A comparison of tests of non-linear cointegration with an application to the predictability of US interest rates using the term structure

Michael P. Clements
Department of Economics
University of Warwick
M.P.Clements@warwick.ac.uk

Ana Beatriz Galvão*
Department of Economics
University of Warwick
ana.galvao@warwick.ac.uk

July 2001

Abstract

We evaluate the forecasting performance of a number of systems models of US short- and long-term interest rates. Non-linearities, including asymmetries in the adjustment to equilibrium, are shown to result in more accurate short horizon forecasts. We find that both long and short rates respond to disequilibria in the spread in certain circumstances, which would not be evident from linear representations or from single-equation analyses of the short-term interest rate.

JEL Classification: C52, C53, E43.

1 Introduction

In this paper we test whether there are non-linearities in the response of short and long-term interest rates to the spread in interest rates. The simple Expectations theory of the term structure entails that the spread has significant predictive content in a linear framework, although there are reasons to expect non-linearities in the responses. We test for non-linearities and asymmetric adjustment using a variety of models and approaches, and carry out a number of forecast comparisons of the out-of-sample predictability of short and long-term interest rates using linear and non-linear models.

Existing studies have focused on possible asymmetries in the response of the short rate to positive and negative spreads, within a single-equation framework. We take as our starting point a systems approach, and show how single-equation techniques for testing and specification can be adapted. A systems approach allows proper ex ante forecasts of short- and long-term rates, and therefore also of the spread. This may be important for a variety of reasons, not least that interest rates, and especially spreads, have been found to give advance warning of recessions in the US (see, inter alia, Estrella and Mishkin, 1998; Hamilton and Kim, 2000; Anderson and Vahid, 2000). We also consider the extent to which the implied patterns of adjustment are consistent with the expectations theory of the term structure of interest rates. In the

*Financial support from the UK Economic and Social Research Council under grant L138251009 is gratefully acknowledged by the first author, and from Capes-Brazil by the second author. We are grateful to Dick van Dijk for helpful comments. This paper is based on chapter 2 of the PhD thesis of the second author. The computations reported in this paper were performed using code written in the Gauss Programming Language.
remainder of this introduction we provide a brief motivation, and discuss some aspects of the recent literature that we address.

A number of authors have investigated the effects of 'linear' cointegration on forecast performance in linear systems (see Clements and Hendry, 1995 and Christoffersen and Diebold, 1998, and the references therein). The inclusion of cointegration especially improves forecasts of cointegrating combinations of the variables. However, the linearity assumption may be too restrictive when, for example, there are transactions costs, so that arbitrage opportunities between two markets only arise when the price differential is large enough to imply net gains to traders. In general, the speed of adjustment to equilibrium may depend on both the sign and size of the disequilibria. Moreover, some economic policies may affect the adjustment to equilibrium (Rudebusch, 1995; Roberds et al., 1996), suggesting a non-linear adjustment process.

Non-linear equilibrium correction models have been used to model the relationship between spot and future prices (Dwyer et al., 1996; Martens et al., 1998; Tsay, 1998), and between interest rates of different maturities (Anderson, 1997; Tsay, 1998; Enders and Granger, 1998; van Dijk and Franses, 2000).

A common finding of a number of recent papers that have considered the effects of non-linearities on economic forecasting in univariate models (e.g., Tiao and Tsay, 1994; Clements and Krolzig, 1998; Rothman, 1998; Montgomery et al., 1998; Stock and Watson, 1999; Lundbergh and Teräsvirta, 2000) is that a superior in-sample fit often fails to translate into out-of-sample forecast gains of a similar magnitude: see also Ramsey (1996). For example, after an evaluation of the forecast performance of univariate non-linear models applied to a large number of macroeconomic series, Stock and Watson (1999) concluded that in the majority of cases forecasts were worse than those of linear models.

Given that cointegration may improve forecasts in linear cointegrated systems, depending on how forecasts are evaluated, and that the improvements in forecasting from allowing for non-linearities in univariate models are on balance rather disappointing, the question we address in this paper is whether allowing non-linearities in the adjustment process in cointegrated systems is beneficial. We consider non-linearities in the term structure of interest rates, which has been a focus in the literature for much of the work on non-linear cointegration, and threshold vector equilibrium correction models (TVECMs). We evaluate the forecast performance of a number of non-linear systems of short and long-term US interest rates based on models proposed in the literature (see, e.g., Anderson, 1997; Tsay, 1998; van Dijk and Franses, 2000). Our models are specified and tested using recently developed techniques, as described in section 3.

The three main contributions of our paper are: to provide a review of the econometric tests of non-linear cointegration, and extensions to systems frameworks; to apply these tests to the US term structure of interest rates; and to assess any gains in out-of-sample predictability from allowing for non-linearities.

The plan of the remainder of the paper is as follows. A brief review of equilibrium correction models of the term structure is presented in section 2. In section 3 we review the specification and testing procedures for threshold equilibrium correction models advanced in the recent literature. Ways of assessing out-of-sample predictability – the forecast accuracy and encompassing tests – are described in section 4. The results of testing for non-linearities are discussed in section 5, and the results of the forecast evaluation exercise are presented in section 6. Section 7 discusses economic implications of our results. Section 8 concludes.
2 Term structure of interest rates and equilibrium correction models

The Expectations Theory implies that at period \( t \), the return on a \( k \)-period asset is the average of the current and a sequence of expected one period yields over the same period, \( t \) to \( t + k \), plus a term premium:

\[
rt(k) = \frac{1}{k} \sum_{j=1}^{k} E_t r_{t+j-1}(1) + L_t(k).
\]  

(1)

\( rt(k) \) is the yield at maturity \( k \); \( E_t \) is the expected value at time \( t \); \( L_t \) is the term premium. The term premium consists of risk and liquidity premia. Assuming that yields are integrated of order one (\( I(1) \)), the possibility of the yield spread being cointegrated follows from rewriting (1) as:

\[
rt(k) - rt(1) = \frac{1}{k} \sum_{i=1}^{k-1} \sum_{j=1}^{i} E_t \Delta r_{t+j}(1) + L_t(k).
\]  

(2)

In the absence of \( L_t(k) \), the RHS of (2) will be \( I(0) \), consisting of a finite sum of \( I(0) \) components (the \( \Delta r_{t+j}(1) \)). Then each yield \( rt(k) \) is cointegrated with \( rt(1) \), and the spreads are stationary, so that spreads between any two yields will be \( I(0) \) (Hall et al., 1992). It follows that a Vector Equilibrium Correction Model (VECM) for any two yields, say, a short-term \( (s) \) and long-term \( (l) \) yield, can be written as:

\[
\Delta r_t = c(L) \Delta r_{t-1} - \alpha(S_{t-1} - \mu) + \epsilon_t
\]  

(3)

where \( r_t = [r_t(s), r_t(l)]' \); \( c(L) \) is a matrix of coefficients in the lag operator; \( S_t = r_t(l) - r_t(s) \) is the spread; \( \mu \) is the equilibrium spread; \( \alpha \) is the adjustment vector to the long-run attractor; and \( \epsilon_t \) is the vector of disturbances. \( L \) is the lag operator, \( L^n x_t = x_{t-n} \), and \( \Delta = 1 - L \) is the difference operator. Yields of different maturities can move apart in the short-run, but in the long-run are ‘tied together’. \( \mu \) may be different from zero because of a risk premium.

The empirical evidence of cointegration between yields of different maturities, and the ability of the ECM to improve forecasts of interest rates, is contested in the literature (Pagan et al., 1996). The usefulness of the spread to forecast short-term rates is found to depend on the maturities (Rudebusch, 1995) and on the monetary policy in operation (Mankiw and Miron, 1986; Rudebusch, 1995; Roberds et al., 1996; Gray, 1996). Tzavalis and Wickens (1998) note that the presence of time-varying risk premia in (3) may mask the predictive power of the spread, and in addition, the dynamics of \( \Delta r(s) \) may depend on its level, as in the regime-switching models considered by Pfann et al. (1996). Moreover, the Expectations Theory ignores transaction cost effects. These considerations – the diversity of monetary policies, the risk premium and the presence of transaction costs – can be accommodated in a non-linear equilibrium correction model, where the speed of adjustment to equilibrium depends on the regime.

Allowing for asymmetric adjustment to positive and negative values of past realizations of \( S_t \), or to changes in \( S_t \), motivates the Threshold Autoregressive (TAR) and Momentum TAR (MTAR) models of Enders and Granger (1998). Anderson (1997) models the effects of transaction costs by supposing that inside a band around \( \mu \), \( \{c_1 = \mu - \tau_1, c_2 = \mu + \tau_1 \} \), there
is no adjustment, where asymmetry arises if \( \tau_l \neq \tau_u \). Outside this band, adjustment may occur at different speeds. Supposing that individual investors have different transaction costs implies that in the aggregate the effect of transaction costs might be smooth, leading to the Smooth Transition Equilibrium Correction model (STEqCM): see Anderson (1997) and van Dijk and Franses (2000). Both these applications are single-equation analyses, although van Dijk (1999, ch. 5, p. 128) discusses bivariate STVECMs (Smooth Transition Vector Equilibrium Correction models).

3 Threshold vector equilibrium correction models

In this section we discuss aspects of testing, model specification and estimation procedures for non-linear equilibrium correction models.

3.1 Non-linearity testing

Threshold vector equilibrium correction models are extensions of univariate Threshold Autoregressive (TAR) models (see, e.g., Tong, 1995), and inherit the non-standard aspects of testing for non-linearity that arise from the presence of nuisance parameters under the null when likelihood-based approaches are used (see, e.g., Hansen, 1996).

The benchmark model is a linear VECM of the form of (3) (the change in notation is to aid comparison with the non-linear models that follow):

\[
\Delta r_t = c + \sum_{j=1}^{p} \Phi_{1,j} \Delta r_{t-j} + \alpha_s t_{-1} + \varepsilon_t,
\]

where the autoregressive order \( p \) is set to minimize an information criteria. A possible non-linear alternative hypothesis is that interest rates follow a threshold VECM (TVECM), such that:

\[
\Delta r_t = c_i + \sum_{j=1}^{p} \Phi_{i,j} \Delta r_{t-j} + \alpha_i s_{t-1} + \varepsilon_t
\]

if \( r_{i-1} < z_{t-d} \leq r_i \), where \( i = 1, 2 \) in the case of a two-regime model (with \( r_0 = -\infty, r_2 = +\infty \)) and \( i = 1, 2, 3 \) for a three-regime model (with \( r_0 = -\infty, r_3 = +\infty \)), and \( \{c_i, \Phi_{i,j}, \alpha_i\} \) depend on the regime \( i \). \( z_{t-d} \) is the transition variable with delay \( d \); \( r_i \) are the thresholds; and \( \alpha_i \) is \((2 \times 1)\), allowing the spread to enter as an explanatory variable in both equations. The \( i \) subscript on \( \varepsilon_t \) indicates that the covariance matrix of the disturbances may depend upon the regime.

3.1.1 Tsay (1998)

The linearity test proposed by Tsay (1998) is the vector extension of the Tsay (1989) test for non-linearity based on an 'arranged regression'. The problem of testing for a threshold becomes that of testing for a change-point. Unlike likelihood-based approaches, which rely on simulated critical values, the test has an asymptotic chi-squared distribution. The arranged regression orders the observations according to the size of \( z_{t-d} \), and assumes that the threshold variable \( (z_t) \) and the delay, \( d \), as well as the autoregressive order, \( p \), are known. The model is then estimated by recursive LS, and the predictive residuals are obtained. Under the null that the model is linear, these residuals are uncorrelated with the explanatory variables in the arranged
regression. The test is constructed by regressing the (standardized) predictive residuals on the explanatory variables, and testing for the significance of the latter: see Tsay (1998, p. 1189–1191), for details. The simulation results presented by Tsay (1998) indicate that this test has good power when \( d \) is correctly specified.

### 3.1.2 Balke and Fomby (1997)

Balke and Fomby (1997) proposed a two step procedure to test non-linearity in cointegrated systems. The first step is to test for cointegration using a standard method, such as OLS and an ADF test, as in Engle and Granger (1987), or the Johansen ML procedure, Johansen (1988). The second step tests for non-linearity in the cointegrating combination, using a test of linearity against SETAR structure, such as Hansen (1996, 2000). An asymptotic approximation to the \( p \)-value of the null of linearity (or of a 2-regime versus a 3-regime structure) can be obtained by simulation, or a small-sample approximation can be calculated by bootstrapping.

### 3.1.3 Enders and Granger (1998) and Enders and Siklos (2001)

Testing for cointegration (as in Step 1 of Balke and Fomby (1997) above) may have low power when the process is \( I(0) \) but exhibits non-linear mean reversion: see, e.g., the Monte Carlo evidence in Balke and Fomby (1997). Enders and Granger (1998) proposed a test for a unit root in a series against asymmetric adjustment under the alternative hypothesis, whereby the process is either a two-regime TAR or an M-TAR (Momentum-TAR). For the former, the rate of autoregressive decay depends on whether the variable is above or below some threshold, and for the latter, on whether the variable is increasing or decreasing. For example, when the threshold is zero, the following regression is run:

\[
\Delta S_t = I_t \rho_1 S_{t-1} + (1 - I_t) \rho_2 S_{t-1} + \epsilon_t
\]  

(6)

where, for the TAR alternative, \( I_t = 1 \) if \( S_{t-1} \geq 0 \), and is zero otherwise, and for testing against the M-TAR alternative, \( I_t = 1 \) if \( \Delta S_{t-1} \geq 0 \), and is zero otherwise. Enders and Granger obtain by simulation critical values for the unit root null that \( \rho_1 = \rho_2 = 0 \) against both these alternatives, and for various modifications of the above set up (e.g., when the attractor is not zero, but has to be estimated, and when lagged \( \Delta S_t \) terms are added in the event that the estimated error, \( \hat{\epsilon}_t \), exhibits autocorrelation). Conditional on rejecting the null and finding \( \rho_1 < 0 \) and \( \rho_2 < 0 \), tests that \( \rho_1 = \rho_2 \) have standard distributions. Enders and Siklos (2001) generalise these ideas to tests for cointegration, that is, when the variable is an estimated residual from an Engle-Granger static regression of one integrated variable on another (or several). In the case of the term structure, the residual is \( \bar{w}_t = r_t(t) - \hat{\theta} r_t(s) - \hat{\mu} \), where \( \hat{\theta}, \hat{\mu} \) are OLS estimates. The null hypothesis is now interpreted as a test that \( r(t) \) and \( r(s) \) are not cointegrated, and the alternative is of ‘asymmetric cointegration’, whereby the series can be shown to be related by a non-linear equilibrium-correction model. They obtain by simulation critical values of two test statistics of the null that \( \rho_1 = \rho_2 = 0 \) (against both TAR and MTAR alternatives). These are an \( F \)-test of \( \rho_1 = \rho_2 = 0 \) and a ‘\( t \text{-max} \)’ statistic, which is the larger of the two individual \( t \)-tests of \( \rho_1 = 0 \) and \( \rho_2 = 0 \). A Monte Carlo evaluation of these tests suggests that the \( F \)-test is to be preferred, and has reasonable power when the process is an M-TAR, but otherwise is dominated by the ADF test! We report both tests. Further complications arise when the threshold is not known (we implicitly assume that it is
zero for \( w_t \). An incorrect assumption concerning the threshold would reduce the power of the test, but we do not pursue this as it turns out that we are able to reject the null against both alternatives.

3.1.4 A generalization of Hansen (2000)

Instead of testing \( S_t \) (or an estimated cointegrating relationship between \( r \) (l) and \( r \) (s)) for non-linearity as outlined above, threshold effects can be tested for by comparing the linear system (4) against the non-linear alternative (5). This requires a multivariate extension of Hansen (2000). Imposing the restriction that the error covariance matrices are the same for the different regimes, the two- and three-regime TVECMs can be written as:

\[
\Delta r_t = \left[ c_1 + \sum_{j=1}^{p} \Phi_{1,j} \Delta r_{t-j} + \alpha_1 S_{t-1} \right] I_{1t}(r) + \left[ c_2 + \sum_{j=1}^{p} \Phi_{2,j} \Delta r_{t-j} + \alpha_2 S_{t-1} \right] I_{2t}(r) + u_{2t}
\]

\[
\Delta r_t = \left[ c_1 + \sum_{j=1}^{p} \Phi_{1,j} \Delta r_{t-j} + \alpha_1 S_{t-1} \right] G_{1t}(r_1, r_2) + \left[ c_2 + \sum_{j=1}^{p} \Phi_{2,j} \Delta r_{t-j} + \alpha_2 S_{t-1} \right] G_{2t}(r_1, r_2)
\]

\[
\Delta r_t = \left[ c_3 + \sum_{j=1}^{p} \Phi_{3,j} \Delta r_{t-j} + \alpha_3 S_{t-1} \right] G_{3t}(r_1, r_2) + u_{3t}
\]

where \( I_{1t}(r) = I(S_{t-1} \leq r) \), \( I_{2t}(r) = I(S_{t-1} > r) = 1 - I_{1t}(r) \), \( G_{1t}(r_1, r_2) = I(S_{t-1} \leq r_1) \), \( G_{2t}(r_1, r_2) = I(r_1 < S_{t-1} \leq r_2) \) and \( G_{3t}(r_1, r_2) = I(S_{t-1} > r_2) \), and \( I(\cdot) \) is the indicator function. We denote the estimated covariance matrices of \( u_{2t} \) and \( u_{3t} \) by \( \hat{\Omega}_2 \) and \( \hat{\Omega}_3 \), and let \( \hat{\Omega}_1 \) be the covariance matrix of the VECM\(^1\). An LR test for linearity against the two-regime specification is:

\[
LR_{12} = T \left[ \ln(\det(\hat{\Omega}_1)) - \ln(\det(\hat{\Omega}_2)) \right]
\]

where \( T \) is the number of observations effectively employed in the estimation.

The asymptotic distribution is an extension of Hansen (1996), as argued by Hansen and Seo (2000). The bootstrap can be used to obtain a finite sample approximation. The bootstrap distribution is calculated from data generated by the linear model by re-sampling its residuals. The residuals are corrected for heteroscedasticity before re-sampling, using a regression of the squared residuals on the squared regressors, as described in Hansen (2000).

A similar procedure can be applied to test linearity against the three-regime alternative:

\[
LR_{13} = T \left[ \ln(\det(\hat{\Omega}_1)) - \ln(\det(\hat{\Omega}_3)) \right]
\]

or the two-regime versus the three-regime model.

3.1.5 Hansen and Seo (2000)

Rather than firstly estimating the cointegrating relationship, Hansen and Seo (2000) suggest a single step approach that jointly estimates the cointegrating vector and the defining character-

\(^1\)The models are estimated on the same sample and have the same autoregressive order \( p \).
istics of the regimes. The non-linear model is:

\[
\Delta r_t = \left[ c_1 + \sum_{j=1}^{p} \Phi_{1,j} \Delta r_{t-j} + \alpha_1 w_{t-1}(\theta) \right] d_{1t}(\theta, r) + \left[ c_2 + \sum_{j=1}^{p} \Phi_{2,j} \Delta r_{t-j} + \alpha_2 w_{t-1}(\theta) \right] d_{2t}(\theta, r) + u_t
\]

(9)

where \( w_t(\theta) = r_t(l) - \theta r_t(s) \) is the equilibrium correction term, \( d_{1t}(\theta, r) = I(w_{t-1}(\theta) \leq r) \) and \( d_{2t}(\theta, r) = I(w_{t-1}(\theta) > r) \). Their method involves estimating (9) at each point in a suitable grid of values defined over both \( \theta \) and \( r \), and choosing the pair \((\hat{\theta}, \hat{r})\) that minimizes \( \log |\Omega(\theta, r)| \) (where \( \Omega \) is the estimated residual covariance matrix). In practice, because of the limitations of the estimation by grid search, the delay is given, and the TVECM is restricted to having two regimes, as above. The authors also proposed an 'LM-like' non-linearity test. LM tests are calculated of the linear model against (9) with \( \theta = \hat{\theta} \), the MLE in the linear VECM, and where \( r \) takes on each of a pre-assigned set of values in the interval \((r_L, r_U)\) (such that a minimum number of observations occur in each regime, say 10%). The test statistic is the supremum:

\[
\text{SupLM} = \sup_{r_L \leq r \leq r_U} \text{LM}(\hat{\theta}, r).
\]

(10)

Hansen and Seo (2000) derive the asymptotic distribution of this statistic, and propose a bootstrap procedure to obtain a finite sample approximation.

### 3.2 Estimation

Tsay (1998) considers estimation of the TVECM by conditional multivariate least squares assuming the number of regimes, the autoregressive order \( p \), and the threshold variable \( z_t \) are known. Equation (5) is estimated for all permissible combinations of the delay \( d \) and the thresholds \( r_1 \) and \( r_2 \) (for the 3-regime model), subject to a minimum number of observations in each regime, \( r_2 > r_1 \), and \( d \) being a low integer. The estimates of the thresholds and \( d \) are those values for which the sum of squares is minimized. The asymptotic results and properties of the estimators are discussed by Tsay (1998). Instead of minimizing the sum of the squares of the errors, Tsay (1998) suggests employing the AIC. When the autoregressive order \( p \) and the number of regimes \( s \) are fixed, minimizing AIC - unlike minimizing the sum of squared errors - is asymptotically equivalent to minimizing the generalised residual variance (see (11) below).

To reduce the computational burden of estimating thresholds in a three-regime model (especially when bootstrapping is employed to calculate the finite sample distribution of the non-linearity test), Hansen (2000) proposes a one-step-at-a-time algorithm, that is, a sequential procedure. The threshold value estimated for a two-regime model is employed as one of the thresholds of the three-regime model, and a grid search for the second threshold is then conducted, with the same delay as in the two-regime model. The objective function of the grid search is \( \hat{\Omega}(r_1, r_2) \), where \( \Omega \) is the estimated variance-covariance matrix of the residuals, given the threshold values \( r_1 \) and \( r_2 \) and assuming constant variance across regimes. This procedure is iterated at least once, to refine the estimation of the threshold values. Bai (1997) proved the consistency of this sequential approach for models of multiple structural breaks. In addition, the cointegration vector and the threshold can be jointly estimated (Hansen and Seo (2000) and section 3.1.5).
3.3 Model specification

The choice of the threshold variable will usually be suggested by subject-matter considerations. The number of regimes of the TVECM may be selected based on theoretical considerations (as in e.g., Kräger and Kugler, 1993 and Anderson, 1997), but as discussed in section 2, in the context of the term structure theoretical arguments can be put forward to support both two and three-regime models (respectively, adjustment depends upon whether the yield curve is normal or inverted, and transaction costs). Tests can be employed to determine whether there are two or three regimes, as for example the Hansen (2000) test for a two-regime versus a three-regime SETAR. We also use a systems generalization: the \( LR_{23} \) test that compares two-with three-regime TVECMs. In addition, Tsay (1998) suggests that the number of regimes \( s \) may be selected along with the lag order \( p \) by using the AIC:

\[
AIC = \sum_{j=1}^{s} \left[ T_j \ln \left( |\hat{\Omega}_j| \right) + 2k(kp + 1) \right],
\]

where \( k = 2 \) in the bivariate system, \( \hat{\Omega}_j \) is the estimated residual covariance matrix of regime \( j \), and \( T_j \) is the number of observations in regime \( j \). Thus the AIC is calculated for each combination of the threshold values, the delay and for \( s = 2, 3 \), given that \( p = 2 \) for the models estimated in this work. Specific tests may also indicate the type of non-linear model required. For example, the Enders and Granger (1998) and Enders and Siklos (2001) tests may suggest either TAR or MTAR.

4 The evaluation of forecast performance

The models are compared on out-of-sample forecast accuracy as judged by their mean squared forecast errors at various horizons. We test whether differences between models are significantly different, and also calculate forecast encompassing tests.

Clements and Hendry (1995) and Christoffersen and Diebold (1998) find that there is a gain in forecast accuracy at longer horizons when the evaluation includes the ability to predict the cointegrating combination (here, the spread). Clements and Hendry (1993) show why forecast evaluation using the standard (root) mean squared forecast error criterion, \((R)MSFE\), may depend on the transformation of the variables adopted (e.g., the levels of the original variables, their first differences or growth rates, or a mixture of first differences and stationary combinations). Of interest here are forecasts of both interest rates and the spread, so we consider these separately for the most part. We look at forecasts of the differences of the interest rates as these should be stationary. The MSFE for the variable \( x \) at horizon \( h \) is denoted by \( MSFE_{x,h} \), and is calculated by averaging the squares of the \( h \)-step ahead forecast errors over a number of forecast origins, \( T \). The MSFEs could be summed for the two rates, for example, to give the trace MSFE, TMSFE, as an estimate of the trace of the MSFE matrix. Rather than doing this, we report an overall measure of systems performance referred to as the GFESM by Clements and Hendry (1993), because this is invariant to whether we evaluate the models in terms of their ability to forecast the changes of the rates, or the short-term rate and the spread, etc.

We test whether the MSFEs of the various models are significantly different from each other using the Diebold and Mariano test of equal forecast accuracy, with the small-sample corrections suggested by Harvey et al. (1997). That is, to test the null of equal accuracy at \( h \) steps-ahead we calculate:
\[ d_t = e_{i,t}^2 - e_{j,t}^2 \]

\[ DM = [\hat{V}(\bar{d})]^{-1/2} \bar{d} \]

\[ ADM = \left[ \frac{n + 1 - 2h + n^{-1}h(h - 1)}{n} \right]^{1/2} DM \]  \hspace{1cm} (12)

where \( t = 1, \ldots, n \) indexes the \( n \) \( h \)-step ahead forecasts that are available, and \( i \) and \( j \) index the two models. So \( e_{i,t} \) is the error in forecasting the value at \( t \), made at \( t - h \), using model \( i \). \( \bar{d} \) is the sample mean of the loss differential \( d_t \), where loss is symmetric, defined in terms of the squares of the errors. The estimated variance of the sample mean is denoted by \( \hat{V}(\bar{d}) \), and depends on the sample autocovariances of \( d_t \). Under the null of equal forecast accuracy, \( DM \) is asymptotically standard normal. Thus values of the statistic in the left tail suggest model \( i \) is more accurate, and values in the right tail that \( j \) is more accurate. Harvey et al. (1997) suggest comparing the Augmented Diebold and Mariano (ADM) test to the \( t \)-distribution to reduce size distortions, and Clark (1999) confirms that these modifications improve the small-sample performance of the test.

We can test whether model \( i \) forecast encompasses model \( j \), that is, whether once we have model \( i \), there is any useful additional information contained in the forecasts of model \( j \), by modifying \( d_t \) to:

\[ d_t = (e_{i,t} - e_{j,t}) e_{i,t} \]  \hspace{1cm} (13)

Harvey et al., 1998 show that this test is equivalent to the forecast encompassing test of Chong and Hendry (1986); see Newbold and Harvey (2001) for a review. The condition that a model forecast encompasses another is more stringent than that the model is more accurate on the DM test. One model may be more accurate than another on DM, but nevertheless, the dominated model contain useful information not incorporated in the superior model.

West (1996) and West and McCracken (1998) draw attention to the impact of parameter uncertainty on the size of the tests of equal forecast accuracy and encompassing, which is enhanced when the models are nested (Clark and McCracken, 2000). By ascribing a reasonable proportion of the observations to the in-sample, model estimation period, we hope it will be of secondary importance here.

### 5 Model estimates

The various TVECMs we employ are described in Table 1. They differ with regard to the testing, estimation or specification procedures used. We analyse the three month treasury bill at secondary market rate \( (r(3)) \) and the 10 year treasury constant maturity rate \( (r(120)) \), both monthly for the period 1960:1 to 2000:4. The data were obtained from the Fred website (www.stls.frb.org/fred/data/irates.html). The data are plotted in Figure 1, along with the spread, \( r(120) - r(3) \), and the NBER nominated recessionary periods. The analysis in this section presents results for the initial estimation period (1960:1 to 1989:12) and for the final estimation period (1960:1 to 1998:4).
5.1 Testing and modelling

Applying the procedures for specification of TVECMs outlined in section 3, we estimated five TVECMs (3 with 2 regimes, 2 with 3 regimes), as described in Table 1. For all models, \( p = 2 \), which minimizes the SIC for the linear VECM (the first model in table 1). For the non-linear models, we set \( d = 1 \), following, e.g., Anderson (1997) and Hansen and Seo (2000), but this assumption is confirmed by the Tsay test presented in panel 11 of Table 2. The threshold(s) are calculated by a grid search, ensuring a minimum of 10\% of the observations are in each regime.

For all but two of the models the spread \( (S_t = r(l) - r(s)) \) is imposed as the equilibrium correction term, and is used as the threshold variable. Table 2 panel 1 indicates that \( r(l) \) and \( r(s) \) are integrated of order 1 on the basis of ADF tests. There is some ambiguity over the spread: on the basis of an ADF test it is \( (1) \), the Phillips-Perron tests indicates that it is \( (0) \) (panel 2), and the Johansen systems-based trace test for cointegration (panel 3) finds that \( r(l) \) and \( r(s) \) are cointegrated, but the restriction that the interest rates have equal and opposite sign (defining the spread) is rejected (not recorded). Following previous work, and taking into account the low power of unit root tests, we use the spread in the linear model and three of the non-linear models. In the other two non-linear specifications the long-run relationship is estimated as described below.

The second model in Table 1, labelled 2R-TVECM is the first 2-regime TVECM. The threshold is chosen to minimize the AIC, \( T_j \ln |\Omega_j| \) with \( p \) set to 2). The third model, MTVECM has as the long-run \( w_t = r(l) - \theta r(s) \), where \( \theta \) is estimated as in Engle and Granger (1987). The tests of Enders and Granger (1998) and Enders and Siklos (2001) for asymmetric adjustment (Table 2, panels 4 and 5) reject no cointegration in favour of both the M-TAR and TAR alternatives. The F-test for \( \rho_1 = \rho_2 \) does reject the null when an M-TAR is specified. However, this is not a strong evidence against M-TAR asymmetries because the F-test assumes that the threshold is equal to zero, while the estimated value 0.39. We choose the M-TAR specification as the basis for the MTVECM to be in line with Enders and Granger (1998) and Enders and Siklos (2001) on similar data sets. 2R-TVECM\( \text{joint} \) also estimates \( \theta \) and here a preliminary estimate of \( \theta \) is used to define a grid of values, and \( \theta \) is then estimated jointly with the threshold value to minimize \( \ln |\Omega| \). The 2R-TVECM\( \text{joint} \) is tested against the linear specification using the Hansen and Seo (2000) SupLM test of the null of linearity against a 2-regime model. We set \( p = 1 \) and \( d = 1 \). The \( p \)-values are calculated by bootstrapping, employing a procedure that corrects for possible heteroscedasticity under the null. Linearity is rejected in favour of a 2-regime model, but the 2-regime SETAR is not rejected in favour of the 3-regime model. Nevertheless, we estimate a 3-regime model, noting that this is selected by AIC and SIC.

Finally, 3R-TVECM is also a 3-regime model, and the thresholds are estimated ‘one-step-at-a-time’, but within the bivariate system. We report the results of a test that extends Hansen...
The LR test p-values presented in Table 2, panels 9 and 10 are based on the bootstrap, employing a correction for heteroscedasticity for the model under the null. Linearity is rejected in favour of 2 regimes, as is 2 regimes in favour of 3, contradicting the findings of the univariate testing of the spread (cf. panel 8).

The 2R-TVECM has regime-dependent error variances because the thresholds are calculated to minimize a criterion such as in \( \sum_{j=1}^{s} T_j \ln |\Omega_j| \) (where \( s \) is the number of regimes) which sums the (log) determinants of the regime-specific error second-moment matrices. For model 3R-TVECM\(_u\), the thresholds are chosen to minimize the sum of squares of a 3-regime SETAR model of the spread. Three sets of regressions are then run on the data belonging to each regime, resulting in regime-specific estimates of the error covariance matrices. The remaining models assume constant error variances across regimes, because the optimization criterion is of the form \( \ln |\Omega| \), where \( \Omega \) is the covariance matrix of the full-sample residuals. In principle, the assumption concerning the variances of the models' errors may also affect the forecasts (and not just via the effect on the parameter estimates), because these are calculated by bootstrapping: for the regime-dependent error variance models we sample with replacement only from the errors appropriate to the regime the model is in at that time; for the constant-variance models we sample from all the past errors. In practice, some experimentation suggested only small differences between these two forms of bootstrapping, and, moreover, that the parameter estimates were not unduly sensitive to whether a criterion such as \( \sum_{j=1}^{s} T_j \ln |\Omega_j| \) or \( \ln |\Omega| \) was minimized over the \( r_i \).

5.2 Analysis of estimation results

The results of estimating the models for both 1960:1-1989:12 (the initial estimation period) and 1960:1-1998:4 (the sample including the forecast period) are presented in Table 3. The short-run dynamics are summarised by the long-run dynamic growth multipliers\(^3\) to save space.

A number of interesting points arise. The estimates of the lower regime threshold are virtually the same for all the models and both periods, at around zero, so the lower regime is typically characterized by \( r(s) > r(l) \). The exceptions are the MTVECM and the 2R-TVECM\(_{joint}\). The threshold variable for the former is the change in the spread, and the threshold is estimated as a 0.36 point increase. For the latter, the threshold value is difficult to interpret, because the threshold variable, at \( r(l) - 1.54r(s) \) and \( r(l) - 1.4r(s) \) for the two samples, is quite different from the spread (with \( \theta = 1 \)). For the 3-regime models the value of the threshold between the middle and upper regimes is around 2\( \frac{\epsilon}{2} \).

The adjustments to equilibrium in the VECM are small in magnitude for both equations, but statistically significant for the short-term rate in both periods. However, allowing for a threshold effect, the coefficient on the spread is much larger at around 0.6 (exempting models MTVECM and 2R-TVECM\(_{joint}\)) for the \( \Delta r(s) \) equation in the lower regime (\( S_l < 0.6 \)). Thus, ceteris paribus, this results in \( \Delta r_{t+1}(s) < 0 \), so that the spread increases. Thus the short-term

\(^3\)The long-run multiplier of the effect of \( \Delta r(s) \) on \( \Delta r(l) \) is regime specific:

\[
\left( \Phi_{1,1,(1,2)} + \Phi_{1,2,(1,2)} \right) / \left( 1 - \Phi_{1,1,(1,1)} - \Phi_{1,2,(1,1)} \right)
\]

where in \( \Phi_{i,(j,l)} \) the subscript \( j \) refers to the regime, \( j \) to the lag, and \( (s,l) \) to the \( s,l^{th} \) element of that coefficient matrix. The multiplier of \( \Delta r(l) \) on \( \Delta r(s) \) is similarly defined.
interest rate falls when the spread is negative, while there is little evidence that \( r(l) \) responds.

In the upper regime of the 3-regime models, \( r(l) \) tends to respond to reduce the discrepancy, and in the case of 3R-TVECM there is a numerically large but not statistically significant coefficient on the spread in the short-term rate equation, suggesting short-term rates will increase. In both 3-regime models the coefficient on the spread is close to zero and statistically insignificant in the middle regime (approximately \( 0 < S < 2/3 \)).

The three-regime TVECMs generally fit better on AIC and SIC than the two-regime models, even discounting the penalties for the inclusion of more parameters (see last column of Table 3).

6 Forecast evaluation

Our smallest estimation period contains 360 monthly observations (1960:1 to 1989:12). For that period, we then generate 1 to 24-step ahead forecasts using a bootstrap\(^4\). In this way, we do not need to make any assumption about the error distribution, only that the errors are independent. The estimation period is increased by one, the models are re-estimated, and forecasts are again generated, and so on until we have 100 forecast origins, and a maximum estimation period of 460 observations (1960:1 to 1998:4). As the coefficients are re-estimated each period, \( d \) and \( p \) are held constant, but the thresholds are re-estimated.

Figure 2 presents the ratios of the MSFEs of the non-linear models to the VECM, for each of the three variables, \( \Delta r(l) \), \( \Delta r(s) \) and \( S \), for forecast horizons of \( h = 1 \) to 24. The VECM MSFE forms the numerator of the ratio, so values in excess of 1 indicate an improved performance relative to the VECM. It is apparent that allowing for non-linearity contributes relatively little to the ability to forecast \( \Delta r(l) \), with gains generally less than 5%. The ADM test results in Table 4 indicate that there are no significant differences in the forecast accuracy of the models at one-step ahead. At a horizon of four-steps the 2R-TVECM is significantly more accurate than the VECM and other non-linear models.

The improvement in accuracy of forecasts of \( \Delta r(s) \) from allowing for non-linearity is much more marked. Models 2R-TVECM, 2R-TVECM\(_{\text{joint}}\) and 3R-TVECM have MSFEs around 50% lower at a horizon of one, and 3R-TVECM is 10% more accurate at four-steps ahead. At one-step ahead, all non-linear models forecast better than the VECM, except for the MTVECM. At four-steps ahead, the 2R-TVECM and the 3R-TVECM forecast significantly better than the other models. In contrast, at \( h = 8 \), non-linear models produce forecasts equivalent to the linear one. These results are in tune with the Pfann et al. (1996) finding of non-linearity in the short-term interest rate, and also the single-equation equilibrium-correction models for the short-term rate (Anderson, 1997 and van Dijk and Franses, 2000).

The MSFE ratios for the spread indicate smaller gains than for \( \Delta r(s) \), but gains that persist as the horizon lengthens. The gains to non-linearity on MSFE are of 15% to 35% at one-step, and continue (albeit at a lower level) for 2R-TVECM and 3R-TVECM. At one-step ahead the 2R-TVECM, the 2R-TVECM\(_{\text{joint}}\) and the 3R-TVECM are statistically more accurate than the other models. Even at eight-steps ahead, the rankings on forecast encompassing tests favour the 2R-TVECM and the 3R-TVECM (Table 4). At horizons in excess of 18 months the 2R-TVECM\(_{\text{joint}}\) records large gains in accuracy. The reasons are probably due to the ‘free’ estimation of \( \theta \) (constrained to be 1 in most models) but why this should matter is unclear.

\(^4\)See e.g., Granger and Teräsvirta (1993) and Clements and Smith (1997).
The 2R-TVECM forecasts well despite the fact that negative spreads do not occur over the forecast period, and that it is during these times that it differs from the VECM model and incorporates a significant 'levels effect'. But notice that the VECM and 2R-TVECM suggest quite different dynamic responses (as summarized in the dynamic multipliers reported in table 3). The failure of the VECM to appropriately model behaviour when the spread is negative affects the dynamic responses of the VECM, and its performance is inferior even at more commonly observed values of the spread.

Accuracy as assessed by the GFESM is recorded in Figure 3. 2R-TVECMjoint has the best performance at all horizons, followed by models 2R-TVECM, 3R-TVECM and 3R-TVECMu, confirming the superiority of non-linear models.

6.1 Forecast evaluation conditional upon a regime

The recent literature suggests that non-linear models often record increased gains over linear comparators in some states of nature, but not others (see, for example, Tiao and Tsay, 1994; Clements and Smith, 1997; Clements and Smith, 1999). Comparing models on MSFE over the whole forecast period, as we have done, is likely to under-estimate the gains that accrue conditional on being in a specific regime.

The form of the estimated non-linear models suggests that responses to the equilibrium correction term differ markedly to those implied by the VECM at low (i.e., negative) and high (in excess of 2.7) values of the spread, so that we would like to consider forecasts of $\Delta r(s)$ and $S$ when the value of the spread at the forecast origin is negative, and forecasts of $\Delta r(l)$ when the value of the spread exceeds 2.7. For the majority of the data points the spread is between these two extremes, so the VECM, whose parameter estimates are effectively an average of the regime-specific values, is characterized by very modest mean reversion, and its forecasts are not too dissimilar to those of the non-linear models. However, the spread is never negative over the forecast period. Alternatively, we could consider only those forecasts made when the models' lower regimes were operative, but even this event seldom happened over the forecast periods. Instead, we report results of a forecast evaluation exercise conditioned on $S_{t-1} < 0$ on the simulated data described in section 6.2.

For large values of the spread, such as those considered in this evaluation exercise, the 3-regime TVECMs suggest the spread exerts significant downward pressure on $r(l)$. Comparing the relative MSFEs for $\Delta r(l)$ in Figures 2 and 4, an improved forecast performance for these models is apparent.

6.2 A simulation exercise

One of the problems of the empirical forecast evaluation exercise is that, as can be observed in Figure 1, the value of the spread is never negative in the out-of-sample period. But we can use Monte Carlo to ensure that the features that occurred historically (during the estimation period) repeat during the forecast period. A major disadvantage of this approach is that it requires that the model being used to generate data is a good representation of the process, otherwise the relevance of the results for 'real world' comparisons is brought into question (but the results will still be informative about the relative forecast performances of the models). As an example, McConnell and Perez-Quiros (2000) argue for a structural break in the volatility of US output growth around 1984. This may explain the long expansion of the 1990s, and would question the comparison of models' forecast performance on simulated data characterized by
recurrent recessions. Similarly, the absence of inverted yield curves during the later period might question the relevance of results based on models which generate such observations.

We generate data from the best 2 and 3-regime models, the 2R-TVECM and the 3R-TVECM, using pseudo random numbers transformed to have the appropriate covariance matrix, and the estimated (full-sample) values of the models' parameters. The number of observations is similar to that in the empirical exercise, and the last 24 observations are kept back for forecasting. These models and the VECM are then estimated, and used to generate forecasts. Our results are based on Monte Carlo estimates of the MSFEs from 2000 replications.

We present the simulation results conditional on the lower regime of negative spreads in the two upper panels of Figure 5. This event characterized around 10% of the simulations, and resulted in gains of around 40 to 50% in predicting $\Delta r(s)$ and $S$ 1-step ahead, highlighting the very different behaviour between the VECM and non-linear models in these circumstances. Notice also that the 2 and 3 regime models behave very similarly because they have the same threshold value for the lower regime. The lower panels in Figure 5 present the MSFE ratios for data generated by the non-linear models, when we consider only those forecasts made (on the roughly 20% of occasions) when the spread exceeded 2.7. The superior performance of the 3-regime model highlights the relevance of allowing for the third regime, at least in the simulation study.

6.3 Comparing with univariate autoregressive models

The choice of the VECM as the linear model comparator may exaggerate the importance of 'non-linearity' from a forecasting perspective if univariate linear autoregressive models are superior to the VECM. So we also generate forecasts from AR models of $\Delta r(s)$, $\Delta r(l)$, and an AR and a random walk model of the spread. The results in terms of MSFEs are plotted in Figure 2 and show that the linear model forecasts are competitive for $\Delta r(s)$, $\Delta r(l)$, but that the TVECMs are superior than an AR for predicting the spread. The random walk provides accurate forecasts (relative to the other models) of the spread at short horizons, matching the finding that the spread is highly persistent except at high or low values. But these forecasts deteriorate at longer horizons. So, at short horizons cointegration does not improve forecasts of the spread, but some improvement from non-linear cointegration is apparent at longer horizons.

7 Economic interpretation of the models

There is a large academic literature on the term structure of interest rates. In this section we discuss the key features of our models in the light of the expectations theory of the term structure. According to Campbell (1995, p.137-8), the expectations theory implies that:

When long rates are unusually high relative to short rates, long rates do not decline to restore the usual yield curve, as one might suppose. Instead long rates tend to rise; the yield spread falls only because short rates rise even faster.

When long rates are sufficiently low relative to short rates, such that the yield curve is inverted, our models are characterized by significant feedback of negative spreads on to lower short rates, and to a lesser extent, on to lower long rates, so that short rates fall, and faster than long rates, consonant with the theory. Negative spreads tend to occur when the current short rate is higher than expected future short rates, so that future short rates would be expected to
be lower than current, a pattern reflected in the models. As noted, the VECM suggests very little response of \( r(s) \) and \( r(l) \) to spreads. From Figure 1 it is apparent that the spread flattens and becomes negative before the economy goes into recession, so that spreads between long- and short-term interest rates act as useful leading indicators of recessions (see, e.g., Hamilton and Kim, 2000).

However, for large positive spreads, our models do not exhibit the positive correlation between long rates and the spread predicted by the theory (see the quotation from Campbell), whereby long rates increase, though less quickly than short rates, in response to the long-rate being too high. But see also the evidence against this in Table 2 of Campbell, p. 139. One strand of argument is that long rates over-react to monetary authority policies aimed at preventing the economy overheating, so that in subsequent periods the disequilibria between the expected values of the short-term rate and the actual long-term rate leads to decreasing long rates, establishing a negative correlation between the spread and long rates.\(^5\)

8 Conclusions

Using a variety of different models and tests for non-linear cointegration, we find strong evidence of non-linearities in the response of interest rates to the spread. An evaluation of the forecasting performance of a number of systems models of US short- and long-term interest rates leads us to conclude that non-linearities, including asymmetries in the adjustment to equilibrium, can result in more accurate short horizon forecasts, especially of the spread (i.e., the difference between long and short rates). However, the gains are not all that large. That non-linearities do not necessarily result in sizable gains is in line with much of the literature. The form of the non-linearity that we identify suggests that linear models will perform relatively well except when either the yield curve is inverted, or spreads are very high, which are not the norm.

As an advance on much of the previous literature in this area, we apply and adapt the testing and specification techniques to a systems analysis of long- and short-term rates, rather than performing a single-equation analysis of the short-term rate. We confirm that the short-term rate responds to narrowing and negative spreads, but we are also able to show that long-term rates adjust when the spread is too high. Between these two extremes, the spread does not help predict either rate.

The out-of-sample forecast exercise does not clearly favour the 3 over the 2-regime model, but we have shown that the former is to be preferred in certain states of nature, viz. when the spread is high. The non-linear models imply different short-term dynamics depending on whether the yield curve is normal or inverted. Even though the latter has not been observed in recent times, a failure to discriminate between the two historical regimes will bias the coefficient estimates and may contribute to inferior forecasts. Comparing the forecasts of the non-linear models with univariate models indicates that linear autoregressions are able to produce competitive forecasts. However, when forecasts for the spread are evaluated, the gains of non-linearity and cointegration are stronger.

A worthwhile focus for future research in this area would be the possibility of a structural break in the way monetary policy was conducted after 1984, and how this might impact on the

\(^5\)Of course, if policy is credible, the long-rate should reflect the expected future low inflation environment brought about by higher short rates. A time-varying risk premium could be another possible explanation for the negative correlation between the spread and long-term interest rate. Hardouvelis (1994) finds little support for this, while Tzavalis and Wickens (1998) take the opposite position.
non-linearity in the system. One starting point would be a model in the spirit of Lundbergh et al. (2000) that allows (possibly smooth) structural breaks in the non-linearity.

References


Table 1  Characteristics of the Models.

<table>
<thead>
<tr>
<th>Model</th>
<th># regimes</th>
<th>Specification and Estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>VECM</td>
<td>1</td>
<td>Linear Vector Equilibrium Correction model with $S_{t-1}$ as the ECM ($\theta = 1$)</td>
</tr>
<tr>
<td>2R-TVECM</td>
<td>2</td>
<td>Threshold estimated by grid search, with $S_{t-1}$ as the ECM. Variances change over regimes</td>
</tr>
<tr>
<td>MTVECM</td>
<td>2</td>
<td>$\theta$ estimated by Engle and Granger (1987) static regression. $w_{t-1}$ is the ECM and $\Delta w_{t-1}$ is the transition variable. Threshold estimated by grid search in MTAR for $\Delta w_t$. Tested for asymmetric adjustment using Enders and Granger (1998) and Enders and Siklos (2001). Short-run dynamics do not depend on the regime.</td>
</tr>
<tr>
<td>2R-TVECM$_{joint}$</td>
<td>2</td>
<td>Joint estimation of $\theta$ and threshold values by grid search within the system with $w_{t-1}(\theta)$ as transition variable. Non-linearity tested with Hansen and Seo (2000).</td>
</tr>
<tr>
<td>3R-TVECM$_{a}$</td>
<td>3</td>
<td>$S_{t-1}$ is the ECM and the transition variable. Thresholds estimated by one-step at-a-time in a univariate SETAR model of $S_t$. SETAR non-linearity tested with Hansen (2000). Variances change over regimes.</td>
</tr>
<tr>
<td>3R-TVECM</td>
<td>3</td>
<td>$S_{t-1}$ is the ECM and the transition variable. Thresholds estimated by grid search one-step-at-a-time within the system. Non-linearity tested with a generalisation of Hansen (2000) for multivariate models.</td>
</tr>
</tbody>
</table>

![Figure 1](image-url)  
Figure 1  Plots of long $r(l)$ (10-year T-bond) and short $r(s)$ (3-month T-bill) rates (top panel), and the spread with NBER-dated recessions marked (bottom panel).
### Table 2 Testing for unit roots and non-linearity.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Augmented Dickey-Fuller</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unit Roots</td>
<td>$r(s), -1.699$</td>
<td>$r(s), -1.890$</td>
</tr>
<tr>
<td></td>
<td>$r(l), -1.572$</td>
<td>$r(l), -1.736$</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s), -8.651^*$</td>
<td>$\Delta r(s), -9.633^*$</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(l), -7.870^*$</td>
<td>$\Delta r(l), -7.109^*$</td>
</tr>
<tr>
<td></td>
<td>$S, -2.497$</td>
<td>$S, -2.795$</td>
</tr>
<tr>
<td>Phillips-Perron Unit Root</td>
<td>$S, -3.716^*$</td>
<td>$S, -3.899^*$</td>
</tr>
<tr>
<td>Johansen cointegration (trace)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>$16.88^*$</td>
<td>$18.6^*$</td>
</tr>
<tr>
<td>Enders and Siklos</td>
<td>$\Phi, 9.84^*$</td>
<td>$\Phi, 10.91^*$</td>
</tr>
<tr>
<td>Asymmetric Cointegration (TAR)</td>
<td>$t - \text{max}, -3.96^*$</td>
<td>$t - \text{max}, -4.26^*$</td>
</tr>
<tr>
<td></td>
<td>$\rho_1 = \rho_2, 2.84[0.09]$</td>
<td>$\rho_1 = \rho_2, 4.42^*[0.04]$</td>
</tr>
<tr>
<td>Enders and Siklos</td>
<td>$\Phi, 8.39^*$</td>
<td>$\Phi, 8.63^*$</td>
</tr>
<tr>
<td>Asymmetric Cointegration (M-TAR)</td>
<td>$t - \text{max}, -3.20^*$</td>
<td>$t - \text{max}, -3.07^*$</td>
</tr>
<tr>
<td></td>
<td>$\rho_1 = \rho_2, 0.11[0.74]$</td>
<td>$\rho_1 = \rho_2, 0.01[0.91]$</td>
</tr>
<tr>
<td>Hansen and Seo</td>
<td>25.183*[0.00]</td>
<td>34.527*[0.00]</td>
</tr>
<tr>
<td>Threshold cointegration</td>
<td>45.17*[0.002]</td>
<td>52.43*[0.00]</td>
</tr>
<tr>
<td>Threshold Cointegration</td>
<td></td>
<td>10.08*[0.14]</td>
</tr>
<tr>
<td>linear vs. two-regime</td>
<td>11.09[0.17]</td>
<td>10.08[0.14]</td>
</tr>
<tr>
<td>two-regime vs. three-regime</td>
<td>11.09 [0.17]</td>
<td>10.08 [0.14]</td>
</tr>
<tr>
<td>System non-linearity (eq. 7)</td>
<td>149.71*[0.01]</td>
<td>182.88*[0.01]</td>
</tr>
<tr>
<td>linear vs. two-regime</td>
<td>149.71*[0.01]</td>
<td>182.88*[0.01]</td>
</tr>
<tr>
<td>two-regime vs. three-regime</td>
<td>149.71*[0.01]</td>
<td>182.88*[0.01]</td>
</tr>
<tr>
<td>System non-linearity</td>
<td>62.24*[0.02]</td>
<td>71.29*[0.03]</td>
</tr>
<tr>
<td>two-regime vs. three-regime</td>
<td>62.24*[0.02]</td>
<td>71.29*[0.03]</td>
</tr>
<tr>
<td>Tsay</td>
<td></td>
<td></td>
</tr>
<tr>
<td>System non-linearity</td>
<td>$d = 1, 116.98^*[0.00]$</td>
<td>$d = 1, 110.54^*[0.00]$</td>
</tr>
<tr>
<td></td>
<td>$d = 2, 68.83^*[0.00]$</td>
<td>$d = 2, 63.94^*[0.00]$</td>
</tr>
<tr>
<td></td>
<td>$d = 3, 105.23^*[0.00]$</td>
<td>$d = 3, 130.13^*[0.00]$</td>
</tr>
<tr>
<td></td>
<td>$d = 4, 87.39^*[0.00]$</td>
<td>$d = 4, 195.43^*[0.00]$</td>
</tr>
</tbody>
</table>

* null hypothesis rejected at least at 5%

Notes: (1) and (3) are based on $p = 7$; (2) is based on truncation lag equal to 5; (4) and (5) are based on positive and negative deviations of the estimated cointegrating combination and $p = 1$; (6), (7), (8) (9) and (10) are based on heteroscedasticity-corrected statistics and on 500 bootstrap samples; (11) is also based on heteroscedasticity-corrected statistic. For all the non-linearity tests we set $d = 1$ and $p = 2$, except for (11) which considers $d = 1, \ldots, 4$. 
Table 3 Results of the estimation of the models.

<table>
<thead>
<tr>
<th>M eq</th>
<th>Lower regime</th>
<th>Middle regime</th>
<th>Upper regime</th>
<th>TVEqCM</th>
<th>M-</th>
<th>joint</th>
<th>2R-</th>
<th>3R-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mult.</td>
<td>$\Delta r(s)$</td>
<td>$\Delta r(l)$</td>
<td>$T_L$</td>
<td>Mult.</td>
<td>$\Delta r(s)$</td>
<td>$\Delta r(l)$</td>
<td>$T_M$</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(l)$</td>
<td>0.099</td>
<td>0.025</td>
<td>0.440</td>
<td>0.058*</td>
<td>0.077</td>
<td>0.022</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s)$</td>
<td>0.400</td>
<td>0.058</td>
<td>-0.007</td>
<td>-0.026*</td>
<td>0.077</td>
<td>0.036</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(l)$</td>
<td>0.411</td>
<td>0.046</td>
<td>0.132</td>
<td>0.037</td>
<td>61</td>
<td>0.059</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s)$</td>
<td>1.217</td>
<td>0.608*</td>
<td>0.0003</td>
<td>0.009</td>
<td>0.122</td>
<td>0.027</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(l)$</td>
<td>0.140</td>
<td>0.067</td>
<td>0.010</td>
<td>0.011</td>
<td>52</td>
<td>-0.006</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s)$</td>
<td>1.669</td>
<td>0.203*</td>
<td>0.003</td>
<td>0.009</td>
<td>83</td>
<td>-0.070</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(l)$</td>
<td>1.669</td>
<td>0.203*</td>
<td>0.199</td>
<td>0.008</td>
<td>52</td>
<td>-0.006</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s)$</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.009</td>
<td>0.127*</td>
<td>(1.14)</td>
<td>92</td>
</tr>
<tr>
<td>2R-</td>
<td>$\Delta r(l)$</td>
<td>0.142</td>
<td>0.067</td>
<td>0.010</td>
<td>0.011</td>
<td>52</td>
<td>-0.006</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s)$</td>
<td>1.669</td>
<td>0.203*</td>
<td>0.003</td>
<td>0.009</td>
<td>83</td>
<td>-0.070</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(l)$</td>
<td>1.669</td>
<td>0.203*</td>
<td>0.199</td>
<td>0.008</td>
<td>52</td>
<td>-0.006</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s)$</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.009</td>
<td>0.127*</td>
<td>(1.14)</td>
<td>92</td>
</tr>
<tr>
<td>3R-</td>
<td>$\Delta r(l)$</td>
<td>0.142</td>
<td>0.067</td>
<td>0.010</td>
<td>0.011</td>
<td>52</td>
<td>-0.006</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s)$</td>
<td>1.669</td>
<td>0.203*</td>
<td>0.003</td>
<td>0.009</td>
<td>83</td>
<td>-0.070</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(l)$</td>
<td>1.669</td>
<td>0.203*</td>
<td>0.199</td>
<td>0.008</td>
<td>52</td>
<td>-0.006</td>
<td>0.010</td>
</tr>
<tr>
<td></td>
<td>$\Delta r(s)$</td>
<td>0.003</td>
<td>0.003</td>
<td>0.003</td>
<td>0.009</td>
<td>0.127*</td>
<td>(1.14)</td>
<td>92</td>
</tr>
</tbody>
</table>

* statistically significant at 5%.

Notes: The models are summarised in Table 1; The long-run multipliers (Mult.) are calculated for each equation treating $\Delta r(s)$ or $\Delta r(l)$ as exogenous. The $\alpha$ for $i = L, M, U$ (lower, middle and upper regime) are the coefficients of the spread variable ($S_t-1$) or the cointegrated relationship ($w_{t-1} = r_{t-1}(l) - \theta r_{t-1}(s)$). $T_t$ refers to the number of observations in each regime.
### Table 4  Rank of tests of forecast accuracy and forecast encompassing.

<table>
<thead>
<tr>
<th></th>
<th>Augmented Diebold and Mariano Forecast Accuracy test – Ranks</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Δr(l)</strong></td>
<td><strong>Δr(s)</strong></td>
<td><strong>S</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>h=1</td>
<td>h=2</td>
<td>h=4</td>
<td>h=8</td>
<td>h=1</td>
<td>h=2</td>
<td>h=4</td>
<td>h=8</td>
<td>h=1</td>
<td>h=2</td>
<td>h=4</td>
</tr>
<tr>
<td>VECM</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>2R-TVECM</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MTVECM</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2R-TVECMjoint</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3R-TVECMu</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>3R-TVECM</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Forecast Encompassing test – Ranks</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>Δr(l)</strong></td>
<td><strong>Δr(s)</strong></td>
<td><strong>S</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>h=1</td>
<td>h=2</td>
<td>h=4</td>
<td>h=8</td>
<td>h=1</td>
<td>h=2</td>
<td>h=4</td>
<td>h=8</td>
<td>h=1</td>
<td>h=2</td>
<td>h=4</td>
</tr>
<tr>
<td>VECM</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>2R-TVECM</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MTVECM</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>2R-TVECMjoint</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>3R-TVECMu</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>3R-TVECM</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Notes:
The ranks for the ADM test are based on the number of models which the model in that row is statistically more accurate than at the 10% level. So a column of 1’s denotes a tie for first place – in pairwise comparisons no model is statistically more accurate (equivalently, less accurate) than any other. The ranks based on the pairwise forecast encompassing tests depend on the number of times that a model is not encompassed by another.
Figure 2  Ratio of MSFEs for VECM against non-linear models, for $\Delta r(l)$, $\Delta r(s)$ and $S$.

Figure 3  Ratio of GFESM for the VECM to the non-linear models.
Figure 4  Ratio of MSFE for $S$ for $\Delta r(l)$ of VECM to the non-linear models, conditional on $S_{t-1} > 2.7$. 
Figure 5  MSFEs for the VECM relative to the 2R-TVECM and the 3R-TVECM for data simulated from one of the TVECMs, conditional on $S_{t-1}$. MSFE for $\Delta r(s)$ conditional on $S_{t-1} \leq 0$, for 2R-TVECM DGP (top left); MSFE for $S$ conditional on $S_{t-1} \leq 0$, for 2R-TVECM DGP (top right); MSFE for $\Delta r(l)$ conditional on $S_{t-1} > 2.7$, for 3R-TVECM (bottom left); MSFE for $S$ conditional on $S_{t-1} > 2.7$, for 3R-TVECM (bottom right).