of the combining methods require a holdout period to calculate the weights used to combine the individual ARDL model forecasts, and we use the first P_0 observations from the out-of-sample period as the initial holdout period. The combination forecasts of y_{t+h}^h made at time t, $\hat{y}_{CB,t+h|t}^h$, typically are a linear combination of the individual ARDL model forecasts:

$$\hat{y}_{CB,t+h|t}^h = \sum_{i=1}^{n} w_{i,t} \hat{y}_{i,t+h|t}^h, \tag{3}$$

simulated out-of-sample forecasts.

[9] Our results are not very sensitive to the maximum lag lengths.

where $\sum_{i=1}^{n} w_{i,t} = 1$. When the weights, $\{w_{i,t}\}_{i=1}^{n}$, are estimated, we use the individual out-of-sample forecasts and y_{t+h}^{h} observations available from the start of the holdout out-of-sample period to time t. For each of the combining methods, we compute combination forecasts over the post-holdout out-of-sample period. This leaves us with a total of $P_h = P - (h-1) - P_0$ combination forecasts, $\{\hat{y}_{CB,t+h|t}^{h}\}_{t=R+P_0}^{T-h}$, available for evaluation.

2.3.1 Simple Combining Methods

We consider three simple combining methods: the mean, median, and trimmed mean. Stock and Watson (1999, 2003, 2004) find that simple combining methods work well in forecasting inflation and output growth using a large number of potential predictors. The mean combination forecast simply involves setting $w_{i,t} = 1/n$ $(i = 1, ..., n)$ in (3), while the median combination forecast is the sample median of $\{\hat{y}_{i,t+h|t}^{h}\}_{i=1}^{n}$. The trimmed mean combination forecast sets $w_{i,t} = 0$ for the individual forecasts with the smallest and largest forecasts at time t, and $w_{i,t} = 1/(n-2)$ for the remaining individual forecasts in (3). Obviously, the simple combining methods do not require a holdout out-of-sample period.

2.3.2 Discount MSFE Combining Methods

Stock and Watson (2004) consider a combining method where the weights in (3) are a function of the recent historical forecasting performance of the individual ARDL models. Their discount MSFE (DMSFE) combining method uses the following weights:

$$w_{i,t} = m_{i,t}^{-1} / \sum_{j=1}^{n} m_{j,t}^{-1}, \tag{4}$$

where

$$m_{i,t} = \sum_{s=R}^{t-h} \psi^{t-h-s} (y_{s+h}^{h} - \hat{y}_{i,s+h|s}^{h})^2 \tag{5}$$

and ψ is a discount factor. Observe that when $\psi = 1$, there is no discounting, and (4) produces the optimal combination forecast derived by Bates and Granger (1969) for the case where the individual forecasts are uncorrelated. When $\psi < 1$, greater importance is attached to the recent forecasting accuracy of the individual ARDL models. We consider ψ values of 1.0 and 0.9.

2.3.3 Cluster Combining Methods

Aiolfi and Timmermann (2006) develop conditional combining methods that incorporate persistence in forecasting performance. We use their $C(K, PB)$ algorithm. The initial combination forecast (for $y_{(R+P_0)+h}^{h}$) is computed by grouping the individual ARDL model forecasts over the initial holdout

out-of-sample period, $\{\hat{y}^h_{i,s+h|s}\}_{s=R}^{R+(P_0-1)-(h-1)}$ $(i = 1, ..., n)$, into K equal-sized clusters based on MSFE, with the first cluster containing the individual ARDL models with the lowest MSFE values, the second cluster containing the ARDL models with the next lowest MSFE values, and so on. The first combination forecast is the average of the individual ARDL model forecasts of $y^h_{(R+P_0)+h}$ in the first cluster. In forming the second combination forecast, we compute the MSFE for the individual ARDL model forecasts, $\{\hat{y}^h_{i,s+h|h}\}_{s=R+1}^{R+(P_0-1)-(h-1)+1}$ $(i = 1, ..., n)$, and again group the individual forecasts into K clusters. The second combination forecast is the average of the individual ARDL model forecasts of $y^h_{(R+P_0+1)+h}$ included in the first cluster. We proceed in this manner through the end of the available out-of-sample period. Observe that the clusters are formed by computing MSFE using a rolling window. Following Aiolfi and Timmermann (2006), we consider $K = 2$ and $K = 3$ in our applications.[10]

2.3.4 Principal Component Combining Methods

Chan, Stock, and Watson (1999) and Stock and Watson (2004) consider generating a combination forecast utilizing the first m principal components of the individual ARDL model forecasts. Let $\hat{F}^h_{1,s+h|s}, ..., \hat{F}^h_{m,s+h|s}$ for $s = R, ..., t$ denote the first m principal components of the uncentered second-moment matrix of the individual ARDL model forecasts, $\hat{y}^h_{i,s+h|s}$ $(i = 1, ..., n;\ s = R, ..., t)$. To form a combination forecast of y^h_{t+h} at time t based on the fitted principal components, we estimate the following regression model:

$$y^h_{s+h} = \phi_1 \hat{F}^h_{1,s+h|s} + ... + \phi_m \hat{F}^h_{m,s+h|s} + \nu^h_{s+h}, \tag{6}$$

where $s = R, ..., t - h$. The combination forecast is given by $\hat{\phi}_1 \hat{F}^h_{1,t+h|t} + ... + \hat{\phi}_m \hat{F}^h_{m,t+h|t}$, where $\hat{\phi}_1, ..., \hat{\phi}_m$ are the OLS estimates of $\phi_1, ..., \phi_m$, respectively, in (6). We use the IC_{p3} information criterion developed by Bai and Ng (2002) to select m (considering a maximum value of four) when calculating combination forecasts using the principal component (PC) method. Bai and Ng (2002) show that familiar information criteria such as the AIC and SIC do not always consistently estimate the true number of factors, and they develop alternative criteria that can consistently estimate the true number of factors under more general conditions. In extensive Monte Carlo simulations, Bai and Ng (2002) find that the IC_{p3} criterion performs well in selecting the correct number of factors for sample sizes close to ours.[11]

[10] Observe that the number of clusters serves to define the size of the first cluster, as none of the other clusters are used in generating the forecast. The greater the number of clusters, the smaller the size of the first cluster.

[11] Using the taxonomy in Huang and Lee (2007), all of the combining methods we consider are classified as "combination of forecasts," while the bagging forecasts—which are formed by including all of the variables in a single

2.4 Forecast Encompassing Tests

We use forecast encompassing tests to compare the information content in the BA model and combination forecasts.[12] Consider forming an optimal forecast of y^h_{t+h} as a convex combination of a combination forecast and the BA forecast:

$$\hat{y}^h_{OPT,t+h|t} = \lambda_{CB}\hat{y}^h_{CB,t+h|t} + \lambda_{BA}\hat{y}^h_{BA,t+h|t}, \tag{7}$$

where $\lambda_{CB} + \lambda_{BA} = 1$. If $\lambda_{CB} = 0$, then the BA forecasts encompass the combination forecasts, as the combination forecasts do not contribute any useful information—apart from that already contained in the BA forecasts—in the formation of the optimal composite forecast. Alternatively, if $\lambda_{CB} > 0$, then the BA forecasts do not encompass the combination forecasts, so that the combination forecasts do contain information—beyond that contained in the BA forecasts—useful in the formation of the optimal composite forecast.

Harvey, Leybourne, and Newbold (1998) develop statistics based on the approach of Diebold and Mariano (1995) to test the null hypothesis that $\lambda_{CB} = 0$ against the one-sided (upper-tail) alternative hypothesis that $\lambda_{CB} > 0$. Let $\hat{u}^h_{k,t+h|t} = y^h_{t+h} - \hat{y}^h_{k,t+h|t}$ $(k = BA, CB)$ denote the forecast error associated with $\hat{y}^h_{k,t+h|t}$ and define

$$\hat{d}^h_{t+h|t} = (\hat{u}^h_{BA,t+h|t} - \hat{u}^h_{CB,t+h|t})\hat{u}^h_{BA,t+h|h}. \tag{8}$$

The Harvey, Leybourne, and Newbold (1998) statistic is given by

$$HLN_h = [\hat{V}(\bar{d^h})]^{-1/2}\bar{d^h}, \tag{9}$$

where $\bar{d^h} = P_h^{-1}\sum_{t=R+P_0}^{T-h}\hat{d}^h_{t+h|t}$, $\hat{V}(\bar{d^h}) = P_h^{-1}(\hat{\omega}_0 + 2\sum_{j=1}^{h-1}\hat{\omega}_j)$, and $\hat{\omega}_j = P_h^{-1}\sum_{t=R+P_0+j}^{T-h}(\hat{d}^h_{t+h|t} - \bar{d^h})(\hat{d}^h_{(t-j)+h|(t-j)} - \bar{d^h})$. HLN_h is asymptotically distributed as a standard normal variate under the null hypothesis that $\lambda_{CB} = 0$. Harvey, Leybourne, and Newbold (1998) recommend using a modified version of the HLN_h statistic, which we denote by $MHLN_h$,

$$MHLN_h = \left[\frac{P_h + 1 - 2h + P_h^{-1}h(h-1)}{P_h}\right] HLN_h, \tag{10}$$

general model—are classified as "combination of information." Another combination of information procedure that could be used is the Stock and Watson (2002) diffusion index, which involves extracting principal components from the potential predictors themselves (instead of extracting principal components from the individual ARDL forecasts). We experimented with diffusion index forecasts and found that the PC combining method forecasts generally perform better. It would be interesting in future research to consider the approach of Armah and Swanson (2007), who adapt the methodology of Bai and Ng (2006) to a forecasting environment and select individual predictors to serve as proxies for estimated principal components in a modified diffusion index forecasting model.

[12]The notion of forecast encompassing is developed in, *inter alia*, Granger and Newbold (1973) and Chong and Hendry (1986). See Clements and Hendry (1998) for a textbook treatment of forecast encompassing.

and the Student's t distribution with $P_h - 1$ degrees of freedom to test the null hypothesis that $\lambda_{CB} = 0$ (recall that $P_h = P - (h-1) - P_0$ and represents the number of out-of-sample forecasts available for evaluation). We use the $MHLN_h$ statistic in our applications in Section 3 below.[13]

In an analogous manner, we can use the $MHLN_h$ statistic based on $\hat{d}^h_{t+h|t} = (\hat{u}^h_{CB,t+h|t} - \hat{u}^h_{BA,t+h|t})\hat{u}^h_{CB,t+h|h}$ to test the null hypothesis that the combination forecasts encompass the BA forecasts ($\lambda_{BA} = 0$) against the alternative hypothesis that the combination forecasts do not encompass the BA forecasts ($\lambda_{BA} > 0$).

3 Empirical Results

3.1 Data

Our full sample is based on monthly data for 1960:01–2005:03. We measure U.S. employment growth as the first differences of the log-levels of seasonally adjusted civilian U.S. employment (multiplied by 100) from the Conference Board's Labor Force, Employment, and Unemployment data base. The individual predictors are comprised of 30 series from the Conference Board. We use the ten leading indicators:

- *Average weekly hours in manufacturing* (weekly manufacturing hours)
- *Average weekly initial claims for unemployment insurance* (unemployment claims)
- *Manufacturers' new orders for consumer goods and materials* (in chained 1982 dollars; manufacturing new orders)
- *Vendor performance*[14]
- *Manufacturers' new orders of nondefense capital goods* (in chained 1982 dollars; manufacturing capital orders)
- *Building permits for new private housing units* (building permits)
- *S&P 500 stock price index* (S&P 500 index)

[13] A word of caution is in order with respect to the use of the $MHLN_h$ statistic in making inferences on the relative information content across forecasting models. Recent research demonstrates that a number of issues—such as the size of the in-sample period relative to the out-of-sample period, type of estimation window (for example, fixed, rolling, or recursive), and whether the models are nested or non-nested—can affect the asymptotic distribution of the test statistics; see Corradi and Swanson (2006) for an informative review of these issues. We recognize that, strictly speaking, all of the conditions required for the validity of the asymptotic distribution may not be met in our applications, so that inferences based on the $MHLN_h$ statistic serve as a rough guide to statistical significance.

[14] Vendor performance is an index that measures how quickly companies receive deliveries from their suppliers. An increase in the index means that it is taking longer for companies to receive deliveries.

- *Real M2 money supply* (in chained 2000 dollars; real M2)
- *Interest rate spread* (10-year Treasury bond yield minus the federal funds rate)
- *Consumer confidence*

We also consider nine variables from the Conference Board's Labor Force, Employment, and Unemployment series:

- *Civilian labor force*
- *Labor force participation rate for males 20 and over* (male labor force part. rate)
- *Labor force participation rate for females 20 and over* (female labor force part. rate)
- *Labor force participation rate for people 16–19 years of age* (teenage labor force part. rate)
- *Average weekly overtime hours in manufacturing* (weekly overtime hours)
- *Index of help wanted advertising in newspapers* (index of help wanted advertising)
- *Ratio of help wanted advertising to number of unemployed* (advertising-unemployment ratio)
- *Civilian unemployment rate* (unemployment rate)
- *Average duration of unemployment in weeks* (average unemployment duration)

We include another four variables from other Conference Board categories:

- *Manufacturing unfilled orders*
- *Index of industrial production* (industrial production)
- *Manufacturing and trade sales* (in chained 2000 dollars)
- *Personal income* (in chained 2000 dollars)

Finally, we include a number of additional financial and price variables:

- *Average prime rate charged by banks* (prime rate)
- *Commercial and industrial loans outstanding* (in chained 2000 dollars; commercial and industry loans)
- *Consumer installment credit outstanding* (consumer credit outstanding)

- *10-year government bond yield* (10-year interest rate)
- *Consumer price index* (CPI)
- *Producer price index* (PPI)
- *Consumer price index, health* (CPI health)

In an effort to use stationary predictors in (1) and (2), we take the first differences of log-levels for all predictors, with the following exceptions: we use levels for unemployment claims, vendor performance, building permits, interest rate spread, consumer confidence, index of help wanted advertising, unemployment rate, and average unemployment duration; in addition, we use first differences for weekly manufacturing hours, male labor force participation rate, female labor force participation rate, teenage labor force participation rate, weekly overtime hours, advertising-unemployment ratio, prime rate, and 10-year interest rate.

We consider four out-of-sample periods for the evaluation of U.S. employment growth forecasts. The first out-of-sample period covers the last decade of the full sample, 1995:04–2005:03. This is an interesting period for evaluation, as it includes much of the prolonged expansion of the 1990s, 2001 recession, and subsequent "jobless" recovery. To analyze the robustness of the results, we include a "long" out-of-sample period comprised of the last 20 years of the full sample, 1985:04–2005:03. We also consider two out-of-sample periods centered on the two most recent NBER-dated business cycle troughs in 2001:11 and 1991:03 and that include the 30 months before and after the trough. This allows us to focus on periods surrounding important turning points in U.S. employment. As discussed above, we require a holdout out-of-sample period in order to compute most of the combination forecasts, and we use the seven years (84 months) before the beginning of the out-of-sample periods given above as the initial holdout out-of-sample period.

3.2 Individual ARDL Model Forecasts

Table 1 reports the MSFE for the benchmark AR model forecasts, along with the ratio of the MSFE for the individual ARDL model forecasts to the MSFE for the AR benchmark forecasts, at horizons of 1, 3, 6, and 12 months over each out-of-sample period. While all 30 of the variables are plausible predictors of U.S. employment growth, the forecasting performance of the individual ARDL models is mixed: many of the MSFE ratios are above unity (indicating that the ARDL model forecasts are less accurate than the AR benchmark forecasts) and many are below unity (indicating that the ARDL forecasts outperform the AR benchmark). A serious problem in identifying reliable predictors in Table 1 is that the relative forecasting performance of a given predictor can change substantially over time. An extreme example is real M2, whose MSFE ratios are 1.67, 1.24, and 1.57 for $h = 12$ and the 1995:04–2005:03, 1985:04–2005:03, and 1999:05–2004:05 out-of-sample periods, respectively, but 0.67 for the 1988:09–1993:09 out-of-sample period. In addition, the relative forecast accuracy of a particular predictor can vary markedly over the different forecast horizons, even for a given out-of-sample period. Overall, the results in Table 1 indicate that it will be difficult to identify *a priori* the particular variable or small set of variables that are the most relevant for forecasting U.S. employment growth for a given period from among the large number of plausible predictors. Bagging models and forecast combining methods offer ways of incorporating and culling information from a large number of potential predictors in this environment.

3.3 Bagging and Combination Forecasts

Table 2 reports MSFE ratios (relative to the AR benchmark model) and encompassing test results for the BA model and combination forecasts over the 1995:04–2005:03 out-of-sample period. Panel A shows that the MSFE for the BA forecasts is lower than the MSFE for the AR benchmark forecasts by 9% at the 1-month horizon, and the BA forecasts have a lower MSFE than all of the combination forecasts, with the exception of the PC forecasts, at this horizon. At horizons of 3 and 6 months in Panels B and C, the BA forecasts reduce the MSFE by 30% and 28%, respectively, relative to the AR benchmark model forecasts, and the BA forecasts are substantially more accurate than all of the combination forecasts at these horizons.[15] However, Panel D of Table 2 shows that the accuracy of the BA forecasts deteriorates at the 12-month horizon, as the MSFE is 26% higher than the AR benchmark MSFE and well above the MSFE for each of the combination forecasts. The PC forecasts deliver the lowest MSFE ratio at the 12-month horizon in Table 2.

We turn next to the encompassing test results in Table 2. The numerous rejections in Panel A for $h = 1$ show that the BA forecasts do not encompass any of the combination forecasts and that none of the combination forecasts encompass the BA forecasts. BA and combination forecasts thus each contain information useful for forecasting U.S. employment growth beyond that contained in the other, so that neither the BA nor combination forecasts can be judged superior in terms of their relative information content. Panels B and C show that the BA forecasts encompass all of the combination forecasts at horizons of 3 and 6 months, while the combination forecasts

[15] Nevertheless, all of the MSFE ratios for the combination forecasts are below unity, indicating that they outperform the AR benchmark model. It is also worth noting that the MSFE ratio for the BA forecasts is lower than the MSFE ratios for all of the individual ARDL model forecasts in Table 1 over the 1995:04–2005:03 out-of-sample period at the 3- and 6-month horizons.

do not encompass the BA forecasts at these horizons, indicating that the BA forecasts contain more information. At the 12-month horizon in Panel D, the BA forecasts encompass all of the combination forecasts, with the exception of the PC method, and all of the combination forecasts encompass the BA forecasts. The inability of the BA forecasts to encompass the PC forecasts and ability of the PC forecasts to encompass the BA forecasts points to greater information content in the PC relative to the BA forecasts at the 12-month horizon.

Table 3 reports forecasting results for the BA and combination forecasts over the "long" 1985:04–2005:03 out-of-sample period. The results for the 1-month horizon in Table 3 are similar to the corresponding results for the 1995:04–2005:03 out-of-sample period in Table 2: the MSFE ratios are below unity for the BA and all of the combination forecasts, the PC forecasts produce the lowest MSFE, and the BA and combination forecasts each fail to encompass the other. At horizons of 3 and 6 months, all of the MSFE ratios are below unity (again matching the corresponding results in Table 2). The PC forecasts have the lowest MSFE at the 3- and 6-month horizons, followed by the BA forecasts. There is an interesting contrast in terms of the encompassing test results in Tables 2 and 3 at the 3- and 6-month horizons. In Table 2, the BA forecasts encompass all of the combination forecasts at these horizons, while the combination forecasts do not encompass the BA forecasts; in Table 3, the BA and combination forecasts fail to encompass each other, meaning that the BA forecasts no longer have a distinct information advantage at the 3- and 6-month horizons. At the 12-month horizon in Table 3, the BA forecasts have an MSFE ratio above unity (as in Table 2), while all of the combination forecasts have an MSFE ratio below unity, with the PC forecasts displaying the lowest MSFE ratio (0.84). The BA and combination forecasts, with the exception of the PC forecasts, each fail to encompass the other. The PC forecasts encompass the BA forecasts (while the BA forecasts do not encompass the PC forecasts), indicating information advantages in the PC forecasts relative to the BA forecasts at the 12-month horizon.

Tables 4 and 5 report forecasting results for out-of-sample periods centered on the two most recent U.S. business cycle troughs in 2001:11 and 1991:03, respectively. Continuing the pattern in Tables 2 and 3, the BA and combination forecasts each fail to encompass the other at the 1-month horizon in Tables 4 and 5. Matching the results in Table 2 at horizons of 3 and 6 months, the BA forecasts encompass all of the combination forecasts, while none of the combination forecasts encompass the BA forecasts, over the 2001 recession in Table 4. These results also hold over the 1991 recession in Table 5, with the exception that the BA forecasts do not encompass the PC forecasts at the 3-month horizon. The BA forecasts have the lowest MSFE ratio at the 3- and

6-month horizons in Table 4; the PC (BA) forecasts have the lowest MSFE ratio at the 3-month (6-month) horizon in Table 5. The accuracy of the BA forecasts improves notably in Tables 4 and 5 (especially Table 5) relative to Tables 2 and 3 at the 12-month horizon: in Tables 2 and 3, the MSFE ratios for the BA forecasts are well above unity; in contrast, they are well below unity in Tables 4 and 5. The BA and combination forecasts encompass each other at the 12-month horizon in Table 4, so that the BA and combination forecasts do not display significant information advantages over each other. However, the BA forecasts encompass the combination forecasts at the 12-month horizon in Table 5, while the combination forecasts fail to encompass the BA forecasts, indicating significant information advantages for the BA forecasts.

The results across the four out-of-sample periods reported in Tables 2–5 can be summarized as follows:

- The BA and all of the combination forecasts are able to consistently outperform an AR benchmark model in terms of MSFE at the 1-month horizon, with the BA and PC forecasts producing the largest reductions in MSFE, ranging from approximately 10%–14%. The BA and combination forecasts are unable to encompass each other at the 1-month horizon, signifying that the BA and combination forecasts each contain information useful for forecasting U.S. employment growth not found in the other.

- The BA forecasts are typically more accurate than the combination forecasts at horizons of 3 and 6 months, while the PC forecasts are the most accurate of the combination forecasts at these horizons. Encompassing tests also often point to information advantages in the BA forecasts relative to the combination forecasts at horizons of 3 and 6 months.

- The BA model's forecasting performance is uneven at the 12-month horizon, as it has an MSFE well above that of the AR benchmark in some cases and well below in others. The PC combination forecasts perform consistently well at the 12-month horizon, with reductions in MSFE of 12%–28% relative to the AR benchmark forecasts.

Our final forecasting exercise is motivated by the encompassing test results, which show that there are a number of cases where the BA and PC forecasts are unable to encompass one another, implying that each forecast contains information relevant for forecasting U.S. employment growth not contained in the other. In light of this, we consider forming a forecast as a simple average of the BA and PC forecasts, and the results are reported in Table 6.[16] In all of the cases where the

[16]For the cases where both of the $MHLN_h$ statistics are significant, the weights on the BA and PC forecasts in

BA and PC forecasts fail to encompass one another, the average of the BA and PC forecasts has a lower MSFE than either of the individual BA and PC forecasts. Further, in the cases where at least one of the forecasts encompasses the other, the average of the BA and PC forecasts still has an MSFE that is typically near that of the individual BA or PC forecast with the lowest MSFE. With one exception (the 1995:04–2005:03 out-of-sample period and $h = 12$), the average of the BA and PC forecasts leads to a reduction in MSFE of 11%–50% relative to the AR benchmark model. From a practical standpoint, using information from both the bagging and combination forecasts appears to be a fruitful approach to forecasting U.S. employment growth.[17]

4 Conclusion

Bagging-augmented pre-testing based on a general dynamic linear regression model and combining forecasts from a large number of individual ARDL models are two approaches to forecasting macroeconomic variables in realistic environment with many potential predictors. In the present paper, we compare these two approaches with respect to forecasting U.S. employment growth. We find that bagging forecasts are often more accurate according to an MSFE metric than a variety of combination forecasts (as well as AR benchmark model forecasts). The performance of bagging forecasts in the present paper, as well as in Inoue and Kilian (2007), indicates that bagging warrants further attention in the macroeconomic forecasting literature. Our results also show that combination forecasts based on principal components perform consistently well, and there are a number of cases where bagging forecasts and combination forecasts based on principal components each contain information useful for forecasting not found in the other. This suggests that bagging and combination forecasts can be profitably used in tandem, and, indeed, we find further gains in forecast accuracy for forecasts formed by taking a simple average of the bagging forecasts and combination forecasts based on principal components. It would be interesting in future research to investigate whether using bagging and combination forecasts in conjunction leads to forecasting gains for other macroeconomic variables.

Tables 2–5 are close to 0.50, so taking the mean of the two forecasts is reasonable. This procedure also avoids having to estimate the weights, making it easy to implement in practice.

[17]We examined the robustness of our results along a number of dimensions and obtained similar results. For example, the results are very similar when we use the AIC instead of the SIC to select the lag lengths in (1) and (2). We also computed combination forecasts for a set of potential predictors that excludes manufacturing capital orders and manufacturing and trade sales, two variables that are available with a one-month lag relative to the other potential predictors (and so are not "coincident" with the other predictors). We again obtain very similar results. The complete results for these robustness checks are available at http://pages.slu.edu/faculty/rapachde/Research.htm.

References

Aiolfi, M., Timmermann, A. (2006). Persistence in Forecasting Performance and Conditional Combination Strategies. *Journal of Econometrics* 135(1-2):31–53.

Armah, N.A., Swanson, N.R. (2007). Seeing Inside the Black Box: Using Diffusion Index Methodology to Construct Factor Proxies in Largescale Macroeconomic Time Series Environments. Manuscript, Rutgers University.

Bai J., Ng, S. (2002). Determining the Number of Factors in Approximate Factor Models. *Econometrica* 70(1):191–221.

Bai, J. Ng, S. (2006). Evaluating Latent and Observed Factors in Macroeconomics and Finance. *Journal of Econometrics* 113(1-2):507–537.

Bates, J.M., Granger, C.W.J. (1969). The Combination of Forecasts. *Operational Research Quarterly* 20(4):451–468.

Breiman, L. (1996). Bagging Predictors. *Machine Learning* 24(2):123–140.

Chan, Y.L., Stock, J.H., Watson, M.W. (1999). A Dynamic Factor Model Framework for Forecast Combination. *Spanish Economic Review* 1(2):91–121.

Chong, Y.Y., Hendry, D.F. (1986). Econometric Evaluation of Linear Macro-Economic Models. *Review of Economic Studies* 53(4):671–690.

Clements, M.P., Hendry, D.F. (1998). *Forecasting Economic Time Series.* Cambridge: Cambridge University Press.

Corradi, V., Swanson, N.R. (2006). Predictive Density Evaluation. In Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), *Handbook of Economic Forecasting.* Amsterdam: Elsevier, pp. 197–284.

Diebold, F.X., Mariano, R.S. (1995). Comparing Predictive Accuracy. *Journal of Business and Economic Statistics* 13(3):253–263.

Figlewski, S. (1983). Optimal Price Forecasting Using Survey Data. *Review of Economics and Statistics* 65(1):13–21.

Granger, C.W.J., Newbold, P. (1973). Some Comments on the Evaluation of Economic Forecasts. *Applied Economics* 5(1):35–47.

Harvey, D.I., Leybourne, S.J., Newbold, P. (1998). Tests for Forecast Encompassing. *Journal of Business and Economic Statistics* 16(2):254–259.

Huang, H., Lee, T.-H. (2007). To Combine Forecasts or to Combine Information? Manuscript, University of California, Riverside.

Inoue, A., Kilian, L. (2007). How Useful is Bagging in Forecasting Economic Time Series? A Case Study of U.S. CPI Inflation. *Journal of the American Statistical Association*, forthcoming.

Lee, T.-H., Yang, Y. (2006). Bagging Binary and Quantile Predictors for Time Series. *Journal of Econometrics* 135(1-2):465–497.

Newey, W.K., West, K.D. (1987). A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica* 55(3):703–708.

Rapach, D.E., Strauss, J.K. (2006). Forecasting U.S. Employment Growth Using Forecast Combining Methods. *Journal of Forecasting*, forthcoming.

Stock, J.H., Watson, M.W. (1999). Forecasting Inflation. *Journal of Monetary Economics* 44(2):293–335.

Stock, J.H., Watson, M.W. (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics* 20(2):147–162.

Stock, J.H., Watson, M.W. (2003). Forecasting Output Growth and Inflation: The Role of Asset Prices. *Journal of Economic Literature* 41(3):788–829.

Stock, J.H., Watson, M.W. (2004). Combination Forecasts of Output Growth in a Seven-Country Data Set. *Journal of Forecasting* 23(6):405–430.

Timmermann, A. (2006). Forecast Combinations. In Elliott, G., Granger, C.W.J., Timmermann, A. (Eds.), *Handbook of Economic Forecasting.* Amsterdam: Elsevier, pp. 135–196.

TABLE 1 Individual ARDL model forecasting results

	A. 1995:04–2005:03 out-of-sample period				B. 1985:04–2005:03 out-of-sample period			
Predictor	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 1$	$h = 3$	$h = 6$	$h = 12$
AR benchmark	9.13	2.99	1.79	1.14	8.75	2.77	1.68	1.19
Weekly manufacturing hours	1.00	1.00	1.00	1.10	1.00	1.00	1.00	1.09
Unemployment claims	0.94	0.84	0.81	0.92	0.96	0.90	0.89	0.99
Manufacturing new orders	0.98	0.87	0.87	0.84	0.97	0.97	1.00	0.98
Vendor performance	0.98	0.95	0.97	1.00	1.00	0.98	1.04	1.06
Manufacturing capital orders	1.00	0.99	0.98	0.99	1.04	1.04	0.99	0.99
Building permits	1.12	1.36	1.57	1.62	1.01	1.09	1.13	1.11
S&P 500 index	1.00	0.96	0.93	0.82	1.01	1.00	1.01	0.99
Real M2	1.05	1.19	1.38	1.67	1.02	1.10	1.17	1.24
Interest rate spread	1.02	1.11	1.18	1.29	1.03	1.21	1.39	1.64
Consumer confidence	1.04	1.07	1.14	1.38	1.00	1.00	1.03	1.16
Civilian labor force	0.98	0.96	0.99	1.01	1.00	1.00	1.02	1.02
Male labor force part. rate	1.00	0.89	0.93	0.95	0.99	0.92	0.91	0.95
Female labor force part. rate	1.00	1.09	1.14	1.13	1.02	1.08	1.12	1.11
Teenager labor force part. rate	0.99	1.00	1.01	1.00	0.99	1.01	1.02	1.01
Weekly overtime hours	0.96	0.93	1.09	0.93	1.01	1.02	1.17	1.12
Index of help wanted advertising	0.92	0.89	0.95	0.87	0.94	0.84	0.89	0.84
Advertising-unemployment ratio	1.07	1.04	1.06	1.18	1.06	1.13	1.10	1.17
Unemployment rate	0.98	0.93	0.88	0.79	1.00	0.97	0.94	0.89
Average unemployment duration	1.02	1.05	1.09	1.14	1.01	1.04	1.07	1.14
Manufacturers unfilled orders	0.98	0.99	0.96	1.00	0.99	1.00	1.01	1.03
Industrial production	0.93	0.82	0.77	0.86	0.96	0.90	0.92	0.97
Manufacturing and trade sales	1.02	0.94	0.90	0.89	1.02	0.97	0.97	0.97
Personal income	0.99	0.97	0.98	1.02	1.01	1.05	1.10	1.09
Prime rate	1.00	0.99	1.11	1.12	1.00	1.00	1.07	1.09
Commercial and industry loans	1.01	1.02	1.07	1.23	1.02	1.06	1.23	1.45
Consumer credit outstanding	1.01	1.02	1.04	1.06	1.01	1.02	1.04	1.04
10-year interest rate	1.10	1.10	1.05	0.99	1.05	1.16	1.19	1.11
CPI	1.00	0.96	1.02	1.36	0.99	0.98	1.04	1.22
PPI	1.01	1.06	1.16	1.26	1.02	1.04	1.08	1.14
CPI health	1.00	1.01	1.09	1.20	1.00	1.02	1.06	1.15

Notes: The entries for the AR benchmark model report the MSFE; the other entries report the MSFE for the ARDL model forecasts indicated on the left to the MSFE for the AR benchmark model forecasts.

TABLE 1 (continued)

Predictor	C. 1999:05–2004:05 out-of-sample period				D. 1988:09–1993:09 out-of-sample period			
	$h = 1$	$h = 3$	$h = 6$	$h = 12$	$h = 1$	$h = 3$	$h = 6$	$h = 12$
AR benchmark	13.36	4.63	3.10	2.23	8.66	2.89	2.28	2.30
Weekly manufacturing hours	1.00	1.00	0.99	1.04	1.00	1.01	1.01	1.14
Unemployment claims	0.96	0.83	0.77	0.87	1.04	1.07	1.10	1.08
Manufacturing new orders	0.99	0.86	0.82	0.80	0.99	1.31	1.33	1.17
Vendor performance	0.98	0.98	0.98	1.00	1.04	1.13	1.19	1.15
Manufacturing capital orders	0.95	0.95	0.97	0.99	1.05	1.03	1.01	0.99
Building permits	1.16	1.41	1.53	1.52	0.84	0.74	0.53	0.62
S&P 500 index	1.00	0.92	0.82	0.69	1.00	0.99	0.98	1.03
Real M2	1.06	1.17	1.33	1.57	0.96	0.85	0.76	0.67
Interest rate spread	1.02	1.11	1.12	1.19	1.06	1.46	1.67	1.61
Consumer confidence	1.01	1.04	1.06	1.26	0.93	0.87	0.83	0.96
Civilian labor force	0.98	0.90	0.91	0.96	1.05	1.07	1.07	1.03
Male labor force part. rate	1.01	0.88	0.91	0.93	1.00	0.88	0.85	0.99
Female labor force part. rate	1.00	1.06	1.12	1.13	1.08	1.18	1.09	1.04
Teenager labor force part. rate	0.99	1.00	1.01	1.00	1.00	1.02	1.02	1.01
Weekly overtime hours	0.97	0.91	1.08	0.84	1.12	1.33	1.46	1.39
Index of help wanted advertising	0.82	0.69	0.78	0.72	0.82	0.59	0.63	0.57
Advertising-unemployment ratio	1.10	1.08	1.07	1.18	1.04	1.37	1.23	1.23
Unemployment rate	0.96	0.87	0.80	0.73	1.01	1.04	1.03	0.97
Average unemployment duration	1.02	1.04	1.05	1.05	1.01	1.03	1.05	1.04
Manufacturers unfilled orders	0.98	0.99	0.97	1.00	0.99	1.06	1.06	1.06
Industrial production	0.93	0.78	0.71	0.81	1.02	1.07	1.14	1.07
Manufacturing and trade sales	1.00	0.84	0.75	0.79	1.07	1.08	1.04	1.02
Personal income	0.98	0.89	0.92	0.92	1.04	1.19	1.06	0.84
Prime rate	1.00	0.99	1.14	1.13	1.00	1.00	1.04	1.13
Commercial and industry loans	1.01	1.03	1.08	1.26	1.02	1.04	1.09	1.29
Consumer credit outstanding	1.00	1.02	1.04	1.07	1.00	1.00	0.97	0.95
10-year interest rate	0.98	0.93	0.92	0.96	1.03	1.34	1.37	1.23
CPI	1.01	0.96	1.01	1.27	1.00	1.00	1.00	1.05
PPI	1.01	1.04	1.10	1.17	1.04	1.02	0.99	1.03
CPI health	1.00	1.01	1.05	1.15	1.00	1.06	1.04	1.12

TABLE 2 Forecasting and encompassing test results for the bagging model and combining methods, 1995:04–2005:03 out-of-sample period

		H_0: BA encompasses CB			H_0: CB encompasses BA		
Combining method	MSFE ratio	$\hat{\lambda}_{CB}$	$MHLN_h$	p-value	$\hat{\lambda}_{BA}$	$MHLN_h$	p-value
A. $h = 1$							
BA model	0.90						
Mean	0.97	0.35	2.42^{**}	0.01	0.65	3.54^{**}.	0.00
Median	0.99	0.32	2.38^{**}	0.01	0.68	3.77^{**}	0.00
Trimmed mean	0.98	0.34	2.39^{**}	0.01	0.66	3.67^{**}	0.00
DMSE, $\psi = 1.00$	0.97	0.35	2.42^{**}	0.01	0.65	3.54^{**}	0.00
DMSE, $\psi = 0.90$	0.97	0.35	2.41^{**}	0.01	0.65	3.55^{**}	0.00
$C(2, PB)$	0.97	0.34	2.34^{*}	0.01	0.66	3.59^{**}	0.00
$C(3, PB)$	0.96	0.36	2.33^{*}	0.01	0.64	3.43^{**}	0.00
PC	0.87	0.59	2.79^{**}	0.00	0.41	2.14^{*}	0.02
B. $h = 3$							
BA model	0.71						
Mean	0.93	0.13	0.79	0.22	0.87	3.30^{**}	0.00
Median	0.96	0.11	0.71	0.24	0.89	3.37^{**}	0.00
Trimmed mean	0.93	0.15	0.94	0.17	0.85	3.43^{**}	0.00
DMSE, $\psi = 1.00$	0.93	0.13	0.78	0.22	0.87	3.27^{**}	0.00
DMSE, $\psi = 0.90$	0.92	0.13	0.78	0.22	0.87	3.29^{**}	0.00
$C(2, PB)$	0.91	0.13	0.70	0.24	0.87	3.05^{**}	0.00
$C(3, PB)$	0.90	0.13	0.73	0.23	0.87	2.87^{**}	0.00
PC	0.85	0.12	0.74	0.23	0.88	3.83^{**}	0.00
C. $h = 6$							
BA model	0.73						
Mean	0.95	0.13	0.37	0.36	0.87	2.84^{**}	0.00
Median	0.97	0.12	0.37	0.36	0.88	2.77^{**}	0.00
Trimmed mean	0.96	0.14	0.43	0.33	0.86	3.01^{**}	0.00
DMSE, $\psi = 1.00$	0.95	0.13	0.36	0.36	0.87	2.79^{**}	0.00
DMSE, $\psi = 0.90$	0.93	0.15	0.40	0.35	0.85	2.80^{**}	0.00
$C(2, PB)$	0.93	0.13	0.35	0.36	0.87	2.60^{**}	0.01
$C(3, PB)$	0.92	0.12	0.29	0.38	0.88	2.52^{**}	0.01
PC	0.87	0.19	0.58	0.28	0.81	3.17^{**}	0.00
D. $h = 12$							
BA model	1.27						
Mean	0.99	0.79	1.22	0.11	0.21	0.81	0.21
Median	0.98	0.79	1.25	0.11	0.21	0.82	0.21
Trimmed mean	1.01	0.75	1.25	0.11	0.25	1.11	0.14
DMSE, $\psi = 1.00$	0.99	0.80	1.22	0.11	0.20	0.77	0.22
DMSE, $\psi = 0.90$	0.98	0.81	1.24	0.11	0.19	0.72	0.24
$C(2, PB)$	0.99	0.81	1.21	0.11	0.19	0.74	0.23
$C(3, PB)$	0.99	0.82	1.17	0.12	0.18	0.66	0.26
PC	0.89	1.23	2.19^{*}	0.02	-0.23	-0.57	0.72

Notes to Tables 2–5: The MSFE ratio reports the ratio of the MSFE for the BA model or combination forecasts indicated on the left to the MSFE for the AR benchmark model forecasts. H_0: BA encompasses CB (H_0: CB encompasses BA) corresponds a test of the null hypothesis that the BA model (combination) forecasts encompass the combination (BA model) forecasts against the one-sided (upper-tail) alternative hypothesis that the BA model (combination) forecasts do not encompass the combination (BA model) forecasts; $\hat{\lambda}_{CB}$ ($\hat{\lambda}_{BA}$) is the OLS estimate of weight on the combination (BA model) forecast in the optimal convex combination forecast given by (7); $MHLN_h$ is the test statistic corresponding to the null hypothesis; 0.00 indicates less than 0.005; †, *, ** indicate significance at the 10%, 5%, and 1% levels, respectively.

TABLE 3 Forecasting and encompassing test results for the bagging model and combining methods, 1985:04–2005:03 out-of-sample period

		H_0: BA encompasses CB			H_0: CB encompasses BA		
Combining method	MSFE ratio	$\hat{\lambda}_{CB}$	$MHLN_h$	p-value	$\hat{\lambda}_{BA}$	$MHLN_h$	p-value
A. $h = 1$							
BA model	0.94						
Mean	0.96	0.47	2.55^{**}	0.01	0.53	4.86^{**}	0.00
Median	0.98	0.44	2.47^{**}	0.01	0.56	5.29^{**}	0.00
Trimmed mean	0.97	0.46	2.53^{**}	0.01	0.54	5.02^{**}	0.00
DMSE, $\psi = 1.00$	0.96	0.48	2.55^{**}	0.01	0.52	4.84^{**}	0.00
DMSE, $\psi = 0.90$	0.96	0.47	2.54^{**}	0.01	0.53	4.86^{**}	0.00
$C(2, PB)$	0.96	0.47	2.47^{**}	0.01	0.53	4.91^{**}	0.00
$C(3, PB)$	0.95	0.50	2.47^{**}	0.01	0.50	4.65^{**}	0.00
PC	0.88	0.66	3.02^{**}	0.00	0.34	2.81^{**}	0.00
B. $h = 3$							
BA model	0.85						
Mean	0.92	0.42	2.44^{**}	0.01	0.58	4.08^{**}	0.00
Median	0.95	0.39	2.39^{**}	0.01	0.61	4.30^{**}	0.00
Trimmed mean	0.93	0.42	2.49^{**}	0.01	0.58	4.19^{**}	0.00
DMSE, $\psi = 1.00$	0.92	0.43	2.45^{**}	0.01	0.57	4.04^{**}	0.00
DMSE, $\psi = 0.90$	0.91	0.43	2.44^{**}	0.01	0.57	4.04^{**}	0.00
$C(2, PB)$	0.90	0.44	2.36^{**}	0.01	0.56	3.72^{**}	0.00
$C(3, PB)$	0.88	0.46	2.32^{*}	0.01	0.54	3.64^{**}	0.00
PC	0.74	0.65	2.94^{**}	0.00	0.35	3.00^{**}	0.00
C. $h = 6$							
BA model	0.77						
Mean	0.93	0.35	$1.52^{\dagger}$	0.06	0.651	3.93^{**}	0.00
Median	0.95	0.33	$1.51^{\dagger}$	0.07	0.67	3.99^{**}	0.00
Trimmed mean	0.94	0.34	$1.54^{\dagger}$	0.06	0.66	4.01^{**}	0.00
DMSE, $\psi = 1.00$	0.92	0.35	$1.53^{\dagger}$	0.06	0.65	3.91^{**}	0.00
DMSE, $\psi = 0.90$	0.92	0.36	$1.51^{\dagger}$	0.07	0.64	3.92^{**}	0.00
$C(2, PB)$	0.90	0.36	$1.47^{\dagger}$	0.07	0.64	3.70^{**}	0.00
$C(3, PB)$	0.89	0.35	$1.34^{\dagger}$	0.09	0.65	3.37^{**}	0.00
PC	0.76	0.51	2.29^{*}	0.01	0.49	4.36^{**}	0.00
D. $h = 12$							
BA model	1.13						
Mean	0.96	0.62	1.87^{*}	0.03	0.38	2.00^{*}	0.02
Median	0.97	0.61	1.86^{*}	0.03	0.39	2.04^{*}	0.02
Trimmed mean	0.97	0.61	1.88^{*}	0.03	0.39	2.15^{*}	0.02
DMSE, $\psi = 1.00$	0.96	0.63	1.87^{*}	0.03	0.37	1.96^{*}	0.03
DMSE, $\psi = 0.90$	0.97	0.62	1.83^{*}	0.03	0.38	1.89^{*}	0.03
$C(2, PB)$	0.96	0.63	1.79^{*}	0.04	0.37	1.71^{*}	0.04
$C(3, PB)$	0.97	0.64	1.72^{*}	0.04	0.36	$1.56^{\dagger}$	0.06
PC	0.84	0.88	2.82^{**}	0.00	0.12	0.59	0.28

TABLE 4 Forecasting and encompassing test results for the bagging model and combining methods, 1999:05–2004:05 out-of-sample period

		H_0: BA encompasses CB			H_0: CB encompasses BA		
Combining method	MSFE ratio	$\hat{\lambda}_{CB}$	$MHLN_h$	p-value	$\hat{\lambda}_{BA}$	$MHLN_h$	p-value
A. $h=1$							
BA model	0.90						
Mean	0.97	0.36	2.00*	0.03	0.64	2.83**	0.00
Median	0.99	0.33	1.99*	0.03	0.67	3.03**	0.00
Trimmed mean	0.98	0.35	1.98*	0.03	0.65	2.94**	0.00
DMSE, $\psi=1.00$	0.97	0.36	1.99*	0.03	0.64	2.84**	0.00
DMSE, $\psi=0.90$	0.97	0.36	2.00*	0.03	0.64	2.82**	0.00
$C(2, PB)$	0.97	0.35	1.92*	0.03	0.65	2.87**	0.00
$C(3, PB)$	0.96	0.36	1.87*	0.03	0.64	2.75**	0.00
PC	0.89	0.57	1.94*	0.03	0.43	1.53†	0.07
B. $h=3$							
BA model	0.65						
Mean	0.92	-0.03	-0.15	0.56	1.03	3.05**	0.00
Median	0.96	-0.05	-0.27	0.61	1.05	3.15**	0.00
Trimmed mean	0.92	0.00	0.03	0.49	1.00	3.12**	0.00
DMSE, $\psi=1.00$	0.92	-0.03	-0.16	0.56	1.03	3.04**	0.00
DMSE, $\psi=0.90$	0.91	-0.03	-0.17	0.57	1.03	3.04**	0.00
$C(2, PB)$	0.91	-0.06	-0.26	0.60	1.06	2.91**	0.00
$C(3, PB)$	0.90	-0.05	-0.21	0.58	1.05	2.78**	0.00
PC	0.76	-0.06	-0.18	0.57	1.06	2.38*	0.01
C. $h=6$							
BA model	0.57						
Mean	0.93	-0.45	-1.42	0.92	1.45	3.55**	0.00
Median	0.96	-0.41	-1.36	0.91	1.41	3.50**	0.00
Trimmed mean	0.93	-0.34	-1.11	0.87	1.34	3.88**	0.00
DMSE, $\psi=1.00$	0.93	-0.46	-1.41	0.92	1.46	3.46**	0.00
DMSE, $\psi=0.90$	0.92	-0.46	-1.42	0.92	1.46	3.49**	0.00
$C(2, PB)$	0.92	-0.50	-1.34	0.91	1.50	3.12**	0.00
$C(3, PB)$	0.90	-0.58	-1.51	0.93	1.58	3.29**	0.00
PC	0.76	-0.67	-1.27	0.89	1.67	2.95**	0.00
D. $h=12$							
BA model	0.88						
Mean	0.97	0.05	0.05	0.48	0.95	0.84	0.20
Median	0.98	0.09	0.08	0.47	0.91	0.87	0.20
Trimmed mean	0.97	0.12	0.12	0.45	0.88	0.94	0.18
DMSE, $\psi=1.00$	0.97	0.07	0.06	0.48	0.93	0.81	0.21
DMSE, $\psi=0.90$	0.96	0.09	0.07	0.47	0.91	0.79	0.22
$C(2, PB)$	0.96	0.00	0.00	0.50	1.00	0.82	0.21
$C(3, PB)$	0.97	-0.09	-0.07	0.53	1.09	0.84	0.20
PC	0.81	0.94	0.89	0.19	0.06	0.07	0.47

TABLE 5 Forecasting and encompassing test results for the bagging model and combining methods, 1988:09–1993:09 out-of-sample period

		H_0: BA encompasses CB			H_0: CB encompasses BA		
Combining method	MSFE ratio	$\hat{\lambda}_{CB}$	$MHLN_h$	p-value	$\hat{\lambda}_{BA}$	$MHLN_h$	p-value
A. $h = 1$							
BA model	0.85						
Mean	0.96	0.34	$1.57^{\dagger}$	0.06	0.66	2.79^{**}	0.00
Median	0.99	0.30	$1.47^{\dagger}$	0.07	0.70	2.97^{**}	0.00
Trimmed mean	0.99	0.29	$1.34^{\dagger}$	0.09	0.71	2.99^{**}	0.00
DMSE, $\psi = 1.00$	0.95	0.34	$1.57^{\dagger}$	0.06	0.66	2.78^{**}	0.00
DMSE, $\psi = 0.90$	0.95	0.33	$1.54^{\dagger}$	0.06	0.67	2.78^{**}	0.00
$C(2, PB)$	0.95	0.34	$1.52^{\dagger}$	0.07	0.66	2.76^{**}	0.00
$C(3, PB)$	0.92	0.37	$1.63^{\dagger}$	0.05	0.63	2.60^{**}	0.01
PC	0.85	0.49	2.02^{*}	0.02	0.51	1.96^{*}	0.03
B. $h = 3$							
BA model	0.65						
Mean	0.95	0.22	0.86	0.20	0.78	2.80^{**}	0.00
Median	0.97	0.21	0.83	0.20	0.79	2.79^{**}	0.00
Trimmed mean	0.98	0.19	0.74	0.23	0.81	2.95^{**}	0.00
DMSE, $\psi = 1.00$	0.93	0.23	0.87	0.20	0.77	2.79^{**}	0.00
DMSE, $\psi = 0.90$	0.91	0.23	0.85	0.20	0.77	2.76^{**}	0.00
$C(2, PB)$	0.92	0.22	0.77	0.22	0.78	2.66^{**}	0.01
$C(3, PB)$	0.88	0.23	0.75	0.23	0.77	2.67^{**}	0.00
PC	0.63	0.52	$1.32^{\dagger}$	0.10	0.48	2.05^{*}	0.02
C. $h = 6$							
BA model	0.49						
Mean	0.94	-0.03	-0.08	0.53	1.03	2.55^{**}	0.01
Median	0.96	-0.04	-0.11	0.54	1.04	2.54^{**}	0.01
Trimmed mean	1.00	-0.06	-0.18	0.57	1.06	2.86^{**}	0.00
DMSE, $\psi = 1.00$	0.92	-0.02	-0.07	0.53	1.02	2.56^{**}	0.01
DMSE, $\psi = 0.90$	0.92	-0.05	-0.12	0.55	1.05	2.48^{**}	0.01
$C(2, PB)$	0.89	-0.06	-0.14	0.56	1.06	2.40^{**}	0.01
$C(3, PB)$	0.90	-0.09	-0.21	0.58	1.09	2.22^{*}	0.02
PC	0.63	0.20	0.54	0.30	0.80	3.38^{**}	0.00
D. $h = 12$							
BA model	0.49						
Mean	0.98	-0.60	-0.88	0.81	1.60	1.94^{*}	0.03
Median	1.00	-0.59	-0.86	0.80	1.59	1.84^{*}	0.04
Trimmed mean	1.01	-0.59	-0.93	0.82	1.59	2.09^{*}	0.02
DMSE, $\psi = 1.00$	0.96	-0.61	-0.88	0.81	1.61	1.94^{*}	0.03
DMSE, $\psi = 0.90$	1.00	-0.62	-0.88	0.81	1.62	1.83^{*}	0.04
$C(2, PB)$	0.99	-0.70	-0.90	0.81	1.70	$1.61^{\dagger}$	0.06
$C(3, PB)$	0.99	-0.75	-0.92	0.82	1.75	$1.55^{\dagger}$	0.06
PC	0.71	0.08	0.38	0.35	0.92	$1.37^{\dagger}$	0.09

TABLE 6 MSFE ratios for forecasts formed as the average of the BA model and PC combination forecasts

	$h=1$	$h=3$	$h=6$	$h=12$	$h=1$	$h=3$	$h=6$	$h=12$
	A. 1995:04–2005:03 out-of-sample period				B. 1985:04–2005:03 out-of-sample period			
BA model	0.90	0.71	0.73	1.27	0.94	0.85	0.77	1.13
PC combination	0.87	0.85	0.87	0.89	0.88	0.74	0.76	0.84
BA encompasses PC?	No	Yes	Yes	No	No	No	No	No
PC encompasses BA?	No	No	No	Yes	No	No	No	Yes
Average	0.85	0.73	0.74	1.02	0.86	0.71	0.68	0.89
	C. 1999:05–2004:05 out-of-sample period				D. 1988:09–1993:09 out-of-sample period			
BA model	0.90	0.65	0.57	0.88	0.85	0.65	0.49	0.49
PC combination	0.89	0.76	0.76	0.81	0.85	0.63	0.63	0.71
BA encompasses PC?	No	Yes	Yes	Yes	No	No	Yes	Yes
PC encompasses BA?	No	No	No	Yes	No	No	No	No
Average	0.87	0.68	0.65	0.82	0.80	0.56	0.50	0.53

Note: The entries in the BA model and PC combination rows report the ratio of the MSFE for these forecasts to the MSFE for the AR benchmark model forecasts. The entries in the BA encompasses PC? (PC encompasses BA?) rows indicate whether the BA model (PC combination) forecasts encompass the PC combination (BA model) forecasts according to the results in Tables 2–5 using a 10% significance level. The entries for the Average rows report the ratio of the MSFE for a forecast formed as a simple average of the BA model and PC combination forecasts to the MSFE for the AR benchmark model forecasts.