WP 49_07

COSTAS MILAS
Keele University, UK
and
The Rimini Center for Economic Analysis, Italy

PHILIP ROTHMAN
East Carolina University, USA

"Out-of-Sample Forecasting of Unemployment Rates with Pooled STVECM Forecasts"

The Rimini Centre for Economic Analysis
Legal address: Via Angherà, 22 – Head office: Via Patara, 3 - 47900 Rimini (RN) – Italy
www.rcfea.org - secretary@rcfea.org

Out-of-Sample Forecasting of Unemployment Rates with Pooled STVECM Forecasts *

Costas Milas†
Department of Economics, Keele University
Rimini Centre for Economic Analysis

Philip Rothman‡
Department of Economics, East Carolina University

October 31, 2007

Abstract

In this paper we use smooth transition vector error-correction models (STVECMs) in a simulated out-of-sample forecasting experiment for the unemployment rates of the four non-Euro G-7 countries, the U.S., U.K., Canada, and Japan. For the U.S., pooled forecasts constructed by taking the median value across the point forecasts generated by the linear and STVECM forecasts appear to perform better than the linear AR(p) benchmark more so during business cycle expansions. Such pooling also tends to lead to statistically significant forecast improvement for the U.K. "Reality checks" of these results suggest that they do not stem from data snooping.

Keywords

nonlinear, asymmetric, STVECM, pooled forecasts, Diebold-Mariano

*We thank two anonymous referees and the editor, Mike Clements, for very helpful comments and suggestions. We also thank Dick van Dijk for kindly providing us with his GAUSS code, Greg Tkacz for supplying some data, and Norm Swanson for help in implementing the "reality check" test.

†Department of Economics, Keele University, Keele, ST5 5BG, UK, `email: c.milas@keele.ac.uk`

‡Department of Economics, Brewster Building, East Carolina University, Greenville, NC 27858 USA, `email: rothmanp@mail.ecu.edu`. Support from an East Carolina University Faculty Senate research grant is gratefully acknowledged.

1 Introduction

Applying the statistical theory of finite-state Markov chains, Neftci (1984) reported evidence showing that the U.S. quarterly unemployment rate is asymmetric in the sense that the probability of a decrease in the series, conditional on two preceding decreases, is greater than the corresponding probability of an increase conditional on two previous increases. One of the primary time series implications of such behavior is that it is inconsistent with a linear data generating process with symmetrically distributed innovations.

Neftci's study inaugurated a near two-decade long research program in which the extent to which key business cycle indicators display varying forms of asymmetric dynamics has been explored; see Clements and Krolzig (2003) for a useful survey of important developments in the business cycle asymmetry literature. Many of these papers have focused specifically on unemployment rates and have frequently documented strong evidence in favor of dynamic asymmetries, often in the form of parametric nonlinear models, for these series. Recent work includes: Altissimo and Violante (2001), who, by way of a threshold vector autoregressive (VAR) model of U.S. output and unemployment with feedback from a Beaudry and Koop (1993)-like "depth of recession" measure, identified nonlinearities in the propagation and persistence of shocks as well as a beneficial long-run effect of recessions on growth; Caner and Hansen (2001), who uncovered threshold autoregressive (TAR) effects in the U.S. unemployment rate via their TAR-based unit root test; and Skalin and Teräsvirta (2002), whose results suggest that smooth transition autoregressive (STAR) models can capture well the asymmetry displayed in many OECD unemployment rate series.

While the vast majority of these papers have been concerned with in-sample fits of linear and nonlinear models to unemployment rate data, researchers have increasingly investigated one of the main practical problems which has motivated the literature, that is, whether out-of-sample unemployment rate forecasts generated by nonlinear time series models can dominate those produced with standard linear models. Rothman (1998) was one of the first to consider this question for the U.S. quarterly unemployment rate. He analyzed the forecasting performance of six nonlinear time series models against linear forecasts, and in many cases the mean squared prediction error (MSPE) associated with the nonlinear forecasts was less than those for the linear forecasts. Montgomery, Zarnowitz, Tsay, and Tiao (1998) found that, at multistep-ahead forecast horizons during business cycle contractions, TAR and Markov-switching autoregressive models outperformed in the MSPE-sense the benchmark linear model in out-of-sample forecasting of the U.S. quarterly unemployment rate. Using artificial neural network (ANN) and logistic STAR (LSTAR) models for a very large data set of U.S. macroeconomic time series, including the monthly unemployment rate, Stock and Watson (1999) showed that linear forecasts generally dominated the nonlinear forecasts. However, following a similar approach with an analogous data set for the Euro area, Marcellino (2006) reported much more favorable results for ANN and LSTAR forecasts; for the Euro area unemployment rates in particular, the ANN and LSTAR forecasts had lower MSPEs two and half times more often than did the linear forecasts.[1]

A common feature of the nonlinear forecasts evaluated in these four papers is that they were all univariate.[2] This marks a significant point of departure for our paper: while we also examine nonlinear forecasts of unemployment rates, the models we use are multivariate. The macroeconomic theoretical motivation behind a multivariate approach is straightforward; through standard arguments it is reasonable to assume that the unemployment rate is interrelated with other important variables. The degree to which a particular nonlinear parameterization of these relationships can be exploited to yield improved forecast improvement is the empirical issue addressed in this paper.

To investigate this question for unemployment rates, we employ multivariate STAR models in which we impose cointegrating restrictions. In doing so, we build upon Skalin and Teräsvirta (2002), who noted that their univariate in-sample analysis can be interpreted as a first step in the specification of a multivariate STAR model of unemployment rates.

We also follow Rothman, van Dijk, and Franses (2001), who used a similar approach to study the Granger-causal relationship between money and output. These authors found strong evidence in favor of STAR-type nonlinearity in a system of output, prices, interest rates, and money. By Okun's Law, comparable results arguably are expected to hold for an analogous model in which output is replaced by the unemployment rate. In addition to our primary focus on unemployment rates, there are several differences between our paper and Rothman *et al.* (2001).

First, our chief concern is evaluation of the out-of-sample forecasting performance of the models, while Rothman *et al.* (2001) concentrated on both in-sample and out-of-sample results to analyze the money-output relationship. Our main in-sample interest is identification of the *transition variables* which govern parameter variation in STAR models. Second, ours is a closer approximation to real-time implementation of these forecasting models. In Rothman *et al.* (2001) specification of the STAR models was done using practically the full sample, such that common specifications were imposed in all rolling windows of data. While this aided interpretation of the results with respect to the Granger causality question under consideration, it effectively allowed the use of post-sample information in generating the forecasts. In contrast, we specify the models for each data window only using data available through the date of each forecast, and thus allow the model specifications to vary across data windows. Though this substantially increases the computational burden, we feel our experimental design offers a better simulation of real-time forecasting practice.[3] Third, in this paper forecasts are computed using two approaches: following the standard route by iterating forward estimated one-step-ahead models; and also, following Stock and Watson (1999) and Marcellino (2006), by estimating directly h-step-ahead models and projecting them forward. This allows a useful comparison of these strategies for forecasting unemployment rates. Rothman *et al.* (2001) did not employ "h-step-ahead projections" for multistep-ahead forecasting. Fourth, we consider some easily-constructed pooled forecasts, whereas Rothman *et al.* (2001) did not use any forecast pooling procedures. Finally, while Rothman *et al.* (2001) only worked with U.S. data, we examine multivariate STAR models with data for the U.S., U.K., Canada, and Japan.

Another paper quite close to ours is Krolzig, Marcellino, and Mizon (2002), who analyzed a Markov-switching vector error correction model (MS-VECM) of the U.K. labor market with quarterly data. Besides our use of STAR as opposed to MS models, there are several differences between our paper and Krolzig *et al.* (2002). First, their unemployment measure was the volume of unemployment, whereas we use unemployment rates. Second, selection of the variables to be included in their system came out of specific focus on the labor market; their four-variable system comprised unemployment, employment, real output, and real wages. In contrast, our choice of variables follows a standard practice in the empirical monetary policy literature; our four-variable system comprises the unemployment rate, the aggregate price level, a monetary aggregate, and a short-term interest rate. Third, in their out-of-sample forecasting exercise, Krolzig *et al.* (2002) only computed one-step-ahead forecasts, and do so only for two estimated versions of their model.[4] In our approach the models used are reestimated for each fixed-length rolling window of data, and we compute one-quarter-ahead through eight-quarters-ahead forecasts.

The paper proceeds as follows. In Section 2 we discuss the multivariate STAR model and outline a specification procedure for such models. The results of linearity testing against STAR alternatives within a multivariate context are also presented in this section. Our out-of-sample forecasting results are examined in Section 3 and Section 4 concludes the paper.

2 Multivariate STAR Models and Linearity Testing

Let $\mathbf{x}_t = (x_{1t}, \ldots, x_{kt})'$ be a $(k \times 1)$ vector time series. In our case we have $\mathbf{x}_t = (u_t, m_t, p_t, i_t)'$, with u_t the log of the unemployment rate, m_t the log of a money supply measure, p_t the log of the producer price index, and i_t a short-term interest rate.[5] We analyze quarterly vector time series for four different countries, the U.S., U.K., Canada, and Japan for the 1959:1-2005:4, 1965:1-2005:4, 1968:1-2005:4, and 1966:4-2005:4 sample periods, respectively.[6] The unemployment rate, money supply, and producer price index series were seasonally adjusted, while the interest rate series were not. The data were obtained from the following sources: the Federal Reserve Bank of St. Louis, the U.S. Bureau of Labor Statistics, the U.K. Office for National Statistics, the Bank of Canada, and the OECD *Main Economic Indicators* and the IMF *International Financial Statistics* databases. Figures 1-4 present time series plots of the data used.

A k-dimensional smooth transition vector error-correction model [STVECM] can be specified as

$$\Delta \mathbf{x}_t = \left(\boldsymbol{\mu}_1 + \boldsymbol{\alpha}_1 \mathbf{z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_{1,j} \Delta \mathbf{x}_{t-j} \right) (1 - G(s_t; \gamma, c)) \\ + \left(\boldsymbol{\mu}_2 + \boldsymbol{\alpha}_2 \mathbf{z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_{2,j} \Delta \mathbf{x}_{t-j} \right) G(s_t; \gamma, c) + \boldsymbol{\varepsilon}_t, \quad (1)$$

where Δ_j denotes the j-th difference operator, defined as $\Delta_j x_t = x_t - x_{t-j}$ for integers $j \neq 0$ and $\Delta_1 \equiv \Delta$, $\boldsymbol{\mu}_i$, $i = 1, 2$, are $(k \times 1)$ vectors, $\boldsymbol{\alpha}_i$, $i = 1, 2$, are $(k \times r)$ matrices, $\mathbf{z}_t = \boldsymbol{\beta}' \mathbf{x}_t$ for some $(k \times r)$ matrix $\boldsymbol{\beta}$ denoting the error-correction terms, $\boldsymbol{\Phi}_{i,j}$, $i = 1, 2$, $j = 1, \ldots, p-1$, are $(k \times k)$ matrices, and $\boldsymbol{\varepsilon}_t = (\varepsilon_{1t}, \ldots, \varepsilon_{kt})$ is a k-dimensional vector white noise process with mean zero and $(k \times k)$ covariance matrix $\boldsymbol{\Sigma}$. The transition function $G(s_t; \gamma, c)$ is assumed to be a continuous function bounded between zero and one. In this paper we allow the transition variable s_t to be either a function of a lagged component of $\mathbf{x}_t$ or a lagged exogenous variable.

The STVECM can be thought of as a regime-switching model that allows for two regimes associated with the extreme values of the transition function, $G(s_t; \gamma, c) = 0$ and $G(s_t; \gamma, c) = 1$, where the transition from one regime to the other is smooth. In this paper we restrict attention to the logistic transition function

$$G(s_t; \gamma, c) = \frac{1}{1 + \exp\{-\gamma(s_t - c)/\hat{\sigma}_s\}}, \qquad \gamma > 0, \quad (2)$$

where $\hat{\sigma}_s$ is the sample standard deviation of s_t. The parameter c in (2) can be interpreted as the threshold or border between the two regimes, in the sense that the logistic function changes monotonically from 0 to 1 as s_t increases, and $G(c; \gamma, c) = 0.5$. The parameter γ determines the smoothness of the change in the value of the logistic function and, thus, the smoothness of the transition from one regime to the other. As γ becomes very large, the logistic function approaches the indicator function $I[s_t > c]$. Hence, the STVECM (1) with (2) nests a two-regime threshold vector error-correction model as a special case; see Balke and Fomby (1997) and Tsay (1998) for discussion. Finally, note that when $\gamma = 0$ the logistic function equals 0.5 for all s_t, such that the STVECM model reduces to a linear

VECM.

The procedure we follow for specifying STVECMs is a straightforward modification of the specification procedure for univariate STAR models put forward by Teräsvirta (1994). We start by specifying a linear VECM for $\mathbf{x}_t$, that is,

$$\Delta \mathbf{x}_t = \boldsymbol{\mu} + \boldsymbol{\alpha} \mathbf{z}_{t-1} + \sum_{j=1}^{p-1} \boldsymbol{\Phi}_j \Delta \mathbf{x}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{3}$$

where the lag order p should be such that the residuals $\hat{\boldsymbol{\varepsilon}}_t$ are approximately white noise and have zero autocorrelations at all lags. To reduce the number of parameters $(4 + (4 \times r) + (4 \times 4 \times (p-1)))$, we decided to use a subset VECM by imposing zero restrictions on coefficients in the $\boldsymbol{\Phi}_j$, $j = 1, \ldots, p-1$, matrices in (3). Use of such subset models simplifies computation of the test statistics required for the linearity tests described below and significantly eases estimation of the STVECMs used.

The subset VECM is specified by following the strategy recommended by Brüggemann and Lütkepohl (2001), which treats the individual equations in the VECM separately. We estimate the parameters in the i-th equation of (3) by ordinary least squares [OLS] and sequentially delete the regressor with the smallest absolute value of the corresponding t-ratios, until all t-ratios of the remaining coefficients are greater than some threshold value τ in absolute value; in each iteration only a single regressor is eliminated. Then the reduced model equation is re-estimated and new t-ratios are computed. We choose the threshold τ as a function of the iteration l as

$$\tau = \tau_l = \sqrt{(\exp(\lambda_T/T) - 1)(T - L + l - 1)}, \tag{4}$$

where T denotes the effective sample size, $L = 1+r+4\times(p-1)$ is the number of parameters in the unrestricted equation and λ_T is a sequence indexed by the sample size. As shown by Brüggemann and Lütkepohl (2001), by setting λ_T equal to the penalty term involved in an information criterion of choice, this procedure leads to the same final model as sequentially removing those regressors whose elimination yields the largest improvement in the value of this particular information criterion. We use the Akaike Information Criterion (AIC), which requires setting $\lambda_T = 2$.[7] It should be noted that we only eliminate lagged first differences from the VECM, and always retain the intercept and error-correction terms.

We set the cointegrating rank $r = 2$ and pre-specify the two cointegrating vectors as $(1, 0, 0, 0)'$ and $(0, 0, 0, 1)'$, that is, the first row of $\mathbf{z}_t$ is the log-unemployment rate and the second row of $\mathbf{z}_t$ is the short-term interest rate. Such pre-specification as opposed to estimation of the cointegrating vectors serves as a simplifying pair of assumptions which allows us to focus on the value-added of allowing STAR-type effects in a multivariate forecasting model of the unemployment rate. In using u_t as an error-correction term, we follow Skalin and Teräsvirta (2002), who assumed that the unemployment rate, a bounded variable, is a globally stationary process. By way of an LSTAR specification, however, asymmetry and local nonstationarity are possible. Our decision to include i_t as an error-correction term follows Rothman *et al.* (2001), who did so in the 'Hendry-style' in that it was based on economic theory; see, for example, Hendry and Mizon (1993) and Söderlind and Vredin (1996). The latter authors showed that the Cooley and Hansen (1995) monetary equilibrium business cycle model implies that the nominal interest rate is stationary.[8]

The next step in the specification procedure is to select a transition variable s_t, which is done via linearity testing of the subset VECM against the alternative of a STVECM. To carry out our forecasting exercise we require a sequence of transition variables for "rolling" fixed-length windows of data, where the first data window runs from the first observation of the data set for each country out to 1991:4, and each successive data window is constructed by shifting the preceding window ahead by one observation. This setup allows us to generate 49 out-of-sample forecasts at forecast horizons $h = 1, \ldots, 8$.

Testing linearity in this context is complicated by the fact that the STVECM contains nuisance parameters which are not identified under the null hypothesis; see, for example, Davies (1987). To circumvent this identification problem, we follow the approach of Luukkonen, Saikkonen, Teräsvirta (1988) and replace the transition function $G(s_t; \gamma, c)$ with a suitable Taylor approximation. The S_1 test is a standard variable addition test based on an auxiliary regression of the residuals from the linear VECM on a set of variables given by a first-order Taylor expansion of $G(s_t; \gamma, c)$. The S_2 test is based on a third-order Taylor approximation of the logistic transition function, and the S_3 test is a parsimonious version of the S_2 test. For the sample sizes we have, it turns out that we lack sufficient degrees of freedom to compute the system-wide version of the S_2 test, such that we only use its parsimonious version.[9]

Given that our VECM residuals tend to be highly heteroskedastic, it makes sense to employ heteroskedasticity-robust versions of the linearity tests; simulations discussed by Rothman *et al.* (2001) showed that the estimated sizes of the non-robust linearity tests in the presence of heteroskedasticity tend to be severely distorted upwards. To this effect, the specification tests developed by Wooldridge (1991) are very helpful, since they can be used in the presence of heteroskedasticity without the need to specify the often unknown form of heteroskedasticity explicitly. The robust versions of the linearity tests we use were obtained by applying 'Procedure 3.1' of Wooldridge (1991).

Simulations discussed by Lundbergh and Teräsvirta (1998) and Rothman *et al.* (2001) suggest, however, that these robust tests are conservative, with estimated sizes less than nominal significance levels and low estimated power. Nonetheless, we follow Rothman *et al.* (2001) and apply these heteroskedasticity-robust versions of the linearity tests since we feel that the ranking across a set of prospective transition variables is valuable information for the STVECM specification process. It is unlikely that such a ranking will be affected by presence of heteroskedasticity in the VECM residuals.

To identify an appropriate transition variable s_t with a linearity test for each data window, we run the test for several candidates, $s_{1t}, \ldots, s_{mt}$, and select the one for which the p-value of the associated test statistic is smallest. Here we consider the following different candidate transition variables for all countries: lagged yearly changes in the log unemployment rate ($\Delta_4 u_{t-d}$), lagged yearly growth rates in the money supply ($\Delta_4 m_{t-d}$), lagged annual inflation rates ($\Delta_4 p_{t-d}$), lagged yearly changes in the short-term interest rate ($\Delta_4 i_{t-d}$), lagged yearly changes in the annual money supply growth rates ($\Delta_4 \Delta_4 m_{t-d}$), lagged yearly changes in the annual inflation rate ($\Delta_4 \Delta_4 p_{t-d}$), and lagged yearly changes in the relative price of oil ($\Delta_4 o_{t-d}$, with $o_t = p_t^{\text{OIL}}/p_t$ and p_t^{OIL} the crude petroleum producer price index). In addition, for the U.S. we also used lagged annual changes in the federal funds rate ($\Delta_4 ff_{t-d}$).[10]

The reason why we use 4-quarter differences as transition variables is that we expect the regimes in unemployment rate dynamics to be more so persistent, because, for example, they might be related to the business cycle or to monetary policy. Using 4-quarter differences effectively eliminates short-run fluctuations which do not necessarily represent changes in regimes. We test linearity with the above-mentioned variables for delays $d = 1, \ldots, d_{\max}$, where we set the maximum value of the delay parameter $d_{\max}$ equal to 4.

The empirical and theoretical literature upon which we base our focus on these particular candidate transition variables is large. Of particular relevance in our STAR context, we note that a good deal of research has been done which suggests that these variables are reasonable measures of either the 'state of the economy' and/or the 'state of policy.' As such, our use of these variables is strongly motivated by much of the macroeconomic research on 'state-dependent' dynamics; see, for example, Caplin and Leahy (1991) and Caballero and Hammour (1994). [11]

3 Out-of-Sample Forecasting

3.1 Forecasting Methods

Our forecasts are produced by 12 forecasting methods for our 49 simulated out-of-sample periods, where we use the term "method" in the sense of Stock and Watson (1999). That is, the sequence of forecasts generated by each method is based on an underlying "primitive model," and we let the specification of each primitive model vary across the 49 simulated in-sample periods. Forecasting Methods 1 and 2 are based on, respectively, identifying a univariate linear autoregressive (AR) model and a linear VECM for each rolling in-sample window, using the AIC and a diagnostic check for residual serial correlation. Given the evidence reported in the literature discussed above in Section 1 on the difficulty of beating linear AR models with nonlinear models in out-of-sample forecasting, below we generally treat Method 1 as the benchmark.

For each sample window the maximum lag length allowed is 4. The AR(p) model is estimated by least squares and the parameters of the VECM are obtained by seemingly unrelated regressions estimation. Mutistep-ahead forecasts are computed by iterating forward the estimated one-step-ahead models; we end up with forecasts for steps $h = 1, \ldots, 8$.

Forecasting Methods 3 through 7 use as the primitive model a STVECM, and generate multistep-ahead forecasts by, as we do with forecasting Methods 1 and 2, iterating forward the estimated one-step model. But since the models are nonlinear, we use bootstrap simulations to help compute the multistep-ahead forecasts. These STVECM forecasting methods differ as to how the transition variable is selected. Methods 3 through 5 use the top-ranked candidate variable as determined by the single-equation S_1, S_2, and S_3 tests, respectively, run on the first-differenced log-unemployment rate equation of a subset linear VECM obtained through the Brüggemann and Lütkepohl (2001) procedure. The Brüggemann and Lütkepohl (2001) algorithm is further applied to the STVECM to facilitate estimation of the model. Methods 6 and 7 do the same with the system-wide S_1 and S_3 tests, respectively.

Forecasting Methods 8 through 10 use STVECM "$h-$step-ahead projections" constructed as follows. First, for each in-sample window, we estimate directly the $h-$step-ahead model for the log-unemployment equation of the STVECM, such that with the dependent variable Δu_t, the first lag allowed amongst the regressors is from observation $t - h$ for forecast step h. This requires that we select a transition variable for each separate forecast step h for each of the 49 in-sample rolling windows of data, which leads to selection of 1,176 (8 forecast steps $\times$ 3 forecasting methods $\times$ 49 data windows) transition variables per country. Methods 8 through 10 base the selection of the transition variable on the single-equation S_1, S_2, and S_3 tests, respectively, run on the first-differenced log-unemployment rate equation of an unrestricted $h-$step-ahead linear VECM; after selection of the transition variable, the Brüggemann and Lütkepohl (2001) procedure is applied to the STVECM log-unemployment rate equation to help identify the appropriate regressors. Second, the forecast of Δu_{t+h} is computed by projecting the estimated equation ahead by h periods.

Stock and Watson (1999) and Marcellino (2006) emphasize that use of such $h-$step-ahead-projections simplifies significantly computation of the multistep-ahead nonlinear forecasts, since no simulations are required for forecast steps $h > 1$. On the other hand, this requires a very large increase in the number of linearity tests run to rank the candidate transition variables for all data windows. Further, these authors point out that $h-$step-ahead projections can reduce the effects of misspecification of the estimated one-step-ahead, since the effects of such misspecification do not propagate through to the multistep-ahead forecasts. Estimation of all STVECMs used in Methods 3 through 10 is done by nonlinear generalized least squares.

In addition, we employ two straightforward pooling procedures. First, Method 11 forecasts are constructed by taking the median forecast value from the nonlinear forecasts produced by Methods 3 through 10. Second, Method 12 uses the median forecast across Methods 1 through 10. Table 1 summarizes all of the forecasting methods used.

3.2 Out-of-Sample Forecasting Ranks

Table 2 presents the average out-of-sample forecasting rankings across the 49 forecasting windows and 8 forecast horizons of these methods for each of the four countries according to two evaluation criteria, the mean squared prediction error (MSPE) and median squared prediction error (MedSPE); note that the "better" or "higher ranked" forecasting methods have "lower" numerical ranks. In examining the average rank results in this table, it is useful to note that if the average rank of Method i is higher than the average rank of Method j according to either the MSPE or MedSPE, then Method i outperforms Method j via the particular criterion for more than 50% of the forecast horizons, that is, for at least 5 out of the 8 forecast horizons used.

The key result for the U.S. is that Method 11, the pooled median forecast across the STVECMs, is the top-ranked forecasting methods according to both the MSPE and MedSPE. So, in addition to dominating the linear AR(p) and VECM forecasts of Methods 1 and 2, median-pooling across the nonlinear forecast methods is superior to such pooling when the linear forecasts are also used. The result that median-pooling across all the nonlinear forecasting methods dominates each of the individual ones suggests that focus on single-primitive-model-based nonlinear forecasting methods may mask the potential gains obtainable by combining these individual nonlinear forecasts. Method 1, based on forecasts from the linear AR(p) model, is the sixth-ranked forecasting method according to the MSPE criterion, and its relative performance decreases substantially, down to twelfth out of the 12 methods, using the MedSPE, the more robust forecast comparison criterion. The linear VECM forecasts of Method 2 dominate the linear AR(p) forecasts according to both the MSPE and MedSPE.

We next discuss the forecasting ranks of Methods 3 through 10 for the U.S., since we are interested in determining whether any particular class of STVECM forecasts used tends to dominate another. As per the definitional scheme given in Table 1, we distinguish three such classes of STVECM forecasts within Methods 3 through 10: Methods 3 through 5; Methods 6 and 7; and Methods 8 through 10. Via the MSPE, Methods 6 and 7, which select transition variables through system-wide linearity tests and generate multistep-ahead forecasts by iterating the one-step-ahead model, outperform the other nonlinear forecasting methods. This result does not carry through, however, when ranking the methods with the MedSPE. Further, the h-step-projections of Methods 8 through 10 do not dominate the other STVECM forecasts using either the MSPE or the MedSPE.

For the U.K., Method 12 is the top-ranked forecasting method according to the MSPE and Method 2 is top-ranked using the MedSPE; the ranks of these two forecasting methods are reversed when the robust MedSPE evaluation criterion is used. In contrast to the U.S. case, Method 12 dominates Method 11 according to both the MSPE and MedSPE, i.e., median-pooling is more helpful when the linear AR(p) and VECM forecasts are included. Using both the MSPE and MedSPE, the VECM forecasts strongly dominate the AR(p) forecasts, which rank seventh and ninth, respectively, via these two criteria.

Further, according to the MSPE, Methods 4, 3, and 5 are top-ranked, respectively, among the nonlinear non-pooling Methods 3 through 10. So, as per the MSPE criterion, system-wide nonlinearity testing for model identification in the U.K. case leads to worse forecasting performance relative to nonlinearity testing restricted to the unemployment rate equation of the STVECM. Using both the MSPE and MedSPE, the $h-$step-ahead forecasts of Methods 8 through 10 perform worst.

For Canada, Method 12 is the top-ranked forecasting method according to both the MSPE and MedSPE, showing that the nonlinear forecasts provide useful information which is not incorporated in the linear forecasts. The second-ranked forecasting methods are Methods 11 and 4 using, respectively, the MSPE and MedSPE. As in the U.S. case, forecasting evaluation via the MedSPE leads to a decrease in the relative performance of the linear forecasting methods. As in the U.K. case, model identification based on system-wide nonlinearity testing leads to higher MSPEs in comparison to selecting the transition variable through such testing only on the STVECM unemployment rate equation. But in contrast to the U.S. and U.K. cases, the univariate AR(p) forecasts dominate those produced by the linear VECM. In addition, according to the MedSPE the forecasts generated by STVECMs of Methods 6 and 7 perform worst.

For Japan, the linear forecasts are top-ranked using both the MSPE and MedSPE. More specifically, the AR(p) forecasts of Method 1 are ranked first according to the MSPE and the VECM forecasts of Method 2 are ranked first according to the MedSPE. As per the MedSPE, Methods 2 and 1 are ranked, respectively, first and second. Median-pooling across all forecasting methods in Method 12 dominates each of the individual nonlinear forecasting methods via both both MSPE and MedSPE. According to these two criteria, Method 6 performs worst. Using the MedSPE, the $h-$step-ahead forecasts dominate the other nonlinear forecasts.

3.3 MSPE Ratios

Tables 3-6 present further details on the relative performance of the alternative forecasting methods on the basis of the MSPE. Using Method 1 as the benchmark, MSPE ratios are reported for each of the forecasting horizons $h = 1, \ldots, 8$; these were computed with Method 1's MSPE in the denominator. The linear AR(p) forecasts of Method 1 are used as the benchmark since much of the literature, e.g., Montgomery *et al.* (1998), has focused on the extent to which nonlinear models can dominate linear univariate models in out-of-sample forecasting. In the last column of each table the average MSPE ratio across the eight forecast steps is reported. This measure helps quantify the extent of any gains to use of the alternatives considered to Method 1.

To examine whether any of the MSPE reductions are statistically significant, we apply the Harvey, Leybourne, and Newbold (1997) modification of the Diebold and Mariano (1995) statistic (DM). The DM test statistic is computed by weighting the forecast loss differentials between the two competing methods equally, where the loss differential for observation t is defined by $d_t \equiv g(e_{i,t|t-h}) - g(e_{j,t|t-h})$, with $g(\cdot)$ some arbitrary loss function, and $e_{i,t|t-h}$ and $e_{j,t|t-h}$ the $h-$step-ahead forecast errors for Methods i and j. That is, the DM test examines whether the following equally-weighted sample mean loss differential, when standardized, is different from zero at some given significance level

$$\bar{d} = \frac{1}{P} \sum_{t=R+h}^{R+P+h-1} d_t, \tag{5}$$

where forecasts have been produced for observations $t = R+h, \ldots, R+P+h-1$, such that there are P out-of-sample point forecasts and R observations have been used for estimation of the model.

Under standard conditions, Diebold and Mariano (1995) established the asymptotic normality of the DM statistic. Two important concerns with the use of DM-type statistics, however, have appeared in the literature and we address those here. First, West (1996, 2001) and West and McCracken (1998) analyzed modification of forecast comparison tests in light of the use of estimated model parameters in the computation of such tests. West (1996) points out, though, that for DM-type tests under quadratic loss, such parameter estimation uncertainty is asymptotically irrelevant. Conditional on this result, van Dijk and Franses (2003) argued that when examining the statistical significance of MSPE reductions (which is what we are interested in), corrections of the type suggested by West (1996, 2001) and West and McCracken (1998) are not necessary.

Second, under the assumption that the estimation sample size R and the number of out-of-sample forecasts P tend to infinity, McCracken (2000) and Clark and McCracken (2001) showed that, if the underlying forecasting models are nested, the asymptotic distribution of the DM statistic is not standard normal. As noted by van Dijk and Franses (2003), these conditions on the parameters R and P effectively mean that expanding windows of data are used for estimation. In contrast, for the case in which R remains finite, as in our use of fixed-length rolling estimation windows, Giacomini and White (2006) proved that the asymptotic distribution of the DM statistic is standard normal when comparing forecasts generated by nested models. It does not necessarily follow, however, that this asymptotic approximation is a good one. Indeed, Clark and West (2007) demonstrate that it can be quite poor.[12]

Simulation evidence has shown that the size of the DM statistic is biased upwards in small samples. As such, Harvey *et al.* (1997) introduced a modification of the DM statistic (M-DM) to correct for this. Following Harvey *et al.* (1997), we use the Student's t distribution with $P-1$ degrees of freedom to obtain critical values for the M-DM tests we run.

van Dijk and Franses (2003) argued that the uniform weighting scheme employed by the M-DM test may be unsatisfactory for frequently encountered situations in which some observations are more important than others. For example, in an unemployment forecasting exercise of the type we analyze, large positive observations for the change in the unemployment rate generally signal a business cycle downturn.

Accordingly, van Dijk and Franses modified the Diebold-Mariano statistic by weighting more heavily the loss differentials for observations that are deemed to be of greater substantive interest. In their approach, the weighted average loss differential is given by

$$\bar{d}_w = \frac{1}{P} \sum_{t=R+h}^{R+P+h-1} w(\omega_t) d_t, \tag{6}$$

where ω_t is the information set available at time t. Letting y_t be the variable to be forecast, two particular cases van Dijk and Franses studied are

$$w_{\mathrm{LT}}(\omega_t) = 1 - \Phi(y_t), \tag{7}$$

where $\Phi(\cdot)$ is the cumulative distribution function of y_t, to focus on the left tail of the distribution of y_t, and

$$w_{\mathrm{RT}}(\omega_t) = \Phi(y_t), \tag{8}$$

to focus on the right tail of the distribution of y_t.

A necessary condition for the associated test statistic to have an asymptotic standard normal distribution under the null hypothesis of equal forecast accuracy is that the weight function $w(\omega_t)$ be a twice continuously differentiable mapping to the [0,1] interval. The weighted Diebold-Mariano statistic is computed as,

$$\text{W-DM} = \frac{\bar{d}_w}{\sqrt{\hat{V}(\bar{d}_w)}}, \tag{9}$$

where $\hat{V}(\bar{d}_w)$ is a consistent estimate of the variance of $\bar{d}_w$.

Following Harvey *et al.* (1997), van Dijk and Franses adjusted the W-DM statistic by way of a small-sample correction. The resulting modified W-DM statistic is given by

$$\text{MW-DM} = \sqrt{\frac{P + 1 - 2h + h(h-1)/P}{P}}\text{W-DM}. \tag{10}$$

Once again following Harvey *et al.* (1997), van Dijk and Franses proposed using the Student's $t-$distribution with $P-1$ degrees of freedom to obtain critical values for the MW-DM test.

To examine the statistical significance of MSPE reductions with greater weight placed on forecast losses associated with, respectively, unemployment rate decreases and increases, we apply the left-tailed and and right-tailed MW-DM tests. $p-$values for the M-DM, left-tailed MW-DM, and right-tailed MW-DM tests, respectively, appear in parentheses under the MSPE ratios in Tables 3-6.

White (2000) emphasized that muti-model out-of-sample forecasting comparisons of the type we carry out are exercises in "data snooping," since we use a given set of data more than once for tests against the benchmark. As such, statistically significant MSPE reductions we obtain may be due to chance, simply reflecting our use of a sufficiently large number of alternatives to the benchmark. White's (2000) "reality check" procedure allows one to account for the effects of data snooping in a study such as ours. The null hypothesis for the reality check test is that the benchmark model performs as well or better than all competitor models, and the alternative hypothesis is that the benchmark is dominated by at least one of the competitors. In the bottom rows of Tables 3-6, we report $p-$values for the Corradi and Swanson (2007) version of White's (2000) reality check test for our M-DM, left-tailed MW-DM, and right-tailed MW-DM forecast comparisons.

3.3.1 U.S. MSPE Ratios

Recalling that Method 11 is ranked first for the U.S. using the MSPE, we see that at each forecast step its MSPE is lower than that of the linear AR(p) benchmark Method 1; the average MSPE reduction across the eight forecast steps is 15%. Method 12, which is ranked second according to the MSPE, also generates a MSPE lower than Method 1's MSPE at each forecast step; the average MSPE reduction is 16% across the forecast steps. For no other forecasting method is the MSPE lower than Method 1's MSPE at each forecast horizon. In two cases, those of Methods 5 and 9, the MSPE performance is quite poor relative to the benchmark; on average the MSPE increase is roughly five-fold and double, respectively, across the forecast steps.

With the M-DM test Method 11's MSPE reduction relative to Method 1 is significant at the 10% significance level at only two of the forecast steps, $h = 4, 5$; in our discussion of the M-DM and WM-DM test results below, if the $p-$value for the test is less than or equal to 10% we say there is a "statistically significant" MSPE reduction. But when the median-pooling is extended across the linear and STVECM forecasts with Method 12, there are statistically significant MSPE reductions relative to the linear AR(p) case with uniform weighting of the forecast loss differentials at four of the forecast horizons, $h = 4, 5, 7, 8$. Our reality check test results, however, suggest that each of the statistically significant MSPE reductions we obtain with Methods 11 and 12 stems from data snooping, since the estimated $p-$values for the M-DM reality check tests at $h = 4, 5, 7, 8$ are all above 0.10.

Using the MW-DM test with greater weight given to unemployment rate decreases, Methods 11 and 12 both generate significant MSPE reductions relative to the linear AR(p) forecasts at all eight forecast steps. So, left-tailed weighting of the forecast loss differentials produces much more evidence of forecast improvement in comparison to uniform weighting. Method 2, based on the linear VECM forecasts, and Method 6 also generate significant MSPE reductions over the benchmark Method 1 at all forecast steps with left-tailed weighting. Statistically significant MSPE reductions relative to Method 1 are also produced by Methods 3, 7, and 8 at many forecast steps; 5, 7, and 5, respectively. Estimated $p-$values for the reality check on the left-tailed MW-DM test results are below

0.10 at five forecast horizons, $h = 1, 2, 5, 6, 7$.[13]

In sharp contrast to our left-tailed weighting results, with right-tailed weighting the MW-DM test produces no evidence of statistically significant MSPE reductions relative to the linear AR(p) benchmark. Similarly, the estimated $p-$values for the reality check on the right-tailed MW-DM results are all quite high.

3.3.2 U.K. MSPE Ratios

For the U.K., Methods 12 and 11, ranked first and third using the MSPE, each generate a lower MSPE relative to the linear AR(p) benchmark Method 1 at each forecast horizon; the average MSPE reductions across the forecast steps are, respectively, 47% and 45%. While we use Method 1 as the benchmark, it is useful to note that the linear VECM forecasts of Method 2 also produce a lower MSPE in comparison to Method 1 at each forecast step; the average MSPE reduction across the forecast horizons is 51%. For no other forecasting method is the MSPE lower than that of Method 1 at each forecast step; but Methods 4 and 5 each have an average MSPE ratio below 1, with an average MSPE reduction of 35% in the case of Method 4 and 17% in the case of Method 5.

With the M-DM test, the MSPE reductions over Method 1 produced by Methods 11 and 12 are statistically significant at all but the first of the eight forecast steps. The same is true for Method 2's linear VECM forecasts. Estimated $p-$values for the reality check test on the M-DM results are below 0.10 at five forecast horizons, $h = 3, 4, 5, 6, 7$.

Using the MW-DM test with left-tailed weighting, Method 12 generates statistically significant MSPE reductions relative to the benchmark Method 1 at all eight forecast steps, and Method 11 does so at seven forecast steps. The MSPE reductions relative to the benchmark produced by the linear VECM forecasts of Method 2 are statistically significant at all eight forecast horizons. But the reality check test $p-$values on the left-tail MW-DM results are below 0.10 at only two forecast horizons, $h = 4, 5$. So, in contrast to what happens in the U.S. case, left-tail weighting of the forecast loss differentials produces less evidence of forecast improvement in comparison to uniform weighting in the U.K. case.

With right-tailed weighting the MW-DM test finds statistically significant MSPE reductions for Method 12 relative to Method 1 at four forecast horizons, $h = 2, 3, 4, 5$, and for Method 11 at three forecast horizons, $h = 2, 3, 4$. But the reality check estimated $p-$values at these forecast steps are all above 0.10.

3.3.3 Canada MSPE Ratios

For Canada, Methods 12 and 11, ranked first and second using the MSPE, produce a lower MSPE relative to the benchmark at, respectively, six and five forecast horizons; the average MSPE reductions for these two pooling forecasting methods are, respectively, 5% and 1%. The average MSPE reductions generated by these forecasting methods are considerably lower in comparison to the U.S. and U.K. cases. For all other forecasting methods, the average MSPE across the forecast horizons is higher than the linear AR(p) benchmark's average MSPE.

With the M-DM test, Method 12 generates statistically significant MSPE reductions relative to Method 1 at two forecast steps, $h = 5, 8$, and Method 11 does so at one forecast step, $h = 5$. The estimated $p-$values for the reality check at $h = 5, 8$ are above 0.10.

Using the MW-DM test with greater weight given to unemployment rate decreases, Method 12 produces statistically significant MSPE reductions at five forecast horizons, $h = 4, 5, 6, 7, 8$, while Method 11 does so at six forecast horizons, $h = 3, 4, 5, 6, 7, 8$. Method 2's linear VECM forecasts also generate statistically significant MSPE reductions at $h = 4, 5, 6, 7, 8$. The estimated $p-$value for the reality check test is below 0.10 at only one of these forecast steps, $h = 3$.

With right-tailed weighting the MW-DM test produces no evidence of statistically significant MSPE reductions in comparison to the linear AR(p) benchmark. The estimated $p-$values for the reality check on these right-tailed MW-DM results are all very high.

3.3.4 Japan MSPE Ratios

For Japan, there are only two cases (out of $88 = 11$ methods $\times$ 8 forecast steps), $h = 4$ for the linear VECM Method 2 and $h = 3$ for Method 12, in which the linear AR(p) benchmark does not generate a lower MSPE relative to the alternative forecasting method. Accordingly, all of the average MSPE ratios reported in Table 6 are greater than 1, a set of results consistent with the linear benchmark Method 1 being top-ranked using the MSPE.

With uniform, left-tailed, and right-tailed weighting of the forecast loss differentials, at no forecast horizon does either of the STVECM pooling methods, Methods 11 and 12, produce a statistically significant MSPE reduction. Further, all of the estimated $p-$values for the reality check tests are above 0.10.

4 Conclusions

In this paper we set out to explore how a set of multivariate STAR models performs, both against a linear benchmark and relative to one another, in simulated real-time out-of-sample forecasting of the four non-Euro G-7 quarterly aggregate unemployment rate series. Consideration of this issue appears warranted in light of work in the empirical literature on business cycle asymmetry, in which a good deal of evidence that the data generating process for many unemployment rate series may indeed be nonlinear has been reported.

Our out-of-sample results show that, according to the MSPE, the top-ranked forecasting method for the U.S., U.K., and Canada is a pooled-median forecasting approach. For the U.S., the top-ranked forecasting method uses the median across the set of STVECM point forecasts; for the U.K. and Canada, forecasting with the median across the set of linear and nonlinear point forecasts performs best. These multivariate pooling results are consistent with those reported by Stock and Watson (1999) and Marcellino (2004) in their analysis of univariate nonlinear models. For Japan, the linear AR(p) forecasts are top-ranked using the MSPE.

The strongest evidence in favor of the pooled-median forecasts generating statistically significant MSPE reductions over the linear benchmark is provided by the U.S. and U.K. cases. For the U.S., this occurs when we use a recently developed test of forecast accuracy which places more weight on the forecast loss differentials associated with extreme values of the unconditional distribution of the unemployment rate first differences. More specifically, when unemployment rate decreases are emphasized, both pooling-median forecasting methods produce a statistically significant lower MSPE than that generated by the linear AR(p) forecasts at all forecast steps examined. A White (2000) reality check of these results suggests that, for five of these eight forecast horizons, this outcome is not due to data snooping. It appears, then, that the STVECM forecasts perform better during business cycle expansions for the U.S.

For the U.K., the strongest evidence of forecast improvement with the pooled-median forecasts occurs with uniform weighting of the forecast loss differentials. In this case, both pooling approaches generate statistically significant MSPE reductions at seven out of the eight forecast steps, and the reality check implies that, for five of these seven forecast horizons, these results do not stem from data snooping. When unemployment rate decreases are weighted more heavily in the forecast accuracy tests, median-pooling across the linear and nonlinear forecasts generates statistically significant MSPE reductions over the linear benchmark at all forecast horizons. However, the reality check on these results suggests that most of them are due to data snooping.

We believe the main message from our forecasting exercise is as follows. While individual nonlinear forecasting methods may rarely dominate a linear approach, forecast improvement seems attainable by combining across the set of linear and nonlinear forecasts; our pooling results with the multivariate models we use mirrors a similar forecast combination finding with univariate nonlinear models obtained by Teräsvirta, van Dijk, and Medeiros (2005). Noting that in this paper we restrict ourselves to STAR-type multivariate models, we speculate that pooling linear forecasts with a larger set of nonlinear alternatives would prove to be useful. We intend to pursue this question in further research.

Among the set of STVECM forecasting methods used, we find that no individual approach tends to outperform the others. In some cases, the top-ranked nonlinear forecasting method employs multi-step-ahead forecasts obtained by iterating the estimated one-step-ahead model. In others, $h-$step-ahead projections dominate. As a result, at least for the data sets examined in this paper, it appears that use of both approaches is warranted. We note that these results stand in contrast with those in Marcellino, Stock, and Watson (2006), who, in their linear study with U.S. macroeconomic time series, found that iterated forecasts generally dominated $h-$step-ahead projections.

In this paper we compare the point forecasts of the models used. Thus, it would be interesting to investigate the robustness of our results with respect to construction and evaluation of both interval and density forecasts. Clements and Hendry (1999, p. 285), for example, suggest that use of interval and density forecasts may indeed show improved forecasting performance for nonlinear models. We note, however, that Clements, Franses, Smith, and D. van Dijk (2003) report simulation results which suggest that the Diebold and Mariano (1995) test is in fact more powerful than interval and density forecast-based tests in discriminating between linear and nonlinear models.

Table 1: Forecasting Method Definitions

Method	Definition
1	Linear AR(p).
2	Unrestricted VECM.
3	STVECM, with transition variable selected by S_1 test run on first-differenced log-unemployment equation of subset linear VECM.
4	STVECM, with transition variable selected by S_2 test run on first-differenced log-unemployment equation of subset linear VECM.
5	STVECM, with transition variable selected by S_3 test run on first-differenced log-unemployment equation of subset linear VECM.
6	STVECM, with transition variable selected by system-wide S_1 test.
7	STVECM, with transition variable selected by system-wide S_3 test.
8	h-step-ahead projection of STVECM's first differenced log-unemployment rate equation, with transition variable selected by S_1 test run on corresponding subset equation of VECM.
9	h-step-ahead projection of STVECM's first-differenced log-unemployment rate equation, with transition variable selected by S_2 test run on corresponding subset equation of VECM.
10	h-step-ahead projection of STVECM's first-differenced log-unemployment rate equation, with transition variable selected by S_3 test run on first-differenced log-unemployment subset equation of VECM.
11	Pooled median forecast from nonlinear methods, i.e., Methods 3 through 10.
12	Pooled median forecast from Methods 1 through 10.

Table 2: Average Out-of-Sample Forecasting Ranks

Method i	U.S.	U.K.	Canada	Japan
		MSPE		
1	6.1	8.0	4.0	1.3
2	6.0	2.5	5.5	2.9
3	7.9	5.4	7.3	7.6
4	9.1	4.9	6.9	7.6
5	11.8	6.8	6.1	7.5
6	4.0	9.0	8.9	10.1
7	4.5	9.0	7.8	9.6
8	6.8	9.5	8.8	8.6
9	10.5	9.0	8.8	8.0
10	7.8	9.4	8.1	7.4
11	1.8	3.0	3.9	4.8
12	1.9	1.6	2.1	2.6
		MedSPE		
1	11.3	9.6	5.0	3.8
2	7.8	3.4	7.3	2.5
3	7.4	4.6	7.8	9.3
4	7.9	4.4	4.6	7.5
5	7.3	5.6	5.6	8.9
6	4.5	5.1	8.6	9.6
7	5.9	6.0	8.5	6.9
8	5.1	10.3	8.0	5.9
9	8.4	10.3	6.6	7.3
10	5.6	9.9	7.1	5.3
11	2.8	5.1	5.0	6.6
12	4.3	3.8	3.9	4.6

The two panels show the average out-of-sample forecasting rank of Method i across the 49 estimation windows and forecasting horizons $h = 1, \ldots, 8$, using the Mean Squared Prediction Error (MSPE) and Median Squared Prediction Error (MedSPE) criteria. See Table 1 for the forecasting method definitions.

Table 3: MSPE Ratios for U.S.

	Forecast horizon h								
Method i	1	2	3	4	5	6	7	8	Avg.
1	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	
2	1.056 (0.60,**0.02**,0.89)	1.090 (0.65,**0.02**,0.96)	1.027 (0.54,∗,0.85)	0.886 (0.34,∗,0.74)	0.890 (0.30,∗,0.76)	1.001 (0.50,∗,0.83)	0.895 (0.13,∗,0.90)	0.874 (**0.04**,∗,0.89)	0.965
3	1.207 (0.73,**0.03**,0.89)	1.355 (0.88,0.43,0.94)	0.769 (**0.10**,∗,0.67)	1.098 (0.58,**0.01**,0.78)	1.095 (0.60,**0.04**,0.82)	1.201 (0.73,**0.01**,0.87)	1.167 (0.80,0.22,0.89)	1.161 (0.93,1.00,0.86)	1.132
4	1.030 (0.56,**0.07**,0.86)	2.116 (0.84,0.13,0.92)	0.816 (0.15,∗,0.61)	1.119 (0.61,**0.04**,0.82)	1.643 (0.74,0.71,0.80)	2.718 (0.84,0.83,0.86)	1.331 (0.84,0.64,0.96)	1.982 (0.85,0.83,0.87)	1.594
5	3.509 (0.90,0.84,0.93)	4.424 (0.86,0.71,0.88)	5.482 (0.83,0.81,0.85)	3.923 (0.81,0.57,0.83)	2.351 (0.78,0.74,0.82)	2.627 (0.81,0.26,0.84)	6.235 (0.83,0.83,0.83)	10.498 (0.84,0.84,0.79)	4.881
6	0.993 (0.49,**0.09**,0.75)	1.031 (0.56,∗,0.93)	0.851 (0.23,∗,0.76)	0.809 (0.17,∗,0.71)	0.810 (0.13,∗,0.76)	0.919 (0.33,∗,0.82)	0.861 (0.16,∗,0.84)	0.894 (**0.09**,∗,0.91)	0.896
7	0.975 (0.45,**0.09**,0.76)	1.035 (0.57,**0.01**,0.93)	0.802 (0.13,∗,0.74)	0.770 (0.12,∗,0.68)	0.775 (**0.08**,∗,0.72)	1.023 (0.54,0.14,0.83)	0.888 (0.22,∗,0.81)	0.905 (0.16,∗,0.89)	0.897
8	1.089 (0.64,0.24,0.81)	1.143 (0.77,**0.03**,0.97)	0.944 (0.40,**0.02**,0.85)	1.126 (0.61,0.42,0.85)	0.974 (0.47,**0.02**,0.90)	0.982 (0.47,**0.06**,0.87)	0.878 (0.15,∗,0.97)	1.061 (0.58,0.11,0.90)	1.025
9	1.198 (0.80,0.29,0.95)	2.405 (0.90,0.60,0.95)	1.051 (0.59,0.11,0.88)	4.494 (0.85,0.82,0.91)	1.784 (0.78,0.75,0.84)	2.506 (0.80,0.78,0.84)	1.145 (0.63,0.44,0.88)	2.052 (0.87,0.77,0.94)	2.079
10	1.270 (0.85,0.52,0.94)	1.014 (0.53,**0.01**,0.93)	0.981 (0.47,**0.03**,0.90)	1.088 (0.58,0.39,0.83)	1.130 (0.64,0.26,0.91)	1.077 (0.60,**0.08**,0.93)	0.893 (0.27,**0.05**,0.77)	1.316 (0.81,**0.08**,0.95)	1.096
11	0.935 (0.38,**0.02**,0.77)	0.964 (0.42,∗,0.92)	0.785 (0.13,∗,0.71)	0.710 (**0.10**,∗,0.62)	0.754 (**0.09**,∗,0.70)	0.896 (0.31,∗,0.78)	0.842 (0.11,∗,0.80)	0.899 (0.11,∗,0.91)	0.848
12	0.900 (0.31,∗,0.75)	0.958 (0.41,∗,0.91)	0.789 (0.12,∗,0.70)	0.715 (**0.09**,∗,0.61)	0.756 (**0.08**,∗,0.67)	0.899 (0.30,∗,0.77)	0.843 (**0.08**,∗,0.82)	0.891 (**0.07**,∗,0.88)	0.844
Snoop	(0.12,**0.08**,0.90)	(0.15,**0.08**,0.81)	(**0.03**,0.18,0.44)	(0.12,0.11,0.42)	(0.11,**0.09**,0.59)	(0.13,**0.09**,0.93)	(0.12,**0.05**,0.93)	(0.16,0.14,0.95)	

First row presents MSPE for forecasting Method 1 at forecasting horizons $h = 1, \ldots, 8$. The top entry in rows 2 through 12 shows, for each forecast step, the ratio of Method i's MSPE to the MSPE of Method 1; the final column lists the average of these MSPE ratios across the forecast steps. The parentheses in these rows report, from left to right, the p-value for the M-DM, left-tailed WM-DM, and right-tailed WM-DM tests; '*****' indicates that the p-value is less than 0.01. Estimated p-values for the Corradi and Swanson (2007) implementation of White's (2000) data snooping "reality check" test of the null hypothesis that the best of Methods 2 through 12 at forecasting horizon h improves over Method 1, according to, respectively, the M-DM, left-tailed WM-DM, and right-tailed WM-DM tests, are given from left to right in the bottom row.

Table 4: MSPE Ratios for U.K.

Method i	Forecast horizon h: 1	2	3	4	5	6	7	8	Avg.
1	0.001	0.001	0.001	0.002	0.002	0.002	0.002	0.002	
2	0.889 (0.30,**0.06**,0.92)	0.579 (**0.04**,**0.01**,0.46)	0.463 (**0.03**,**0.02**,0.20)	0.399 (**0.02**,**0.01**,0.16)	0.372 (**0.02**,**0.01**,0.14)	0.396 (**0.03**,**0.01**,0.16)	0.398 (**0.03**,**0.01**,0.20)	0.426 (**0.02**,*,0.21)	0.490
3	0.871 (0.29,0.15,0.74)	0.556 (**0.01**,**0.01**,0.14)	0.431 (**0.02**,**0.03**,**0.07**)	0.576 (0.15,**0.07**,0.38)	0.833 (0.39,0.14,0.65)	1.250 (0.62,0.48,0.72)	1.097 (0.55,0.92,0.68)	1.108 (0.56,0.44,0.69)	0.840
4	1.297 (0.81,0.68,0.96)	0.635 (**0.05**,**0.01**,0.47)	0.556 (**0.06**,**0.03**,0.37)	0.458 (**0.02**,**0.01**,**0.10**)	0.469 (**0.02**,**0.02**,**0.09**)	0.486 (**0.05**,**0.04**,0.11)	0.637 (0.16,0.14,0.26)	0.641 (0.17,0.12,0.33)	0.647
5	0.928 (0.37,0.19,0.76)	0.595 (**0.04**,**0.02**,0.29)	0.468 (**0.03**,**0.02**,0.18)	0.615 (0.15,**0.05**,0.54)	0.844 (0.38,0.12,0.70)	1.031 (0.52,0.21,0.72)	1.037 (0.52,0.33,0.73)	1.150 (0.59,0.45,0.76)	0.834
6	0.926 (0.35,0.14,0.90)	0.576 (**0.02**,**0.01**,0.36)	0.667 (0.19,**0.02**,0.69)	1.498 (0.66,**0.01**,0.80)	1.985 (0.73,**0.04**,0.81)	1.591 (0.69,0.48,0.78)	1.231 (0.62,0.39,0.77)	3.054 (0.78,0.72,0.81)	1.441
7	0.966 (0.44,0.20,0.90)	0.656 (**0.10**,**0.06**,0.39)	0.626 (0.16,**0.02**,0.65)	1.504 (0.66,**0.01**,0.80)	1.835 (0.70,**0.02**,0.80)	1.430 (0.64,0.38,0.76)	1.153 (0.57,0.33,0.73)	2.593 (0.74,0.64,0.79)	1.345
8	2.087 (0.99,0.95,1.00)	1.147 (0.68,0.42,0.97)	0.979 (0.46,0.15,0.95)	2.412 (0.79,0.71,0.84)	1.213 (0.70,0.54,0.87)	0.956 (0.44,0.36,0.57)	0.703 (0.20,0.12,0.41)	4.404 (0.93,0.94,0.89)	1.738
9	1.962 (0.95,0.88,1.00)	0.824 (0.19,**0.03**,0.82)	1.393 (0.74,0.39,0.93)	0.939 (0.45,0.20,0.74)	1.944 (0.82,0.76,0.87)	0.906 (0.38,0.33,0.45)	0.766 (0.23,0.23,0.28)	2.631 (0.86,0.87,0.85)	1.421
10	1.841 (0.98,0.89,1.00)	1.073 (0.59,0.38,0.93)	1.384 (0.73,0.38,0.93)	2.033 (0.75,0.61,0.83)	1.298 (0.75,0.66,0.84)	0.818 (0.26,0.22,0.34)	0.734 (0.21,0.12,0.45)	4.474 (0.94,0.94,0.93)	1.707
11	0.875 (0.28,0.14,0.75)	0.472 (*,*,**0.09**)	0.437 (**0.02**,**0.02**,**0.09**)	0.375 (*,*,**0.04**)	0.513 (**0.04**,**0.01**,0.20)	0.570 (**0.08**,**0.03**,0.23)	0.568 (**0.06**,**0.04**,0.17)	0.597 (**0.08**,**0.04**,0.19)	0.551
12	0.810 (0.17,**0.08**,0.61)	0.469 (*,*,**0.07**)	0.426 (**0.01**,**0.02**,**0.07**)	0.372 (*,*,**0.03**)	0.464 (**0.02**,**0.01**,**0.08**)	0.550 (**0.06**,**0.03**,0.19)	0.555 (**0.05**,**0.03**,0.13)	0.567 (**0.05**,**0.02**,0.16)	0.527
Snoop	(0.24,0.18,0.67)	(0.81,0.92,0.14)	(**0.01**,0.91,0.22)	(**0.08**,**0.07**,0.12)	(**0.09**,**0.02**,0.15)	(**0.05**,0.87,**0.10**)	(**0.03**,0.91,**0.07**)	(0.22,0.22,0.15)	

See notes for Table 3.

Table 5: MSPE Ratios for Canada

Method i	Forecast horizon h: 1	2	3	4	5	6	7	8	Avg.
1	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	
2	1.227 (0.80,0.29,0.99)	1.142 (0.69,0.21,0.98)	1.068 (0.61,0.11,0.95)	1.018 (0.53,**0.05**,0.94)	1.116 (0.65,**0.09**,0.96)	1.131 (0.71,**0.09**,0.98)	1.026 (0.56,∗,0.90)	0.812 (**0.10**,∗,0.69)	1.068
3	1.952 (0.96,0.78,0.99)	1.597 (0.96,0.57,0.99)	1.571 (0.96,0.46,0.99)	2.820 (0.94,0.85,0.98)	0.997 (0.49,∗,0.89)	1.038 (0.63,∗,0.96)	0.893 (**0.01**,∗,0.65)	0.895 (∗,**0.02**,0.38)	1.470
4	1.704 (0.92,0.76,0.99)	1.600 (0.90,0.56,0.96)	1.551 (0.90,0.78,0.96)	3.901 (0.98,0.93,0.99)	0.811 (∗,0.50,**0.09**)	0.839 (∗,**0.01**,0.83)	1.008 (0.60,0.59,0.54)	0.982 (0.39,0.45,0.44)	1.550
5	1.902 (0.95,0.76,0.98)	1.139 (0.69,0.23,0.99)	1.446 (0.83,0.64,0.99)	3.743 (0.93,0.91,0.98)	0.888 (0.17,∗,0.90)	1.185 (0.85,0.34,0.96)	0.877 (∗,∗,0.69)	0.863 (∗,∗,0.49)	1.505
6	1.224 (0.80,0.35,1.00)	1.112 (0.66,0.20,0.99)	1.243 (0.85,0.19,1.00)	2.032 (0.90,0.71,1.00)	2.618 (0.85,0.77,1.00)	1.823 (0.81,0.64,0.98)	2.798 (0.83,0.72,0.92)	8.383 (0.83,0.81,0.85)	2.654
7	1.196 (0.77,0.27,0.99)	1.160 (0.74,0.25,1.00)	1.119 (0.69,0.13,0.99)	1.191 (0.76,0.18,1.00)	1.224 (0.89,0.23,1.00)	1.277 (0.83,0.24,0.99)	1.626 (0.88,0.64,0.97)	14.510 (0.84,0.82,0.85)	2.913
8	2.225 (0.89,0.79,0.98)	1.323 (0.81,0.23,0.99)	0.778 (0.17,**0.03**,0.90)	1.557 (0.90,0.22,0.99)	1.075 (0.71,0.36,0.94)	3.785 (0.98,0.95,1.00)	1.938 (0.98,0.91,1.00)	2.268 (0.93,0.83,0.93)	1.869
9	2.114 (0.88,0.84,0.93)	1.381 (0.94,0.76,0.95)	0.884 (0.24,**0.08**,0.77)	1.548 (0.92,0.73,0.93)	1.173 (0.94,0.83,0.84)	3.420 (1.00,0.99,0.98)	1.838 (1.00,0.98,0.99)	2.223 (0.97,0.93,0.96)	1.823
10	2.258 (0.90,0.77,0.99)	1.301 (0.81,0.24,0.99)	0.774 (0.16,**0.03**,0.91)	1.352 (0.71,0.13,0.87)	1.265 (0.87,0.48,0.92)	2.626 (0.96,0.92,1.00)	1.796 (0.99,0.92,1.00)	2.120 (0.94,0.82,0.95)	1.687
11	1.154 (0.72,0.25,0.99)	1.086 (0.63,0.15,0.98)	0.898 (0.32,**0.05**,0.96)	0.965 (0.42,**0.01**,0.99)	0.855 (**0.07**,**0.01**,1.00)	0.992 (0.47,**0.04**,0.97)	1.030 (0.63,**0.09**,0.96)	0.901 (0.15,**0.01**,0.74)	0.985
12	1.113 (0.67,0.21,0.99)	1.053 (0.59,0.14,0.97)	0.853 (0.25,**0.93**,**0.07**)	0.921 (0.30,∗,0.99)	0.839 (**0.04**,**0.01**,1.00)	0.971 (0.40,**0.02**,0.97)	0.996 (0.48,**0.02**,0.94)	0.855 (**0.03**,∗,0.59)	0.950
Snoop	(0.66,0.11,0.87)	(0.41,**0.03**,0.82)	(0.24,**0.05**,0.70)	(0.21,0.15,0.54)	(0.18,0.11,0.44)	(0.32,0.31,0.32)	(0.37,0.20,0.69)	(0.22,0.14,0.24)	

See notes for Table 3.

Table 6: MSPE Ratios for Japan

Method i	Forecast horizon h: 1	2	3	4	5	6	7	8	Avg.
1	0.001	0.001	0.001	0.001	0.001	0.002	0.002	0.002	
2	1.213 (0.97,0.96,0.88)	1.089 (0.84,0.86,0.64)	1.037 (0.70,0.78,0.67)	0.960 (0.28,**0.06**,0.46)	1.018 (0.58,0.29,0.66)	1.033 (0.61,0.17,0.76)	1.053 (0.65,0.24,0.72)	1.021 (0.54,0.12,0.70)	1.053
3	1.207 (0.93,0.90,0.87)	1.372 (0.84,0.86,0.824)	1.143 (0.88,0.84,0.83)	1.728 (0.89,0.88,0.74)	1.635 (0.96,0.94,0.99)	2.473 (0.89,0.88,0.85)	2.928 (0.94,0.91,0.98)	10.581 (0.95,0.92,0.87)	3.008
4	1.664 (0.90,0.80,0.90)	1.497 (0.90,0.89,0.90)	1.195 (0.97,0.91,0.96)	1.258 (0.98,0.98,0.97)	1.400 (0.99,0.99,0.95)	1.715 (0.96,0.94,0.86)	2.403 (0.95,0.86,0.97)	10.165 (0.85,0.82,0.83)	2.662
5	1.142 (0.89,0.86,0.83)	1.474 (0.89,0.94,0.86)	1.019 (0.59,0.70,0.49)	1.829 (0.89,0.89,0.87)	1.776 (0.98,0.95,0.99)	2.549 (0.91,0.90,0.96)	4.051 (0.96,0.92,0.50)	8.125 (0.97,0.92,0.89)	2.746
6	1.690 (0.91,0.93,0.89)	1.518 (0.90,0.95,0.87)	1.062 (0.82,0.76,0.79)	2.071 (0.91,0.88,0.95)	1.940 (0.98,0.94,0.99)	2.601 (0.94,0.90,0.90)	4.067 (0.97,0.92,0.99)	13.044 (0.92,0.94,0.85)	3.499
7	1.664 (0.90,0.92,0.88)	1.524 (0.88,0.92,0.86)	1.153 (0.85,0.74,0.87)	2.117 (0.89,0.87,0.89)	1.812 (0.94,0.89,0.88)	2.729 (0.95,0.90,0.88)	3.562 (0.93,0.91,0.98)	4.158 (0.93,0.86,0.95)	2.340
8	1.407 (0.98,0.97,0.94)	1.280 (0.97,0.96,0.90)	1.146 (0.98,0.97,0.67)	3.302 (0.84,0.83,0.84)	3.462 (0.86,0.88,0.85)	7.342 (0.84,0.85,0.84)	2.384 (0.92,0.93,0.90)	1.555 (0.75,0.79,0.71)	2.735
9	1.275 (0.93,0.89,0.92)	1.351 (0.98,0.90,0.86)	1.244 (0.91,0.90,0.89)	1.383 (0.91,0.97,0.88)	3.210 (0.86,0.90,0.84)	2.103 (0.85,0.91,0.83)	5.483 (0.85,0.86,0.84)	1.686 (0.81,0.91,0.74)	2.217
10	1.344 (0.96,0.97,0.94)	1.402 (0.99,0.97,0.90)	1.032 (0.59,0.77,0.48)	1.435 (0.76,0.82,0.67)	3.301 (0.84,0.84,0.84)	7.340 (0.84,0.85,0.83)	2.368 (0.91,0.93,0.90)	1.546 (0.75,0.84,0.89)	2.471
11	1.095 (0.79,0.72,0.75)	1.360 (0.83,0.86,0.81)	1.012 (0.56,0.64,0.50)	1.307 (0.92,0.90,0.82)	1.517 (0.99,0.98,0.88)	1.800 (0.96,0.93,0.80)	2.245 (0.96,0.95,0.93)	3.317 (0.97,0.91,0.69)	1.707
12	1.061 (0.75,0.65,0.73)	1.097 (0.75,0.74,0.74)	0.990 (0.36,0.54,0.23)	1.044 (0.86,0.82,0.82)	1.245 (0.91,0.98,0.84)	1.318 (0.87,0.91,0.78)	1.282 (0.90,0.86,0.79)	1.441 (0.93,0.99,0.81)	1.185
Snoop	(0.94,0.67,0.85)	(0.98,0.90,0.85)	(0.20,0.29,0.16)	(0.20,0.29,0.24)	(0.55,0.34,0.87)	(0.54,0.22,0.91)	(0.65,0.21,0.61)	(0.34,0.26,0.69)	

See notes for Table 3.

Notes

[1]See Table 11 of Marcellino (2006).

[2]Other papers which study the performance of nonlinear time series models in forecasting unemployment rate fluctuations include Peel and Speight (2000), Terui and van Dijk (2002), and Proietti (2003).

[3]In both Rothman *et al.* (2001) and here, revised as opposed to real-time or preliminary data are used; see, for example, Amato and Swanson (2001) for discussion of the distinction between these. The main reason we use revised data is that real-time data sets for all of the systems we estimate are not available. While this certainly warrants the standard caveat about our results, what we do is also consistent with, for example, Stock and Watson (1999) and Stock and Watson (2003).

[4]In this paper, which appears in a special issue of *Empirical Economics* devoted to recent developments in modelling business cycle and financial data with regime-switching models, the one-step-ahead forecasts were computed for a given parametric structure via an updating of the regime-dependent probabilities.

[5]We decided to work with log-unemployment rate data following the theoretical framework presented in Nickell (1998) and in order to reduce the heteroskedasticity of the residuals in our estimated models.

[6]The money supply and interest rate series used are M2 and the 90-day Treasury bill rate for the U.S., M4 and the 90-day Treasury bill rate for the U.K., M2 and the 90-day commercial paper rate for Canada, and M2 and the lending rate for collateral and overnight loans in the Tokyo call money market for Japan. We use M4 data for the UK since the M2 time series is incomplete and inconsistent. We use the overnight Tokyo call rate for Japan since no sufficient long 90-day rate is available; see, for example, Table 9 of Stock and Watson (2003).

[7]This procedure leads to the model that would be selected by applying the AIC to each equation individually. But it is not guaranteed that this model also minimizes the AIC for the system as a whole. The simulation evidence in Brüggemann and Lütkepohl (2001), however, shows that the difference between the models selected by this single equation approach and a comparable system approach is generally small.

[8]We found, by way of generating alternative sets of forecasts, first with u_t removed from $\mathbf{z}_t$ and then with i_t removed from $\mathbf{z}_t$, that our results are robust with respect to the assumption that the unemployment rate and interest rate are $I(1)$ instead of $I(0)$. In addition, we found that our results are not sensitive to our $I(0)$ assumption for the inflation rate by examining forecasts using the change in inflation instead of its level.

[9]More details on these system-wide versions of the Luukkonen *et al.* (1988) tests can be found in Rothman *et al.* (2001).

[10]The reason why we do not use analogues of the federal funds rate for the U.K. and Canada is that, if we were to do so, this would shorten considerably the available time series; see, once again, Table 9 of Stock and Watson (2003). Also, as noted earlier, our

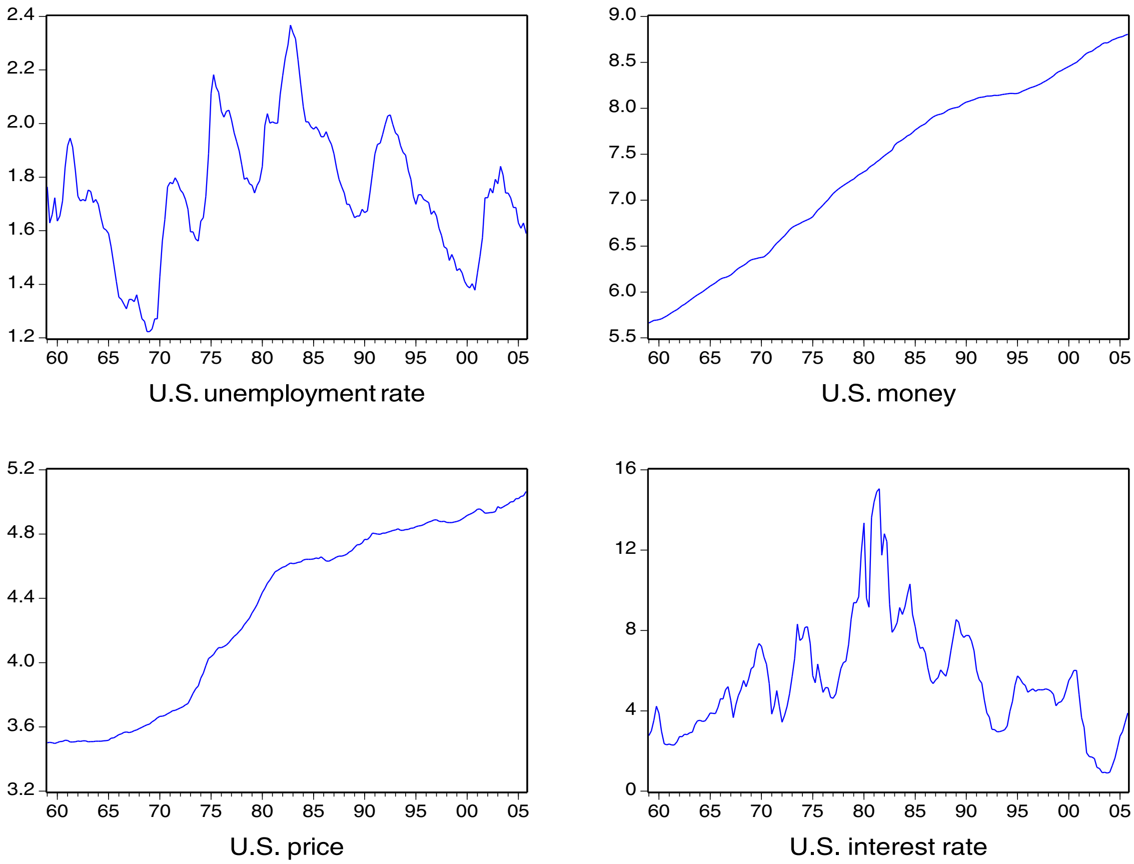
short-term interest rate for Japan is an overnight rate.

[11]Details on all top-ranked transition variables used in this paper, along with p-values of the associated linearity tests, are available upon request. Since our primary interest in this paper is on out-of-sample forecasting, we do not focus on the linearity testing results here. That said, we note that the battery of linearity tests run reveal strong evidence in favor of STAR-type nonlinearity and that the rankings of the candidate transition variables vary a good deal across the particular tests employed and the individual unemployment rate series examined.

[12]We thank one of our referees for pointing this out to us. Our decision to use sequences of fixed-length rolling windows, as opposed to sequences of expanding windows in which the sample size for estimation is increased by one observation in each sucessive window, was based on the Giacomini and White (2006) result that the DM statistic asymptotically distributed standard normal. However, we obtain quite similar results when we use expanding windows of data. On the fixed-length rolling window versus expanding window question, we note: (a) Stock and Watson (2005, p. 26) report that, for the representative macroeconomic dataset they study, "recursive forecasts are more accurate than the rolling forecasts"; and (b) in contrast, Giacomini and White (2006, p. 1566) find that a "rolling window procedure can result in substantial forecast accuracy gains relative to an expanding window for important economic time series."

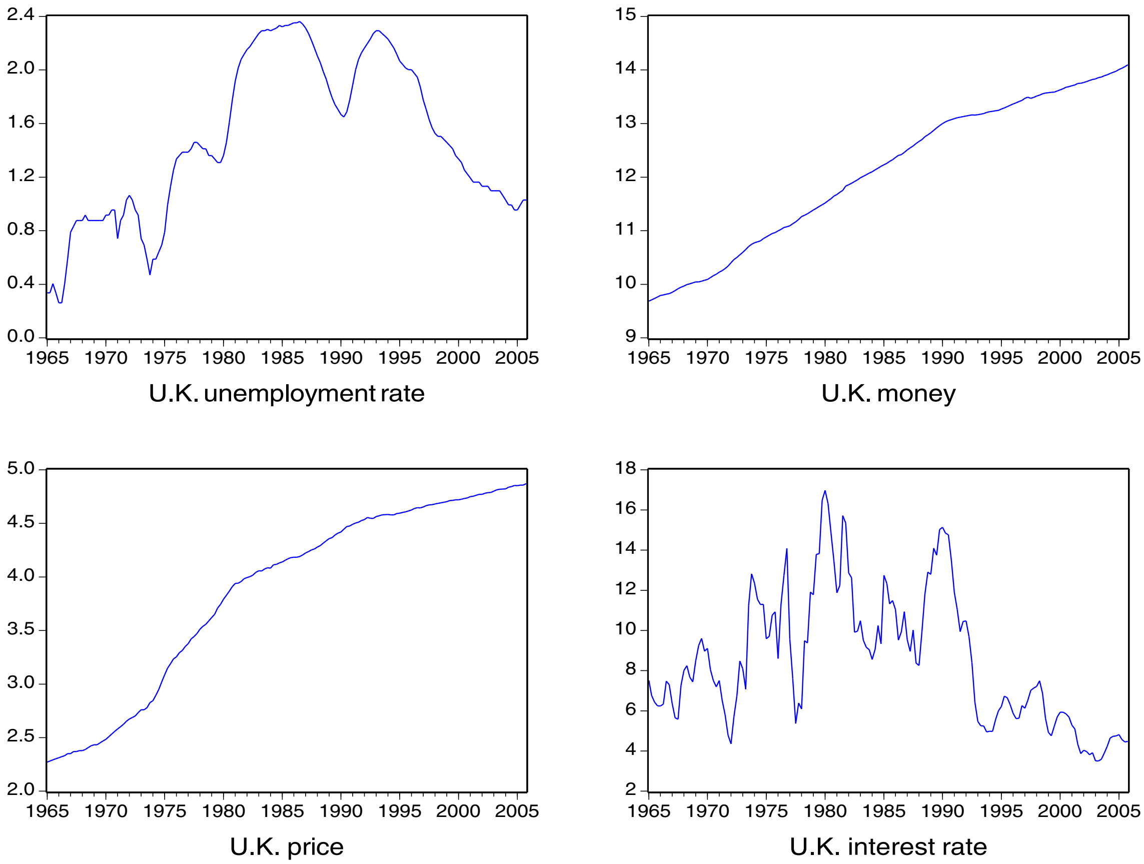
[13]Following up on a suggestion made by a referee, we examined whether the stronger evidence of statistically significant MSPE reductions over the linear AR(p) benchmark with left-tailed as opposed to uniform weighting of the forecast loss differentials might be explained by forecast bias. However, we did not find strong evidence of forecast error bias for the pooling forecasting methods, Methods 11 and 12, for the U.S.; there is only one case (out of $16 = 2$ methods $\times$ 8 forecast steps) in which the $p-$value for the forecast bias test is below 0.10.

Figure 1: Time Series Plots of U.S. Data



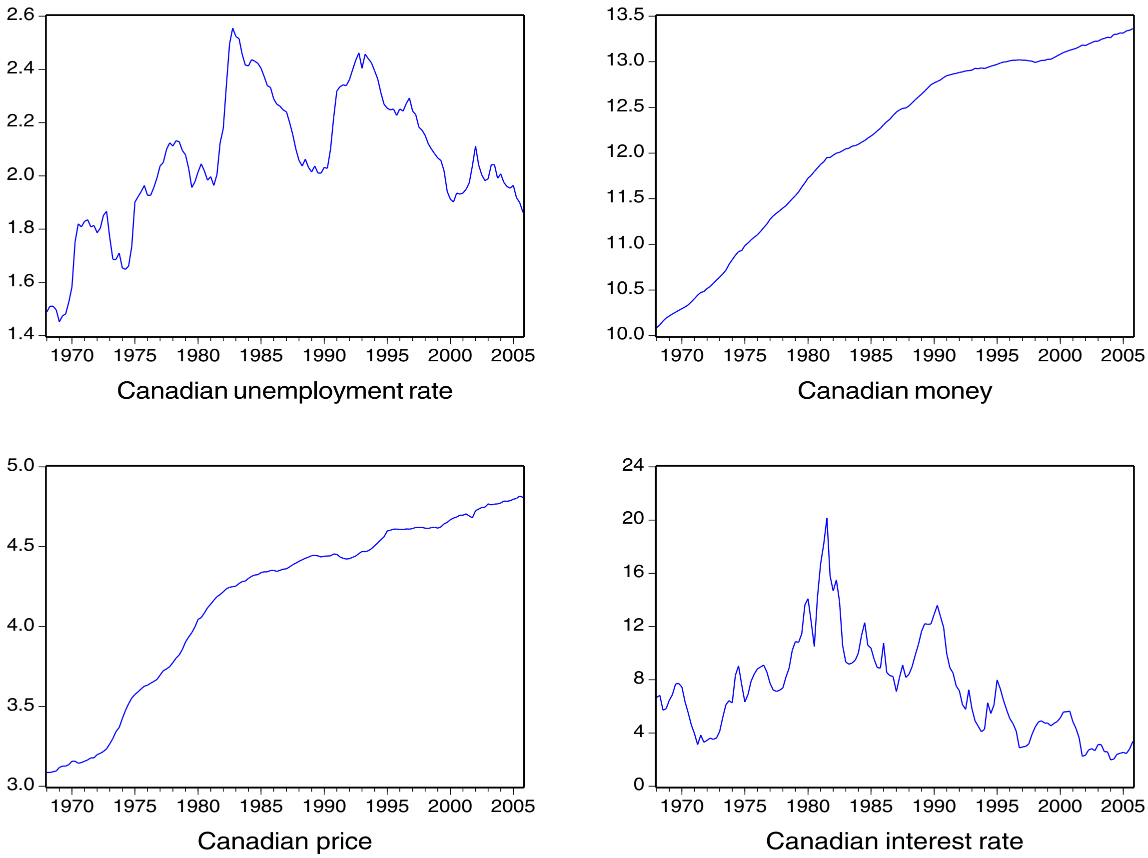
Time series plots of u_t, m_t, p_t, and i_t for the U.S. are presented in, respectively, the top-left, top-right, bottom-left, and bottom-right panels.

Figure 2: Time Series Plots of U.K. Data



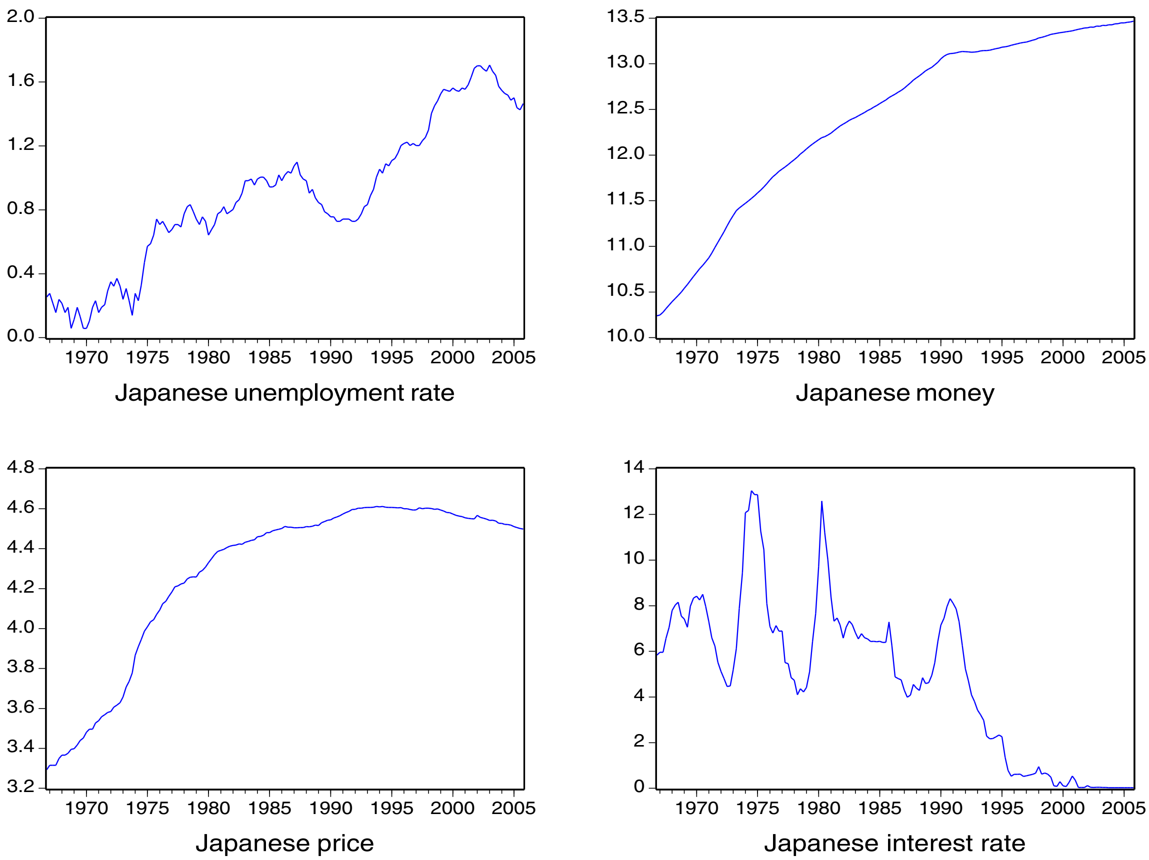
Time series plots of u_t, m_t, p_t, and i_t for the U.K. are presented in, respectively, the top-left, top-right, bottom-left, and bottom-right panels.

Figure 3: Time Series Plots of Canadian Data



Time series plots of u_t, m_t, p_t, and i_t for Canada are presented in, respectively, the top-left, top-right, bottom-left, and bottom-right panels.

Figure 4: Time Series Plots of Japanese Data



Time series plots of u_t, m_t, p_t, and i_t for Japan are presented in, respectively, the top-left, top-right, bottom-left, and bottom-right panels.

References

Altissimo, F. and G.L. Violante (2001), "The Non-Linear Dynamics of Output and Unemployment in the U.S.," *Journal of Applied Econometrics* 16, 461–486.

Amato, J.D. and N.R. Swanson (2001), "The Real-Time Predictive Content of Money for Output," *Journal of Monetary Economics* 48, 3–24.

Balke, N.S. and T.B. Fomby (1997), "Threshold Cointegration," *International Economic Review* 38, 627–646.

Beaudry, P. and G. Koop (1993), "Do Recessions Permanently Affect Output?" *Journal of Monetary Economics* 31, 149–163.

Brüggemann, R. and H. Lütkepohl (2001), "Lag Selection in Subset Var Models with an Application to a U.S. Monetary System," in Friedman, R., L. Knüppel, H. Lütkepohl (eds.), *Econometric Studies: A Festschrift in Honour of Joachim Frohn*, Münster: LIT-Verlag, 107–128.

Caballero, R. and M. Hammour (1994), "The Cleansing Effects of Recession," *American Economic Review* 84, 1350–1368.

Caner, M. and B.E. Hansen (2001), "Threshold Autoregression with a Unit Root," *Econometrica* 69, 1555–1596.

Caplin, A.S. and J. Leahy (1991), "State-Contingent Pricing and the Dynamics of Money and Output, *Quarterly Journal of Economics* 106, 683–708.

Clark, T. E. and M.W. McCracken (2001), "Tests of Equal Forecast Accuracy and Encompassing for Nested Models," *Journal of Econometrics* 105, 85–110.

Clark, T.E. and K.D. West (2007), "Approximately Normal Tests for Equal Predictive Accuracy in Nested Models," *Journal of Econometrics* 138, 291-311.

Clements, M.P., P.H. Franses, J. Smith, and D. van Dijk (2003), "On SETAR Non-Linearity and Forecasting," *Journal of Forecasting* 22, 359–375.

Clements, M.P. and D.H. Hendry (1999), *Forecasting Non-stationary Economic Time Series*, Cambridge, MA: MIT Press.

Clements, M.P. and H.-M. Krolzig (2003), "Business Cycle Asymmetries: Characterization and Testing Based on Markov-Switching Autoregressions," *Journal of Business and Economic Statistics* 21, 196–211.

Cooley, T.F. and G.D. Hansen (1995), "Money and the Business Cycle," in Cooley, T.F., (ed.), *Frontiers of Business Cycle Research*, Princeton, NJ: Princeton University Press, 175–216.

Corradi, V. and N.R. Swanson (2007), "Nonparametric Bootstrap Procedures for Predictive Inference Based on Recursive Estimation Schemes," *International Economic Review* 48, 67-109.

Davies, R.B. (1987), "Hypothesis Testing When a Nuisance Parameter is Present Only Under the Alternative," *Biometrika* 74, 33–43.

Diebold, F.X. and R.S. Mariano (1995), "Comparing Predictive Accuracy," *Journal of*

Business and Economic Statistics 13, 253–263.

Giacomini, R. and White, H. (2006), "Tests of Conditional Predictive Ability," *Econometrica* 74, 1545-1578.

Harvey, D.I., S.J. Leybourne, and P. Newbold (1997), "Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting* 13, 281–291.

Hendry, D.F. and G.E. Mizon (1993), "Evaluating Dynamic Econometric Models by Encompassing the VAR," in Phillips, P.C.B. (ed.), *Models, Methods, and Applications: Essays in Honor of A.R. Bergstrom*, Cambridge, MA: Basil Blackwell, 272–300.

Krolzig, H.-M., M. Marcellino, and G.E. Mizon (2002), "A Markov-Switching Vector Equilibrium Correction Model of the UK Labour Market," *Empirical Economics* 27, 233–254.

Lundbergh, S. and T. Teräsvirta (1998), "Modelling Economic High-Frequency Time Series with STAR-GARCH Models," Working Paper Series in Economics and Finance No. 291, Stockholm School of Economics.

Luukkonen, R., P. Saikkonen and T. Teräsvirta (1988), "Testing Linearity Against Smooth Transition Autoregressive Models," *Biometrika* 75, 491–499.

Marcellino, M. (2006), "Instability and Non-Linearity in the EMU," in Milas, C., P. Rothman, and D. van Dijk (eds.), *Nonlinear Time Series Analysis of Business Cycles*, Amsterdam: Elsevier, 151–174.

Marcellino, M. (2004), "Forecast Pooling for Short Time Series of Macroeconomic Variables," *Oxford Bulletin of Economics and Statistics* 66, 91–112.

Marcellino, M., J.H. Stock, and M.W. Watson (2006), "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series ," *Journal of Econometrics* 135, 499–526.

McCracken, M. W. (2000), "Robust Out-of-Sample Inference," *Journal of Econometrics* 99, 195–223.

Montgomery, A.L., V. Zarnowitz, R.S. Tsay, and G.C. Tiao (1998), "Forecasting the U.S. Unemployment Rate," *Journal of the American Statistical Association* 93, 478–493.

Neftci, S.N. (1984), "Are Economic Time Series Asymmetric Over the Business Cycle?" *Journal of Political Economy* 92, 307–328.

Nickell, S. (1998), "Unemployment: Questions and Some Answers," *The Economic Journal* 108, 802–816.

Peel, D.A. and A.E.H. Speight (2000), "Threshold Nonlinearities in Unemployment Rates: Further Evidence for the UK and G3 Economies," *Applied Economics* 32, 705–715.

Proietti, T. (2003), "Forecasting the US Unemployment Rate," *Computational Statistics and Data Analysis* 42, 451–476.

Rothman, P. (1998), "Forecasting Asymmetric Unemployment Rates," *Review of Economics and Statistics* 80, 164–168.

Rothman, P., D. van Dijk, and P.H. Franses (2001), "A Multivariate STAR Analysis of the Relationship Between Money and Output," *Macroeconomic Dynamics* 5, 506–532.

Skalin, J. and T. Teräsvirta (2002), "Modeling Asymmetries and Moving Equilibria in Unemployment Rates," *Macroeconomic Dynamics* 6, 202–241.

Söderlind, P. and A. Vredin (1996), "Applied Cointegration Analysis in the Mirror of Macroeconomic Theory," *Journal of Applied Econometrics* 11, 363–381.

Stock, J.H. and M.W. Watson (2005), "Forecasting with Many Predictors," forthcoming in Elliott, G., W.J. Granger, and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Amsterdam: Elsevier.

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

Stock, J.H. and M.W. Watson (1999), "A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series," in Engle, R. and H. White (eds.), *Cointegration, Causality, and Forecasting: A Festschrift in Honor of Clive W.J. Granger*, Oxford: Oxford University Press, 1–44.

Teräsvirta, T. (1994), "Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models," *Journal of the American Statistical Association* 89, 208–218.

Teräsvirta, T., D. van Dijk, and M.C. Medeiros (2005), "Linear Models, Smooth Transition Autoregressions, and Neural Networks for Forecasting Macroeconomic Time Series: A Re-examination," *International Journal of Forecasting* 21, 755-774.

Terui, N. and H.K. van Dijk (2002), "Combined Forecasts from Linear and Nonlinear Time Series Models," *International Journal of Forecasting* 18, 421–438.

Tsay, R.S. (1998), "Testing and Modeling Multivariate Threshold Models," *Journal of the American Statistical Association* 93, 1188–1202.

van Dijk, D. and P.H. Franses (2003), "Selecting a Nonlinear Time Series Model Using Weighted Tests of Equal Forecast Accuracy," *Oxford Bulletin of Economics and Statistics* 65, 727-744.

West, K. D. (2001), "Tests for Forecast Encompassing When Forecasts Depend on Estimated Regression Parameters," *Journal of Business and Economic Statistics* 19, 29–33.

West, K. D. (1996), "Asymptotic Inference About Predictive Ability," *Econometrica* 64, 1067–1084.

West, K. D. and M.W. McCracken (1998), "Regression Based Tests of Predictive Ability," *International Economic Review* 39, 817–840.

White, H. (2000), "A Reality Check For Data Snooping," *Econometrica* 68, 1097-1127.

Wooldridge, J.M. (1991), "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variance, *Journal of Econometrics* 47, 5–46.