Evaluating forecasts of interest rates in linear and non-linear error-correction frameworks*

Eric M. Scheffel

December 16, 2008

*The author wishes to extend his gratitude to Dr. Mike Clements for his helpful comments made in his capacity as dissertation supervisor

Abstract

The following paper's aim is to evaluate the respective forecast performance of three distinct but related error-correction models. Coefficient estimates and forecasts are being obtained from a standard linear error-correction, a non-linear threshold error-correction and a non-linear logistic smooth transition error-correction model. These are then compared with a simple AR-model in terms of their forecast accuracy. The algorithms used in estimating the non-linear models are based on two- and three-dimensional grid-searches over appropriate parameters aimed at minimising the RSS-function. The algorithms were coded in GAUSS. The variable to be forecasted is the change in the 3-month U.S. treasury bill and the motivation for non-linear model specification stems from possible transaction costs incurred in switching from one financial instrument into another.

Contents

List of Tables

List of Figures

1 Introduction

The following paper presents a forecast comparison of three distinct, but interrelated, error-correction models. 1-step ahead prediction errors will be generated over a variety of forecast horizons for a standard "linear" error-correction model (ECM), a 3-regime threshold error-correction model (BAND-Tar) and finally a logistic smooth transition error-correction model (LSTAR). The variable to be predicted is the change in the 3-month U.S. treasury bill, which will be explained by past realisations of itself, lagged changes in the 6-month treasury bill and the one-period lagged spread between the levels of 3- and 6-month T-bills. Preliminary statistical analysis with the end of justifying both the application of error-correction as well as non-linear model specification will be carried out. Motivated by a theory of non-negligible transaction costs present in financial markets, possible non-linearities in the data-generating process will be modelled by the above-mentioned model specifications. Coefficient estimates of the various non-linear models will be obtained using two- and three-dimensional grid-searches over pre-specified trust regions aimed at minimising the residual-sum-of-squares (RSS)-function. An interpretation of the asymmetries encountered in the obtained model structures will be offered, followed by the forecast exercise employing various criteria to compare and contrast prediction errors obtained from a simple benchmark "linear" ECM-model with those from an ordinary AR-model as well as the BAND-Tar and LSTAR non-linear model specifications. The findings indicate a statistically weak but otherwise consistent improvement in forecast accuracy of the ECM- over the ordinary AR-model. The non-linear models are shown to reduce the median but to increase the mean forecast variability compared to the standard "linear" ECM-model. Further tests consolidate the view that none of the non-linear models exhibit statistically significant difference in prediction to the ECM-model.

1.1 Literature Review

One major motivation for specifying models in order to forecast financial variables such as the 3-month U.S. treasury bill stems from the potential benefit of being able to better predict looming recessions. This idea was first empirically explored by Estrella and Mishkin (1998) who investigated the usefulness of various financial variables in predicting U.S. recessions, using a Probit modelling approach. Their findings conclude, among other, that the 3-month treasury bill appears to exhibit some statistical power in predicting recessions. More interestingly, their results show that the 3-month bill is decisively better (for very short horizons) at predicting recessions than some longer-maturity financial instrument such as a 6-month commercial paper or a 10-year government bond. For this reason, the following discussion will mostly focus on modelling and forecasting the 3-month, instead of the 6-month bill. Finally, mention should be given to one of Estrella and Mishkin's most important finding, underscoring the superior applicability of the spread in forecasting recessions. This result has been given further attention in papers by Hamilton and Kim (2000) and Anderson and Vahid (2000).

Modelling of a variable such as the 3-month U.S. treasury bill may traditionally have been conducted by specifying a simple Box-Jenkins AR-process, using (for the purposes of this paper) lags of the first difference of the 3-month and the 6-month bill in the autoregressive structure. However, more recent developments in time-series econometrics revolving around the modelling of non-stationary time-series in the presence of cointegration (*see* Engle and Granger, 1987) suggest using an error-correction model specification instead. A relatively recent investigation of the usefulness of error-correction models for forecasting U.S. treasury bills of varying maturities is provided by Hall, Anderson and Granger's (1992) paper, "A Cointegration Analysis of Treasury Bill Yields". The paper concludes that such "linear" ECMs are indeed stable for periods during which the Federal Reserve specifically targeted short-term interest rates, but become unstable during periods of policy regime changes.

However, one possible criticism of such standard "linear" ECM-models is that the error-correction term ensures a permanent attraction effect, even for extremely small deviations from the long-run cointegrating relationship of the variables in question. This could be viewed as a relatively implausible assumption when considering the presence of transaction costs in markets. Simulations carried out by Marshall (1994), who makes use of a general equilibrium asset pricing model, show that extremely small costs of consumption adjustment at the individual level can cause large differences between observed aggregate consumption and the prediction derived from a frictionless model. Evidence from simulation results generated by such structural models indicates that research based on observed asset market behaviour should not overlook these possible effects.

Balke and Fomby's (1997) article "Threshold cointegration" successfully transported the econometric methodology of non-linear adjustment in so-called threshold autoregressions (Tong, 1978) into the realm of error-correction models (ECMs). Since then, applications of non-linear ECMs have proliferated in various exercises aimed at re-assessing hypotheses such as the expectations theory of the term structure of interest rates (Anderson, 1997; Tsay, 1998; Enders and Granger, 1998; Franses and van Dijk, 2000b) or the law of one price (LOP) (Obstfeld and Taylor, 1997; O'Connell and Wei, 1997; Lo and Zivot, 2001; Trinkler and Wolf, 2003). One of the most common motives for the vast majority of these and other related contributions for using non-linear instead of "linear" error-correction models is to allow for the possibility of transaction costs in the empirical modelling process. Accounting for adjustment costs incurred by arbitrageurs trading in financial or commodity markets permits a much more prolonged deviation of time series data like goods prices or investment returns from some externally defined long-run "anchor", as long as these deviations do not become too pronounced in their amplitude. This has been a particularly appealing feature for researchers interested in the law of one price or purchasing power parity, since previous work in these fields has either been ambiguous or outright devastating in their verdict on these fundamental economic laws.

Finally, another motivation for comparing the forecast accuracy of linear with non-linear error-correction models is to assess whether building in such non-linearities into this class of models can significantly improve out-of-sample forecast accuracy. Such an investigation is necessary, because many of the empirical findings on the non-linear modelling of univariate time-series by means of threshold autoregressive (TARs) and self-exciting threshold autoregressive models (SETARs), as well as artificial neural networks and other types of non-linear models, have been rather disappointing in terms of forecast improvements (Stock and Watson, 1999). One explanation for this disappointing record might be that many of these non-linear models are not explicitly derived from economic theory and thus are relatively a-theoretic in nature. It is therefore hardly surprising that a superior in-sample fit seldomly translates into an equally superior out-of-sample forecast performance. Such criticism cannot as such be raised for non-linear error-correction models, since they can be explicitly derived from an economic theory of financial arbitrage in the presence of transaction costs. On the outset, these types of models should therefore theoretically be more promising to deliver on average more accurate forecasts than standard linear error-correction models.

The present paper is adopting a similar line of investigation as that illustrated in Anderson's (1997) paper on transaction costs and nonlinear adjustment in the U.S. treasury bill market. Her paper investigates whether allowing for non-linearities conditional on the sign and value of the (de-meaned) spread between a long- and short-maturity bill can significantly improve out-of-sample forecast performance when compared to a standard "linear" error-correction model.

1.2 Expectations Theory of the Term Structure

The expectations hypothesis of the term structure of interest rates uses a long-run non-arbitrage condition which is a common theme found in many areas of financial theory. According to this condition, annualised returns gained on otherwise homogeneous financial instruments of different maturities should be equalised in the long-run, as temporary differences in returns occurring in the short-run should be quickly arbitrated away by raising the demand and thus price of the financial instrument momentarily yielding the higher return. Applying this condition to our present discussion involving the 3- and 6-month treasury bills would lead to the following relationship:

$$r_t(k) = \frac{1}{k}\left[\sum_{j=1}^{k} E_t r_{t+j-1}(1)\right] + L_t(k) \tag{1}$$

Letting t be equal to the period of three months, then the above relationship simply postulates that the annualised return on the current period's 6-month treasury bill ($r_t(k)$ should be equal to the average of the expected annualised returns on the current period's ($r_t(1)$) and next period's ($r_{t+1}(1)$) 3-month treasury bill plus a term premium, reflecting risk and liquidity premia. Now, provided that both the 3- and 6-month treasury bill time series are integrated of order 1 (I(1)), it can be shown, given the expectations hypothesis is true, that the spread between the 3- and 6-month bill should be of order 0 (I(0)) and therefore represent the cointegrating linear combination between the two series:

$$r_t(k) - r_t(1) = \frac{1}{k}\left[\sum_{i=1}^{k-1}\sum_{j=1}^{i} E_t \Delta r_{t+j}(1)\right] + L_t(k) \tag{2}$$

Notice that equation (2) follows from subtracting the yield of the 3-month bill from both sides of equation (1). Furthermore, the cointegrating property of the spread can be derived analytically by noticing that the spread of the two bills is equal to the finite sum of a series of I(0) terms (the $\Delta r_{t+j}(1)$). In addition to that, this result can be further qualified by observing that the I(0) property given by the right-hand side of equation (2) is valid for $r_t(k) - r_t(1)$ for *any* given t, so that the spread for *any* two yields will be I(0). According to Granger's (1987) representation theorem, and given that both 3- and 6-month bills are I(1) and their spread is I(0), it is possible to write down the following vector error-correction model (VECM):

$$\mathbf{\Delta r_t} = c(B)\mathbf{\Delta r_{t-1}} - \gamma[Z_{t-1}] + \epsilon_\mathbf{t} \tag{3}$$

where $\mathbf{\Delta r_t} = [\Delta Y_t \;\; \Delta X_t]'$ and ΔY_t and ΔX_t are equal to the first difference of the 3-month and 6-month treasury bill, respectively. Notice that in order to simplify any subsequent analysis, $r_t(1)$ has been set equal to Y_t and $r_t(k)$ has been set equal to X_t. Z_{t-1} is simply equal to the error-correction term given by the deviation of the lagged spread from its long-run average value, i.e. $Z_{t-1} = S_{t-1} - \mu$, where $\mu = \frac{1}{n}\sum_{t=1}^{n} S_t$. Finally, $c(B)$ is a backward operator polynomial, i.e. $c(B)^k = 1 + B + B^2 + \ldots + B^k$ and $\epsilon_\mathbf{t}$ is the trivial 2×1 column vector containing the Gaussian error terms.

2 Model specification

2.1 A standard linear ECM model

The preceding section's discussion of the expectations hypothesis of the term structure of interest rates motivated the formulation of a VECM in the 3- and 6-month treasury bills using the linear combination of $[-1\ \ 1]'$ as the cointegrating vector. Unless otherwise stated, most of the following analysis will be focusing on the statistical as well as economic properties of only the first line of this VECM and therefore on a model specification similar to:

$$\Delta Y_t = \phi_0 + \phi_1 \Delta Y_{t-1} + \phi_2 \Delta X_{t-1} - \gamma[Z_{t-1}] + \epsilon_t \qquad (4)$$

where for the sake of expositional simplicity the maximum lag order has been limited to 1. Model (4) is a simple univariate error-correction model, modelling the change in the yield on the 3-month bill by using lagged first differences of the yield on the 3- and 6-month bill as well as the error-correction term as regressors.

Before obtaining OLS estimates of model (4), a variety of statistical properties will have to examined in order to justify the above model specification in the first place. First of all, augmented Dickey-Fuller-type (ADF) tests will have to be carried out on both time series under consideration, in order to ensure the necessary I(1) property. Secondly, the cointegrating vector will have to be estimated. This can either be done by running the "long-run" regression to estimate β

$$Y_t = \beta X_t + \epsilon_t \qquad (5)$$

and subsequently testing the obtained residuals for remaining non-stationarity or by using a Johansen-type (1988) trace test, which offers the benefit of allowing the researcher to test the restriction that the cointegrating vector is indeed $[-1 \;\; 1]'$. Finally, the appropriate lag-lengths for ΔX_t and ΔY_t have to be determined by using information criteria such as the Akaike or Schwartz-Bayesian Information Criterion, given by:

$$\begin{aligned} AIC &= ln\left(\frac{RSS}{d.f.}\right) + \frac{2\times k}{d.f.} \\ BIC &= ln\left(\frac{RSS}{d.f.}\right) + \frac{k\times ln(n-d.f.)}{d.f.} \end{aligned} \tag{6}$$

where RSS is the residual sum of squares, k the number of regressors and $d.f.$ the degrees of freedom. Typically, since the AIC is penalised to a lesser extent for including more regressors it usually suggests using a more elaborate model than that selected by the BIC, which typically selects a more parsimonious lag structure. Also, regardless of the criteria's suggestions, serial correlation should be absent in the final model specification.

2.2 A BAND-Tar model with discrete jumps

2.2.1 Model Description

The first non-linear model to be estimated and subsequently used in obtaining forecasts is a variant of the standard threshold autoregressive (TAR) model as initially proposed by Tong (1978) and Tong and Lim (1980). The general idea underlying the simple TAR is that of a k-regime model in which the value of a pre-chosen threshold variable (with delay parameter d) determines which regime the regressand should follow. Balke & Fomby (1997) extended this technique to the class of error-correction models, commonly associated with the work of Engle & Granger (1987), by allowing the value of the error-correction term to endogenously determine the regime currently in effect. Applying this framework to our present discussion calls for the following model specification:

$$\Delta Y_t = \begin{cases} \phi_{0,1} + \phi_{1,1}\Delta Y_{t-1} + \phi_{2,1}\Delta X_{t-1} - \gamma_1[Z_{t-1}] + \epsilon_t & \text{if} \quad Z_{t-1} \geq \tau_2 \\ \phi_{0,2} + \phi_{1,2}\Delta Y_{t-1} + \phi_{2,2}\Delta X_{t-1} - \gamma_2[Z_{t-1}] + \epsilon_t & \text{if} \quad \tau_2 > Z_{t-1} > \tau_1 \\ \phi_{0,3} + \phi_{1,3}\Delta Y_{t-1} + \phi_{2,3}\Delta X_{t-1} - \gamma_3[Z_{t-1}] + \epsilon_t & \text{if} \quad Z_{t-1} \leq \tau_1 \end{cases} \tag{7}$$

where ΔY_t and ΔX_t are the first difference of the short- and long-term bill, respectively; Z_{t-d} assumes both the role of the disequilibrium error term as well as the threshold variable with delay parameter $d = 1$ and the number of regimes is set equal to $k = 3$. Notice that for the sake of expositional simplicity the lag order has been limited to $p = 1$, whereas in any subsequent model estimations this will be adjusted accordingly.

Contrary to the allegedly a-theoretic nature of self-exciting threshold autoregressive (SETAR) models, which typically follow a k-regime univariate distributed-lag autoregressive process and have particularly proliferated in applications of exchange rate modelling, the threshold error-correction model discussed in this section can be derived from a hypothesis of non-negligible transaction costs in financial markets. Assuming that profit-maximising investors face equal and hence homogeneous transaction costs equal to τ incurred in swapping funds between two financial instruments with cointegrated annualised returns, the above general model (7) with the additional restrictions of $\tau = \tau_1 = -\tau_2$ and $\gamma_2 = 0$ should hold.

$$\Delta Y_t = \begin{cases} \phi_{0,1} + \phi_{1,1}\Delta Y_{t-1} + \phi_{2,1}\Delta X_{t-1} - \gamma[Z_{t-1}] + \epsilon_t & \text{if} \quad |Z_{t-1}| \geq \tau \\ \phi_{0,2} + \phi_{1,2}\Delta Y_{t-1} + \phi_{2,2}\Delta X_{t-1} + \epsilon_t & \text{if} \quad |Z_{t-1}| < \tau \end{cases} \tag{8}$$

Similar to the linear model specification, the essential argument used in arriving at the non-linear TAR model (8) is that of a non-arbitrage condition in efficient financial markets. However, due to homogeneous transaction costs, investors will only rush to swap their funds from one financial instrument into another, once the divergence in annualised returns is bigger than the transaction cost involved in swapping, i.e. $|Z_{t-1}| \geq \tau$. For any divergence in returns smaller than the transaction cost, i.e. $|Z_{t-1}| < \tau$, investors should allow such disequilibria to occur without engaging in arbitrage leading to the eventual equalisation of returns. The resulting model permits the existence of a so-called "band of inaction" in which the short-term bill is allowed to follow a simple AR-process. Only after the threshold variable has either touched or exceeded the critical upper or lower boundary does the attractor effect set in, or in other words, will the error-correction term be "switched on". The final model employed in any subsequent estimation or forecast exercises following below will be a mid-way solution between model (7) and (8), given by:

$$\Delta Y_t = \begin{cases} \phi_{0,1} + \phi_{1,1}\Delta Y_{t-1} + \phi_{2,1}\Delta X_{t-1} - \gamma_1[Z_{t-1}] + \epsilon_t & \text{if} \quad Z_{t-1} \geq \tau_2 \\ \phi_{0,2} + \phi_{1,2}\Delta Y_{t-1} + \phi_{2,2}\Delta X_{t-1} + \epsilon_t & \text{if} \quad \tau_2 > Z_{t-1} > \tau_1 \\ \phi_{0,3} + \phi_{1,3}\Delta Y_{t-1} + \phi_{2,3}\Delta X_{t-1} - \gamma_3[Z_{t-1}] + \epsilon_t & \text{if} \quad Z_{t-1} \leq \tau_1 \end{cases} \tag{9}$$

which is exactly equal to model (7), except for the omission of the error-correction term in regime 2, thus imposing the restriction of a "band of inaction" in the central regime. The reason for favouring the less restrictive model (9) over model (8) is that the former allows for the possibility of investors incurring different costs as they trade bills of different maturities as well as asymmetries conditional on the sign of the error-correction term.

2.2.2 Model estimation method

The method employed in obtaining estimates for the error-correction TAR outlined in section 2.2.1 is that suggested by Leybourne, Newbold, and Vougas (1998), which amounts to using a direct search over an n-dimensional grid, where n depends on the number of parameters over which the objective function has to be minimized/maximized.

The method exploits the fact that for given values of τ_1 and τ_2 model (9) is essentially a linear model which can be estimated using conditional OLS. The procedure requires the generation of 3 indicator variables, one for each regime, in order to "filter out" and assign individual observations with their corresponding regressors to their respective regime. The original (complete) data matrix is then post-multiplied by each of the three indicator variables to give three equally-dimensioned new data matrices which are then horizontally concatenated to give the final data matrix which is being used in estimating values for the coefficient matrix β via OLS. Notice that $HConc$ in (10) represents a function signifying horizontal concatenation.

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \underset{i=1}{\overset{3}{HConc}} \begin{pmatrix} 1 & Y_{1,t-1} & \dots & Y_{1,t-k} & X_{1,t-1} & \dots & X_{1,t-k} & Z_{1,t-1} \\ 1 & Y_{2,t-1} & \dots & Y_{2,t-k} & X_{2,t-1} & \dots & X_{1,t-k} & Z_{2,t-1} \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Y_{n,t-1} & \dots & Y_{n,t-k} & X_{n,t-1} & \dots & X_{n,t-k} & Z_{n,t-k} \end{pmatrix}_{(i)} \times \begin{pmatrix} I(Z_{1,t-1}) \\ I(Z_{2,t-1}) \\ \vdots \\ I(Z_{n,t-1}) \end{pmatrix}_{(i)} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix} \tag{10}$$

Notice that for any arbitrarily chosen pair of arguments (τ_1, τ_2), the above model (10) can be estimated via OLS permitting the computation of the residual sum of squares (RSS). The grid-search method then essentially amounts to an OLS estimation and RSS computation for any conceivable pair of arguments (τ_1, τ_2) within a pre-specified range, all arranged in a two-dimensional grid given by:

$$\begin{pmatrix} (max[Z_{t-1}], max[Z_{t-1}]) & \dots & (min[Z_{t-1}], max[Z_{t-1}]) \\ \vdots & \ddots & \vdots \\ (max[Z_{t-1}], min[Z_{t-1}]) & \dots & (min[Z_{t-1}], min[Z_{t-1}]) \end{pmatrix} \tag{11}$$

The size of this argument matrix depends on how fine the researcher chooses the grid-search to be. This can be controlled by specifying in advance the number of increments by which τ_1 and τ_2 will be updated for each and every single "OLS-run" that is being computed. In addition to that, once the pair of arguments that minimises the RSS-function

$$Q_n(\tau_1, \tau_2) = \underset{(\tau_1,\tau_2)}{argmin} \sum_{i=1}^{n} [y_t - F(x_t; \tau_1, \tau_2)]^2 \tag{12}$$

has been found, the researcher can choose to carry out a finer search in the neighbourhood of the pair $(\tau_1, \tau_2)_{min}$ that minimises the RSS-function, in order to arrive at more precise values of the minimising arguments.

Compared to other optimisation algorithms, like the Gauss-Newton, the Newpton-Raphson or the method of steepest descent, which all make use of some form of the gradient and Hessian of the objective function in order to converge on the optimising parameter values quicker, the direct search method is computationally much more expensive. However, it is also fool-proof, removes the problems associated with local and global optima and with the computation power of currently available personal PCs still produces results in a relatively timely fashion.

2.2.3 Testing for threshold non-linearity

The most commonly used test for threshold non-linearity is that first devised by Tsay (1989). In his paper he remarks that the problem of finding an appropriate test capable of detecting threshold non-linearity can be solved by using an “arranged autoregression”, which permits the use of Chow-type or CUSUM tests for structural breaks in detecting threshold non-linearities. Since the regime currently in effect is determined by the value and sign of the spread, Tsay suggests ordering each and every observation, together with their respective regressors, in ascending order of the value of the spread. Using the following simplified example (using 10 observations), the test is shown to work as follows:

$$\begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{pmatrix} \sim \begin{pmatrix} Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \end{pmatrix} \sim \begin{pmatrix} Y_1 & X_1 & Z_1 \\ Y_2 & X_2 & Z_2 \\ Y_3 & X_3 & Z_3 \\ Y_4 & Y_4 & Z_4 \\ Y_5 & Y_5 & Z_5 \\ Y_6 & Y_6 & Z_6 \\ Y_7 & Y_7 & Z_7 \\ Y_8 & Y_8 & Z_8 \\ Y_9 & Y_9 & Z_9 \end{pmatrix} \tag{13}$$

The above is a horizontal concatenation (illustrated by the $\sim$) of an index column vector, an independent observations vector and a regressor matrix (commonly referred to as the data matrix). For simplicity and expositional clarity the Δs have been omitted. Following Tsay's method, the entire concatenated matrix should now be sorted in ascending order of the value of the spread, which could lead to the following constellation:

$$\begin{pmatrix} 4 \\ 9 \\ 6 \\ 1 \\ 3 \\ 8 \\ 2 \\ 5 \\ 9 \end{pmatrix} \sim \begin{pmatrix} Y_5 \\ Y_{10} \\ Y_7 \\ Y_2 \\ Y_4 \\ Y_9 \\ Y_3 \\ Y_6 \\ Y_8 \end{pmatrix} \sim \begin{pmatrix} Y_4 & X_4 & Z_4 \\ Y_9 & X_9 & Z_9 \\ Y_6 & X_6 & Z_6 \\ Y_1 & Y_1 & Z_1 \\ Y_3 & Y_3 & Z_3 \\ Y_8 & Y_8 & Z_8 \\ Y_2 & Y_2 & Z_2 \\ Y_5 & Y_5 & Z_5 \\ Y_7 & Y_7 & Z_7 \end{pmatrix} \tag{14}$$

After the entire matrix has been sorted in ascending order of the spread (Z_t), it is now possible to test the hypothesis, that - for instance - every observation bounded between the lowest spread value Z_7 and Z_3 belongs to one specific regime, and that all the other observations are generated by either one or even multiple other data generating processes. In order to test this, the observations allegedly belonging to the one regime, have to be "sliced off" or separated from the entire matrix, to give the following:

$$\begin{pmatrix} 3 \\ 8 \\ 2 \\ 5 \\ 9 \end{pmatrix} \sim \begin{pmatrix} Y_4 \\ Y_9 \\ Y_3 \\ Y_6 \\ Y_8 \end{pmatrix} \sim \begin{pmatrix} Y_3 & Y_3 & Z_3 \\ Y_8 & Y_8 & Z_8 \\ Y_2 & Y_2 & Z_2 \\ Y_5 & Y_5 & Z_5 \\ Y_7 & Y_7 & Z_7 \end{pmatrix} \tag{15}$$

If the selected observations are indeed best described by one specific regime, then estimating this regime over the selected range (up until the hypothesised threshold value Z_3) could provide a starting model from which Chow-type or CUSUM tests could be carried out. This can be done, because any predictive residuals obtained from 1-step ahead recursive forecasts beyond the hypothesised threshold value should violate the orthogonality condition of being uncorrelated with the regressors, since Z_3 should mark a regime-break. Therefore, it should be possible to test for this orthogonality condition by regressing a series of predictive residuals obtained via recursive OLS on the regressors and then computing a standard F-test. Beware however, that before estimating this starting model, which only uses the hypothesised observations belonging to one specific regime, the original dynamic (chronological) ordering will have to be restored. This can be achieved by sorting the above matrix (15) in ascending order of the index column vector from top to bottom, resulting in:

$$\begin{pmatrix} 2 \\ 3 \\ 5 \\ 7 \\ 8 \end{pmatrix} \sim \begin{pmatrix} Y_3 \\ Y_4 \\ Y_6 \\ Y_8 \\ Y_9 \end{pmatrix} \sim \begin{pmatrix} Y_2 & Y_2 & Z_2 \\ Y_3 & Y_3 & Z_3 \\ Y_5 & Y_5 & Z_5 \\ Y_7 & Y_7 & Z_7 \\ Y_8 & Y_8 & Z_8 \end{pmatrix} \tag{16}$$

This dynamically restored matrix could then (omitting the index vector) be used to estimate the starting model from which to obtain 1-step ahead predictive residuals, by using the remaining observations (Z_1 to Z_4 in matrix (13)), which would also have to be dynamically re-sorted into their original time-consistent ordering.

In his paper on "Testing and Modeling Threshold Autoregressive Processes", Tsay does not only show how his test can be used to detect threshold non-linearity (even when the threshold is not known in advance), but how subsequent graphical analysis can aid in identifying which values of the threshold variable are most likely to constitute break-points separating possible multiple regimes. However, if one wishes to impose the a-priori assumption of a 3-regime threshold error-correction model, as is presently the case, one could also take a reversed modelling strategy. That is, one could first optimise the RSS-function over a pre-chosen model specification (i.e. three regimes, hence two thresholds), and subsequently use the obtained RSS-minimising threshold values as the hypothetically "true" values in the above-described F-test for non-linearity.

A further, but statistically much more complicated, method applicable to detecting threshold non-linearity is that devised by Bruce Hansen (1997; 2000). This test, which can be cast as a likelihood ratio or F-test, looks at how much of a reduction in the residual variance has occurred under the alternative of a non-linear model when compared to the residual variance under the null, that is under linear specification. The test is of the following form:

$$F(\hat{\tau}_1, \hat{\tau}_2) = n \left(\frac{\tilde{\sigma}^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \right) \tag{17}$$

Notice that because τ_1 and τ_2 are both estimated, the above test statistic represents the so-called "supremum" of a number of dependent statistics, each of which follow an asymptotic χ^2-distribution. Therefore, "...because the exact form of the dependence between the different $F(\hat{\tau}_1, \hat{\tau}_2)$ is difficult to analyse or characterise, critical values are most easily defined by means of simulation" (Franses and van Dijk, 2000a). However difficult the computation of critical values for this test-statistic may be, the researcher interested in the validity of her various non-linear model specifications should always compare and contrast the σ^2 obtained from the linear model with those of her non-linear models, and be suspicious if they are close to being equal.

2.3 A LSTAR model with smooth adjustment

2.3.1 Model Description

The second type of non-linear model used to obtain forecasts for the first difference of the 3-month T-bill is the so-called logistic smooth transition autoregressive model (LSTAR), which is commonly associated with the work of Chang and Tong (1986) and Teräsvirta (1994). As opposed to the non-linear threshold error-correction model discussed in section 2.2.1, which described the attraction effect of the error-correction term as an "on-off" property depending on the size and sign of Z_{t-1}, the LSTAR version of the error-correction model permits a completely smooth modelling of the adjustment effect towards long-run equilibrium. The particular model described in this section will be the same as that put forth by Anderson (1997) in her paper on nonlinear adjustment in the U.S. bills market. She chooses that specification as a way to model adjustment towards equilibrium in financial markets with heterogeneous transaction costs, as opposed to the homogeneous transaction costs case illustrated in section 2.2.1. When transaction costs are assumed to differ among varying types of market participants, then the "on-off" property of the TAR-model will have to be replaced by the notion of blurred regions around the threshold(s) in which the error-correction term's dominance gradually increases the further the series at hand diverge from their long-run cointegrating relationship. The proposed model specification will take on the following form:

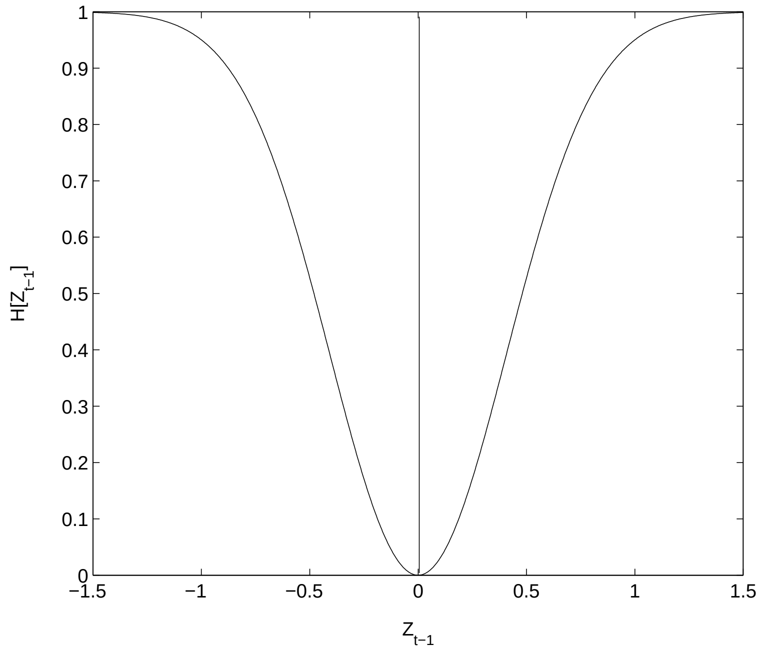
$$\Delta Y_t = \phi_{0,1} + \phi_{1,1}\Delta Y_{t-1} + \phi_{2,1}\Delta X_{t-1} +$$
$$[\phi_{0,2} + \phi_{1,2}\Delta Y_{t-2} + \phi_{2,2}\Delta X_{t-1} - \gamma Z_{t-1}] \times [H(|Z_{t-1}|)] + \epsilon_t \quad (18)$$

The characteristic feature of this particular model is the logistic transition function given by $[H(|Z_{t-1}|)]$. It can be thought of as a cumulative probability distribution of the entire market participants' "trigger thresholds". In other words, for a given absolute value of $|Z_{t-1}|$, the function specifies the proportion of market participants who, given their individual transaction cost τ_i, will find it profitable to engage in equilibrating arbitrage. Therefore, if $|Z_{t-1}|$ is very small and the market has only diverged slightly from its long-run equilibrium, very few traders will benefit from low enough transaction costs to engage in profitable arbitrage. On the other hand, if the spread is very large, almost everyone will be able to engage in profitable arbitrage, leading to a much stronger attraction effect back towards the long-run equilibrium. A typical specification of this logistic function is given below.

$$H(Z_{t-1}) = 1 - \exp[-\lambda(Z_{t-1})^2]\,, \qquad \lambda > 0 \tag{19}$$

Using this particular response function, however, only permits the modelling of perfectly symmetrical responses conditional on the de-meaned spread (i.e. only the absolute value of Z_{t-1} is of relevance, whether the short-term bill is over- or under-priced does not play any role). This can be observed more clearly in figure 1, in which we have set $\lambda = 3$, which is a measure of the steepness of the response (or transition) function. The graphical illustration allows the reader to better comprehend how the response function works. It demonstrates the above-described properties of $H(Z_{t-1})$ and clearly shows how the function is bounded between 0 and 1, gradually approaching the upper bound of 1, the larger the (absolute) value of Z_{t-1} becomes.

Figure 1: Symmetric logistic transition function

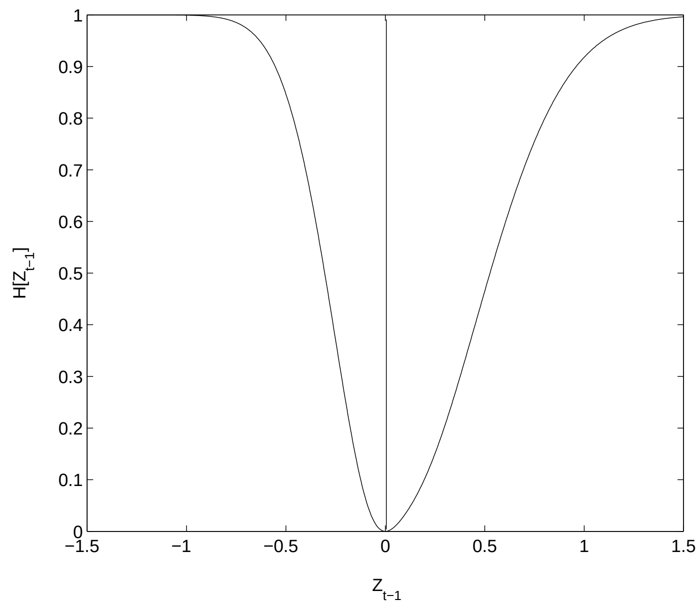


Anderson (1997) uses a modified version of this response function, which permits the modelling of asymmetric adjustment conditional on the sign of the error-correction term. The form of this modified $H(.)$ is as follows:

$$\begin{aligned} H(Z_{t-1}) &= 1 - \exp[-\lambda(Z_{t-1})^2] \times h(Z_{t-1}), \quad \text{with} \\ h(Z_{t-1}) &= 0.5 + [1 + \exp(-\delta(Z_{t-1}))]^{-1} \end{aligned} \tag{20}$$

Comparing figure 2 generated by the modified response function (20) with the preceding figure 1 immediately reveals how the former is capable of accounting for asymmetrical responses conditional on the sign of Z_{t-1}. Notice that this is achieved by the inclusion of an additional parameter, δ, which will also have to be determined in any estimation procedure.

Figure 2: Asymmetric logistic transition function



For expositional purposes, the above illustration has been generated using $\lambda = 5$ and $\delta = -15$. Thus, in this example, the proportion of traders willing to engage in arbitrage when the short-term bill is over-priced is much larger for any given (absolute) market disequilibrium, than would be the case for a situation in which the short-term bill was under-priced.

Finally, the specification of the LSTAR-type error-correction model is completed by pointing out one more important difference between the previous non-linear model and the one described in this section. Following Anderson (1997), the specification of the LSTAR-model is further "loosened" compared to the error-correction TAR by allowing another parameter to be estimated instead of holding it fixed. This parameter is the centrality parameter, denominated by c (denominated by μ for the threshold model), which gives the long-run equilibrium value of the spread, around which the same is allowed to fluctuate. So instead of using $c = \mu = \frac{1}{n}\sum_{t=1}^{n} S_t$, as was previously assumed, c will now be included in the set of coefficients to be estimated.

2.3.2 Model estimation method

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} =$$

$$\begin{pmatrix} 1 & Y_{1,t-1} & \dots & Y_{1,t-k} & X_{1,t-1} & \dots & X_{1,t-k} \\ 1 & Y_{2,t-1} & \dots & Y_{2,t-k} & X_{2,t-1} & \dots & X_{1,t-k} \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Y_{n,t-1} & \dots & Y_{n,t-k} & X_{n,t-1} & \dots & X_{n,t-k} \end{pmatrix} \sim$$

$$\begin{pmatrix} 1 & Y_{1,t-1} & \dots & Y_{1,t-k} & X_{1,t-1} & \dots & X_{1,t-k} & Z_{1,t-1} \\ 1 & Y_{2,t-1} & \dots & Y_{2,t-k} & X_{2,t-1} & \dots & X_{1,t-k} & Z_{2,t-1} \\ 1 & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & Y_{n,t-1} & \dots & Y_{n,t-k} & X_{n,t-1} & \dots & X_{n,t-k} & Z_{n,t-1} \end{pmatrix} \times \tag{21}$$

$$\begin{pmatrix} H(|Z_{1,t-1}|) \\ H(|Z_{2,t-1}|) \\ \vdots \\ H(|Z_{n,t-1}|) \end{pmatrix} \times \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

The estimation method employed in obtaining values for the set of coefficients in question is essentially the same as the one used in section 2.2.1. The only difference in case of the LSTAR-model, is that the RSS-function will have to be minimised over three arguments, namely λ (the steepness-argument), δ (the asymmetry-argument) and c (the centrality-argument). Except for this slight difference, estimates will be obtained as usual via conditional OLS by exploiting the fact that for fixed values of λ, δ and c, the response function $H(Z_{t-1})$ will be unambiguously determined for each and every value of Z_{t-1}. Therefore, for any given triple of arguments (λ, δ, c), it is possible to write down the model given in equation (21), which is precisely the matrix representation of the model specification given by equation (18), and estimate it via OLS. Notice that first the data matrix containing the regressors of the outer regime (the matrix containing the error-correction terms) will have to be post-multiplied (row-by-row) with the column vector containing the values generated by the response-function and only then be concatenated with the data matrix belonging to the central regime. The concatenation sign is, as above, illustrated by the tilde character. Analogous to section 2.2.1, any arbitrarily chosen triple (λ, δ, c) permits estimation of the above model via OLS and the subsequent computation of the residual sum of squares. However, now that the RSS-function requires minimisation over three parameters, the grid-search will have to be carried out over a RSS-matrix generated by a three-dimensional argument matrix. Since GAUSS does not support "information-cubes", such as for instance $M = [x, y, z]$ directly, multiple layers of two-dimensional matrices will have to be searched for the smallest RSS-value. The algorithm employed works as follows:

$$\text{start[c]} \longrightarrow \begin{pmatrix} (\text{start}[\lambda], \text{start}[\delta]) & \ldots & (\text{end}[\lambda], \text{start}[\delta]) \\ \vdots & \ddots & \vdots \\ (\text{start}[\lambda], \text{end}[\delta]) & \ldots & (\text{end}[\lambda], \text{end}[\delta]) \end{pmatrix} \longrightarrow \underset{\lambda,\delta,c}{argmin} \quad [RSS]$$

$$\vdots$$

$$\begin{matrix}\text{start[c]} \\ -\text{inc}[c]\end{matrix} \longrightarrow \begin{pmatrix} (\text{start}[\lambda], \text{start}[\delta]) & \ldots & (\text{end}[\lambda], \text{start}[\delta]) \\ \vdots & \ddots & \vdots \\ (\text{start}[\lambda], \text{end}[\delta]) & \ldots & (\text{end}[\lambda], \text{end}[\delta]) \end{pmatrix} \longrightarrow \underset{\lambda,\delta,c}{argmin} \quad [RSS]$$

$$\vdots$$

$$\text{end[c]} \longrightarrow \begin{pmatrix} (\text{start}[\lambda], \text{start}[\delta]) & \ldots & (\text{end}[\lambda], \text{start}[\delta]) \\ \vdots & \ddots & \vdots \\ (\text{start}[\lambda], \text{end}[\delta]) & \ldots & (\text{end}[\lambda], \text{end}[\delta]) \end{pmatrix} \longrightarrow \underset{\lambda,\delta,c}{argmin} \quad [RSS]$$

After having pre-specified in advance the number of increments for λ, δ and c over which the search should be carried out, the algorithm simply fixes the value for c, searches two-dimensionally over λ and δ and finally records the lowest attained RSS-value for the current "c - layer". After that c is reduced by one increment, held fixed at its new value, and a new search over λ and δ is being carried out. The iterative process comes to an end, once the terminal value of c has been reached. After that, one only needs to pick out the lowest RSS-value from the "layer - vector" containing the lowest RSS-value for each and every single "c - layer" in order to obtain the final optimised RSS-value, including its argument triple, $(\lambda, \delta, c)_{min}$.

2.3.3 Testing for STAR non-linearity

The test used in determining the validity of a smooth transition non-linear model specification is that first discussed in Luukkonen, Saikkonen, and Teräsvirta (1988). The original test has been devised for standard smooth transition autoregressive (STAR) processes and will have to be modified accordingly in order to allow for the spread to endogenously determine any possible non-linearities present in the data-generating process. The following is going to attempt to provide more of an intuitive explanation of how the test works instead of a rigorous derivation from first principles. As it turns out, the test procedure is somewhat similar in spirit to the standard Ramsey (1969) RESET test for model mis-specification.

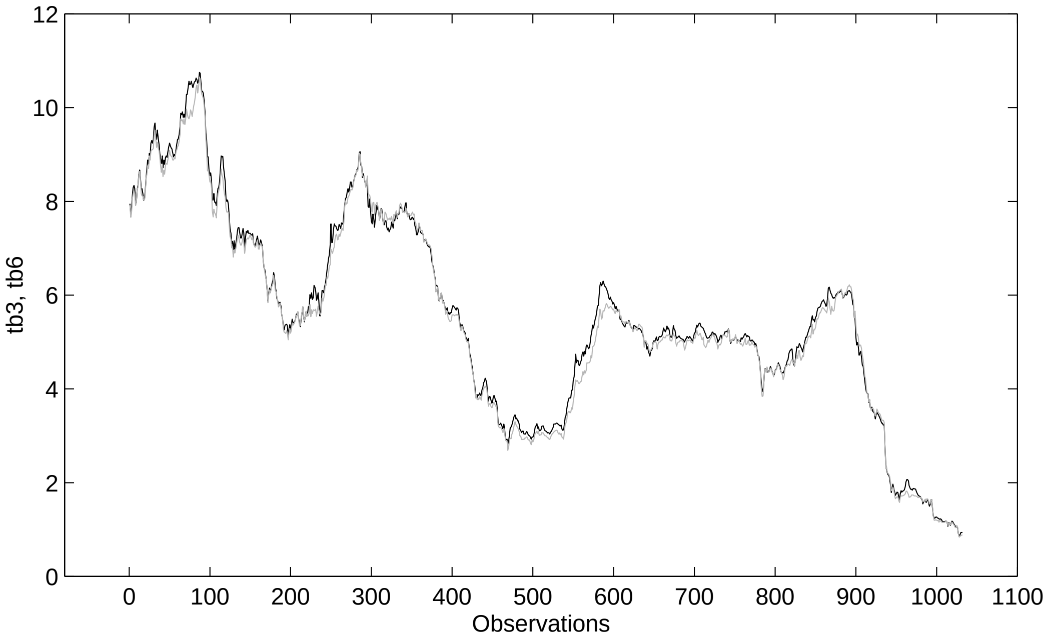
Luukkonen, Saikkonen, and Teräsvirta (1988) approach the problem of testing for linearity against smooth transition non-linearity by the usual formulation of a regression nesting both the alternative (non-linear model) and the null (linear specification) in one equation with the only caveat of having a so-called "unidentified nuisance parameter" present under the null. This simply means that the non-linear model contains a certain parameter which is not restricted under the null hypothesis and which is not present in the linear model. In our case this nuisance parameter is λ, which is unidentified (i.e. $\lambda = 0$) under the null hypothesis of a linear model specification. Luukkonen, Saikkonen, and Teräsvirta circumvent this problem by using a third-order Taylor approximation around $\lambda = 0$ of the above-discussed logistic transition function, leading to the formulation of the following auxiliary regression:

$$\Delta Y_t = \gamma_0 + \beta_{0,i}\mathbf{\Delta X_{t-1}} + \beta_{1,i}\mathbf{\Delta X_{t-1}}S_{t-1} + \beta_{2,i}\mathbf{\Delta X_{t-1}}S_{t-1}^2 + \beta_{3,i}\mathbf{\Delta X_{t-1}}S_{t-1}^3 \quad (22)$$

where $\mathbf{\Delta X_{t-1}} = [\Delta Y_{t-1} \ \ \Delta X_{t-1} \ \ S_{t-d}]'$ with $d = 1$ and the order of lags has been limited to $p = 1$ for expositional purposes. Also, notice that depending on the chosen lag-order for ΔY_t and ΔX_t, the βs are not single coefficients but column vectors containing a varying number of coefficients. In the above case of $p = 1$, the individual βs each contain three coefficients, one for ΔY_t, one for ΔX_t and another one for S_{t-1}. It is now clearer to see how the above test specification can be related to the standard RESET test. Where the RESET test is an omitted-variables test for $\hat{Y}^2$ and $\hat{Y}^3$, here the null hypothesis to be tested is $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$. In other words, if the data were generated by a "linear" ECM model, then adding the additional terms $\mathbf{\Delta X_{t-1}}S_{t-1}$, $\mathbf{\Delta X_{t-1}}S_{t-1}^2$ and $\mathbf{\Delta X_{t-1}}S_{t-1}^3$ should not improve the fit (drastically minimise the RSS) compared to the "linear" ECM model specification. Therefore, the above test can simply be cast as a standard F-test for the omission of β_1, β_2 and β_3. Notice that testing for this particular restriction is exactly the same as testing $H_0 : \lambda = 0$ in the smooth transition error-correction model presented in the previous section.

3 Data and forecast evaluation methods

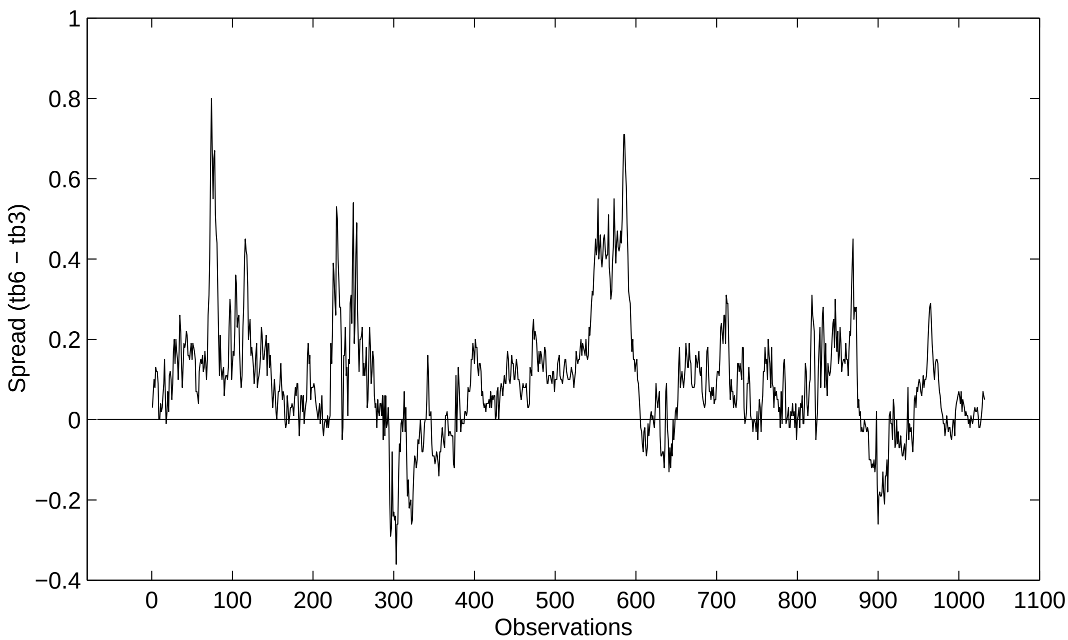
Figure 3: 3-month and 6-month T-bills



3.1 The data

The data used in the present analysis are weekly series of annualised yields (in percentages) of the 3- and 6-month U.S. treasury bills as traded in the secondary market. Furthermore, the data have been compiled using an average of the business days in a week and are quoted on a discount basis. The source of the time-series has been the St. Louis Federal Reserve's online database FREDII, which holds, among other types of data, a large variety of interest rate series, many of which are available in different frequencies (daily, weekly, monthly, quarterly and annually) and which date back as early as the mid-1950s. Following Anderson's (1997) (who uses a different data set obtained from the Bloomberg database) line of reasoning, the choice of higher frequency data than that provided by quarterly or monthly series, reflects the researcher's intention of wanting to capture adjustments brought about by arbitrage "in a market in which investors are known to react quickly" (Anderson, 1997). It is therefore more sensible and appropriate to make use of weekly rather than quarterly data, which is not suitable for capturing such short-run dynamics as the ones investigated in the present paper. More importantly though, not the entire available data span has been exploited in generating the below given results, but only data starting from 1983 has been used in model specification and forecast exercises. This is due to Anderson's observation that error-correction models for the U.S bills market using pre-1983 data points appear to be unstable. The period to be investigated ranges from 01/07/1983 - 07/18/2003, where unusual observations (outliers) pertaining to the period of 10/23/1987 - 07/22/1988, which roughly describes the 1987 stock market crash and its knock-on effects, have been excluded from the final sample.

Figure 4: Spread between 3-month and 6-month T-bills



3.2 Forecast evaluation methods

Once estimates of the set of coefficients in question have been obtained via the above described non-linear least-squares method, 1-step ahead forecast errors will be generated for various pre-specified benchmark periods, over which all three models' forecast performances will be compared with one another. The most common criterion used in assessing the accuracy and robustness of forecasts obtained from various model specifications, is the mean-squared prediction error, or short MSPE, given by:

$$\text{MSPE} = \frac{1}{m}\sum_{j=1}^{m}(\hat{y}_{n+j|n+j-1} - y_{n+j})^2 \tag{23}$$

where the first term in brackets is the 1-step ahead forecast of the dependent variable, i.e. y_{t+1} conditional on the information set known at time t. The intuition behind evaluating forecasts using the MSPE is that of categorising the "seriousness" of various forecast errors according to a symmetric quadratic loss function, such that more ample deviations of predictions from their actual realisations are penalised exponentially. In some circumstances, it may also be desirable to compute the median-squared prediction error (MedSPE) instead. This particular forecast criterion is often advocated on the grounds that it may be more appropriate for evaluating forecasts of financial returns, as they can display rather erratic behaviour. Notice that a typical distribution of squared prediction errors (SPEs) for the data and models used in the present paper is heavily negatively skewed (*see* figures 10 and 11, section A.2), since there exists a disproportionately large number of comparatively small prediction errors. Therefore, if one wishes to place a greater emphasis on evaluating the quality of forecast errors within the median of the total forecast errors' distribution, using the MedSPE should be preferred to using the MSPE. Another very similar criterion focuses on the mean-absolute prediction error (MAPE), rather than the mean-squared prediction error. This criterion is computed as:

$$\text{MAPE} = \frac{1}{m}\sum_{j=1}^{m} |\hat{y}_{n+j|n+j-1} - y_{n+j}| \tag{24}$$

One reason to use the MAPE is that it does not attach excessive weight to very large prediction errors, by squaring them. Generally speaking, it represents a measure of the average variability of prediction. Notice that akin to the MSPE and MedSPE, it is also possible to compute the median-absolute prediction error (MedAPE), by simply taking the median of the absolute prediction errors.

Moreover, in trying to directly compare and contrast the forecast performances of various models considered in the present paper, Diebold and Mariano (1995) have suggested using the so-called "loss-differential" (of squared prediction errors) in the computation of various forecast criteria, which is given by:

$$d_j = e^2_{n+j|n+j-1,A} - e^2_{n+j|n+j-1,B} \tag{25}$$

where $e^2_{n+j|n+j-1,A}$ and $e^2_{n+j|n+j-1,B}$ represent the squared prediction errors at time t+j generated by forecasts from models A and B, respectively. The obtained loss-differential for two model specifications can then be incorporated in the computation of various test statistics aimed at detecting possible qualitative differences (or there lack of) in forecasts. One of such statistics is the sign test statistic, given by:

$$S = \frac{2}{\sqrt{m}}\sum_{j=1}^{m}\left(I[d_j > 0] - \frac{1}{2}\right) \overset{a}{\sim} N(0,1) \tag{26}$$

This test statistic postulates as the null that the median loss differential for two identical forecast processes should be equal to zero. Therefore, a test statistic close to zero should be interpreted as qualitative equality in forecasts of the two models being compared.

However, the above-given sign test statistic compares only the relative magnitude of prediction errors between two models A and B. Therefore, Diebold and Mariano (1995) have also developed a test statistic comparing the absolute magnitudes by testing whether the average loss differential $\bar{d} = \frac{1}{m}\sum_{j=1}^{m} d_j$ is significantly different from zero. The test statistic is given by:

$$DM = \frac{\bar{d}}{\sqrt{w}} \overset{a}{\sim} N(0,1) \qquad \text{where} \qquad \hat{w} = \sum_{i=-(h-1)}^{h-1} \hat{\gamma}_i(d) \tag{27}$$

which, self-evidently, uses as the null the postulate that two identical forecast processes should have an average loss differential equal to zero. Notice that Diebold and Mariano propose estimating the asymptotic variance (w) of $\hat{d}$ by an unweighted sum of autocovariances of d_j, denoted above by $\hat{\gamma}_i(d)$. However, for the purposes of the present discussion, which only involves the computation of 1-step ahead forecast errors, $\hat{w}$ is simply the variance of d_j given by $\hat{\gamma}_0(d)$. In addition to that, it should be mentioned that this test statistic has the advantage of not having to conform to a number of assumptions underlying many other forecast criteria (*see* Clements and Hendry, 1998).

It can also be helpful to compare different forecast processes by evaluating their success in predicting simply the direction of movement of the dependent variable under consideration. This may be particularly relevant for asset returns as investors may be more interested in accurate forecasts of the direction in which, e.g. the stock market is moving, than in the exact magnitude of change. For this purpose it may be useful to compute the following so-called success-ratio (SR):

$$SR = \frac{1}{m}\sum_{j=1}^{m} I_j[y_{n+j} \cdot \hat{y}_{n+j|n+j-1} > 0] \tag{28}$$

which simply calculates the fraction of the m forecasts of $\hat{y}_{n+j|n+j-1}$ that have the same sign as the actual values, y_{n+j}. Given the above SR-statistic, Pesaran and Timmermann (1992) have devised a statistic, which tests whether the value of SR differs significantly from the success ratio that would be obtained where y_{n+j} and $\hat{y}_{n+j|n+j-1}$ are independent. Following Pesaran and Timmermann, define:

$$P = \frac{1}{m}\sum_{j=1}^{m} I_j[y_{n+j} > 0] \tag{29}$$

and further

$$\hat{P} = \frac{1}{m}\sum_{j=1}^{m} I_j[\hat{y}_{n+j|n+j-1} > 0] \tag{30}$$

Using these two expressions, the success rate in case of independence (denominated by SRI) of y_{n+j} and $\hat{y}_{n+j|n+j-1}$ can be computed via the following formula:

$$SRI = P\hat{P} + (1-P)(1-\hat{P}) \tag{31}$$

which has variance given by:

$$\begin{aligned} \mathrm{var}(SRI) = \frac{1}{m}\Big[&(2\hat{P}-1)^2 P(1-P) + \ldots \\ &(2P-1)^2\hat{P}(1-\hat{P}) + \frac{4}{m}P\hat{P}(1-P)(1-\hat{P})\Big] \end{aligned} \tag{32}$$

Furthermore, an expression for the variance of the actual success ratio (i.e. SR) is simply given by:

$$\text{var}(SR) = \frac{1}{m} SRI(1 - SRI) \tag{33}$$

Finally, the so-called Directional Accuracy (DA) test of Pesaran and Timmermann can be calculated as follows:

$$DA = \frac{(SR - SRI)}{\sqrt{\text{var}(SR) - \text{var}(SRI)}} \overset{a}{\sim} N(0, 1) \tag{34}$$

where the asymptotic standard normal distribution is obtained under the null hypothesis that y_{n+j} and $\hat{y}_{n+j|n+j-1}$ are independently distributed. This concludes the present discussion of forecast evaluation criteria, which are going to be applied to any forecast analysis to be conducted and reported below.

4 Results

This section will be focusing on two kinds of computational results. First of all, various statistical tests motivating the application of error-correction models as well as non-linear model specification, will be reported below. Secondly, both obtained model estimates and forecast evaluations will be summarised leading to some first tentative interpretations of the properties of the estimated model coefficients and their respective forecast performances.

4.1 Time-series properties of the data

The following subsection is going to provide a summary of the statistical properties of the time series under investigation. The first part will be focusing on stationarity issues as well as cointegration properties whereas the second part's emphasis will lie on a statistical investigation of the appropriateness of non-linear model specification.

4.1.1 Stationarity and Cointegration

Before carrying out any estimation and forecast exercises, a series of statistical prerequisites, justifying the application of error-correction models, will have to be met. Most of the test statistics reported in this section, which by now have become standard preliminary tests in the time-series literature, have been generated using either EViews4.1 or PcGive10. First of all, the short- and long-term U.S. treasury bill series will have to be tested for non-stationarity. This is commonly accomplished by running an augmented Dickey-Fuller (ADF) regression of the following type:

$$\Delta Y_t = \psi_0 Y_{t-1} + \sum_{i=1}^{p-1} \psi_i \Delta Y_{t-i} + \mu + \delta t + \omega_t \tag{35}$$

where both a linear deterministic time trend and an intercept term have been included. The lag-length is automatically chosen according to the Akaike Information Criterion and the null to be tested is $H_0 : \psi_0 = 0$.

Table 1: ADF-test on levels

Series	t-stat	10%-value	Prob.
tb3	-1.721384	-3.129207	0.7413
tb6	-1.780854	-3.129207	0.7136

The obtained statistics clearly permit to confidently proceed with any further analysis assuming that both series under consideration contain a unit root, since the reported t-ratios are well below the 10%-critical value. However, in order to know the exact order of integration it is useful to first-difference the series under consideration and carry out the same test as above in order to investigate possible I(2) properties.

Table 2: ADF-test on Δ_1

Series	t-stat	1%-value	Prob.
Δtb3	-12.08628	-3.967009	0.0000
Δtb6	-12.31662	-3.967009	0.0000

After having firmly established that the hypotheses of tb3 $\sim$ I(1) and tb6 $\sim$ I(1) cannot be rejected, attention will now turn to attempting to detect and quantify any possible cointegrating linear combination of the two series at hand. In order to achieve this, system-based Johansen (1988) cointegration-statistics will be reported below.

Table 3: Unrestricted Cointegration Rank test

Hypothesised No. of CE(s)	Eigenvalue	Trace Statistic	5% Critical Vale	1% Critical Value
None	0.051332	55.33899	15.41	20.04
At most 1	0.001187	1.21973	3.76	6.65

The above given Trace-statistic generated in EViews reports that there exists exactly one cointegrating relationship. This confirms the a-priori belief that the 3- and the 6-month bill should share only one cointegrating linear combination, namely their spread. More interestingly though, as part of the Johansen cointegration test, EViews also reports efficient estimates of the cointegrating vector, including standard errors. It is therefore possible to test the restriction that the cointegrating vector is indeed equal to $\beta = [-1 \;\; 1]'$.

Table 4: Estimated and restricted beta

Series	Est. Coint. Comb.	Standard Error	Restr. Coint. Comb.	$\chi^2(1)$	Prob.
tb3	-1.001389	(0.00828)	-1.000000	0.000319	0.985744
tb6	1.000000	(–)	1.000000	(–)	(–)

Again, the statistics reported above only confirm a-priori beliefs derived from the Expectations hypothesis of the term structure of interest rates given in section 1.2. The estimated cointegrating linear combination is only negligibly different from $[-1 \;\; 1]'$, a view that is further underscored by the likelihood ratio-test reporting a probability of close to 99% that the a-priori hypothesised cointegrating linear combination cannot be rejected.

4.1.2 Detection of Non-Linearity

This subsection is going to report and interpret results obtained from the above-discussed methods of detecting possible non-linearities in the data-generating process. First of all, various F-tests aimed at revealing structural breaks (regime changes) will be computed in order to justify the use of a 3-regime BAND-Tar model. Secondly, the above-described statistic for testing linearity against possible STAR non-linearity will be applied to the data set at hand in an attempt to detect any possible smooth-transition non-linearities in the data-generating process. Notice that the lag-order of any of the following regressions has been determined by using the Schwartz-Bayesian Information Criterion, which has selected $p_1 = 4$ and $p_2 = 1$ for ΔY_t and ΔX_t, respectively. The lag order has been computed via a grid-search for the standard "linear" error-correction model searching over a maximum of 50 lags for both regressors. In order to avoid extremely time-consuming and computationally expensive algorithms, this lag specification is assumed to hold as well for the various regimes of the non-linear model specifications.

As mentioned above, the methodology used in detecting threshold non-linearity (conditional on the spread) in the present paper will entail using the optimal threshold values obtained from the non-linear least squares procedure. These values will be taken as the hypothetically true regime-delimiting thresholds and subsequently be used in constructing "regime models" for both the upper and lower regime. Then, 1-step ahead predictive errors obtained from observations beyond the thresholds (i.e. sequentially obtained from moving into the central regime) will be computed and subsequently used in constructing F-tests with the end of verifying the orthogonality condition described in section 2.2.3.

Table 5: Threshold non-linearity test

PredErr	10	20	30	40	50	60	70	80	90	100
U.Regime	256.67	97.50	38.09	22.63	8.52	7.63	6.06	3.60	4.36	4.66
(Prob.)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L.Regime	269.98	19.37	15.13	21.37	25.09	23.41	21.01	18.20	19.00	15.30
(Prob.)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5 summarises the F-statistics obtained from evaluating the joint significance of the original set of regressors in explaining the 1-step ahead predictive residuals. Conditional on the RSS-minimising thresholds, $\tau_{1,min}$ and $\tau_{2,min}$, increments of 10 predictive errors, starting from 10 and ending with 100, have been used in the computation of the various F-statistics. Notice that in all of the cases the resulting probabilities allow for the rejection of the null hypothesis of linearity (orthogonal predictive errors). This result appears to be particularly robust for the lower regime, in which the F-statistics never drop below two digits. The upper regime, on the other hand, drops to as low a level as 3.60, which is equivalent to a probability of 0.002 of not rejecting the null of linearity.

Finally, using the methodology described in section 2.3.3, which is based on a third-order Taylor approximation around $\lambda = 0$, an F-statistic will be computed evaluating the statistical significance of the higher-order terms added to the "linear" ECM-specification. The F-statistic obtained is equal to 3.408, which corresponds to a probability insignificantly different from zero of committing a Type1-error of rejecting the null when it is true. Again, this result permits the application of a LSTAR non-linear error-correction model, since the higher-order terms included in the standard ECM-specification do have statistically significant power in explaining variations in the regressand.

After having summarised the evidence speaking in favour of, on the one hand, using error-correction models instead of standard Box-Jenkins-type AR-models, and, on the other hand, using non-linear instead of standard linear model specifications, attention will now turn to estimation and forecast evaluation exercises to be following below.

4.2 Estimation and forecast results

The purpose of the following section is twofold. First of all, coefficient estimates obtained from either ordinary LS or non-linear conditional OLS (i.e. the grid-search method) will be reported for each of the three models under consideration. Furthermore, in order to better compare and contrast the merits of all models, a further standard AR-model (i.e. excluding the error-correction term) has been added as a benchmark specification. Also, as already mentioned above, the lag order for each and every model (including regimes within) has been selected by the Schwartz-Bayesian Information Criterion to be $p_1 = 4$ and $p_2 = 1$ for ΔY_t and ΔX_t, respectively. In addition to that, graphs illustrating the fitted values have been generated, allowing for visual inspection of the differences in "fit" generated by each of the models under consideration.

Secondly, and more importantly, various forecast criteria, which have been discussed in section 3.2, will be reported and compared in order to arrive at an informed decision concerning the relative merit in forecast accuracy of each of the four models. Three forecast horizons have been chosen as benchmark periods, which will be used both individually and jointly for forecast evaluation purposes.

4.2.1 Linear AR and ECM models

Table 6 given below summarises the results obtained from estimating a simple AR-model via ordinary LS. The figures in round brackets are the corresponding t-ratios for each and every coefficient estimate. Notice also that for every one of the following model summaries, the residual variance, denominated by σ^2, will be reported.

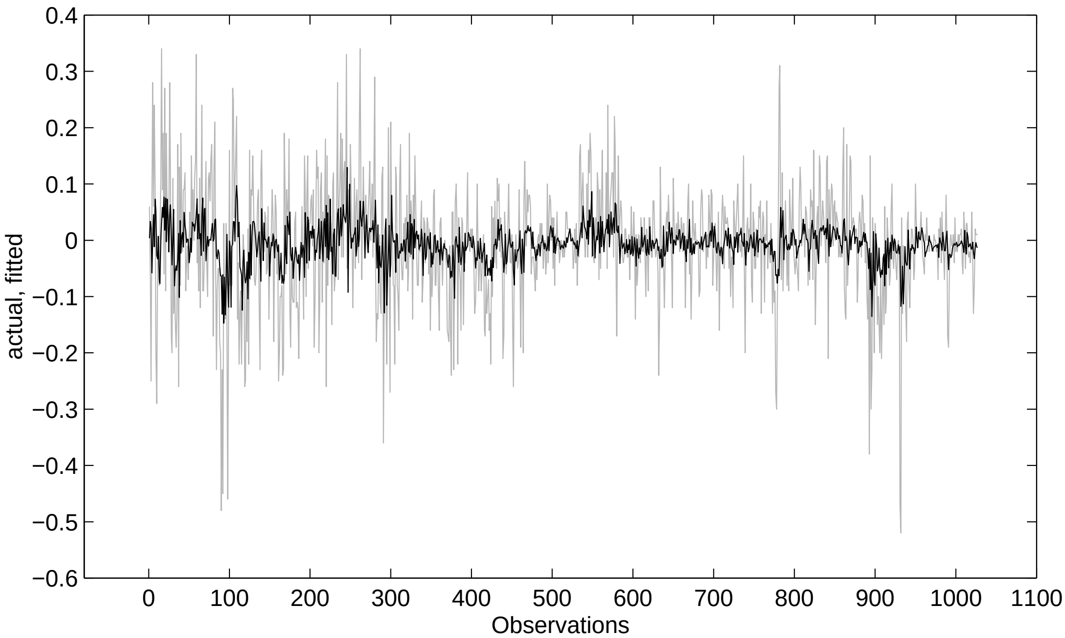
Table 6: Est. Coefficients on AR-model

	Const	ΔY_{t-1}	ΔY_{t-2}	ΔY_{t-3}	ΔY_{t-4}	ΔX_{t-1}	Z_{t-1}
ΔY_t	-0.0042	0.0060	-0.0198	0.0533	0.1216	0.2501	0.0000
	(1.383)	(0.105)	(0.633)	(1.704)	(3.986)	(4.457)	(-)

$\sigma^2 = 0.00948$

Provision of figure 5 allows the reader to visually inspect and compare the actual realisations of the dependent variable with the estimated fitted values obtained from the simple AR-model. It shows how the simple AR-model is somewhat successful in modelling the direction of various characteristic troughs and peaks of the series; it does not, however, do very well in capturing the extremity in amplitude of most of the observations colour-coded in the graph as light-grey.

Figure 5: ΔY and $\widehat{\Delta Y}$ of AR-model



Moving on to the standard "linear" error-correction model, table 7 summarises the coefficient estimates of this ECM-model, which only distinguishes itself from the above simple AR-model by the inclusion of the error-correction term. Noteworthy here are the sign and the statistical significance of the error-correction term, given by Z_{t-1}.

First of all, a positive sign corresponds to an a-priori expectation that if the (de-meaned) spread is positive (i.e. the short-term bill is under-priced relative to the long-term bill), the yield on the short-term bill should rise relative to the long-term bill in order to establish the long-run (non-arbitrage) equilibrium. Secondly, the spread, with a reported t-ratio of 4.445, is statistically significant within this particular linear estimation framework. This represents confirming evidence for the hypothesis of the expectations theory of the term structure of interest rates outlined in section 1.2.

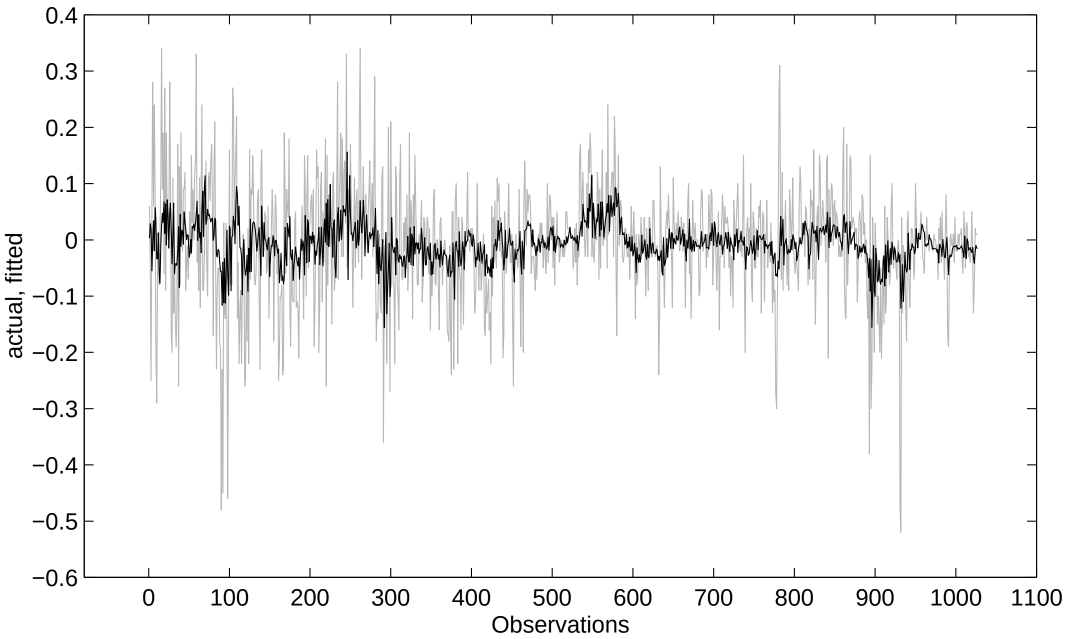
Table 7: Est. Coefficients on Linear ECM-model

	Const	ΔY_{t-1}	ΔY_{t-2}	ΔY_{t-3}	ΔY_{t-4}	ΔX_{t-1}	Z_{t-1}
ΔY_t	-0.0046	0.0391	-0.0286	0.0418	0.1037	0.1984	0.0969
	(1.524)	(0.681)	(0.922)	(1.347)	(3.401)	(3.493)	(4.445)

$\sigma^2 = 0.00930$

Visual discrimination between the AR- and the ECM-graphs is somewhat difficult, as the fitted values of the two models do not seem to differ too much from one another. However, upon closer examination, one could gain the impression, that the ECM-model is more flexible around its mean and that it seems to follow the series slightly better. This impression can also be further strengthened by comparing the σ^2s of the two models, which reveal a slight drop in the variance of the residuals when moving from the simple AR to the ECM model specification.

Figure 6: ΔY and $\widehat{\Delta Y}$ of ECM-model



4.2.2 BAND-TAR model

In this subsection, model estimates of the first non-linear model, the BAND-Tar model, will be reported. Again, the results given below are based on the assumption of an invariant lag-structure for each of the three regimes, with the additional restriction of a zero error-correction term in the central regime. Furthermore, neither of the three regimes has been allowed to contain less than 15% of the total number of observations. This further restriction ensures meaningful regression results for all of the regimes being estimated. Also, the grid-search over τ_1 and τ_2 has been carried out using 200 increments bounded between $\max[Z_{t-1}]$ and $\min[Z_{t-1}]$. The search has been further refined by carrying out an identically incremented search in the neighbourhood (defined as $(\tau_1, \tau_2)_{min} \pm 2 \times incr$) of the first RSS-minimising pair $(\tau_1, \tau_2)_{min}$.

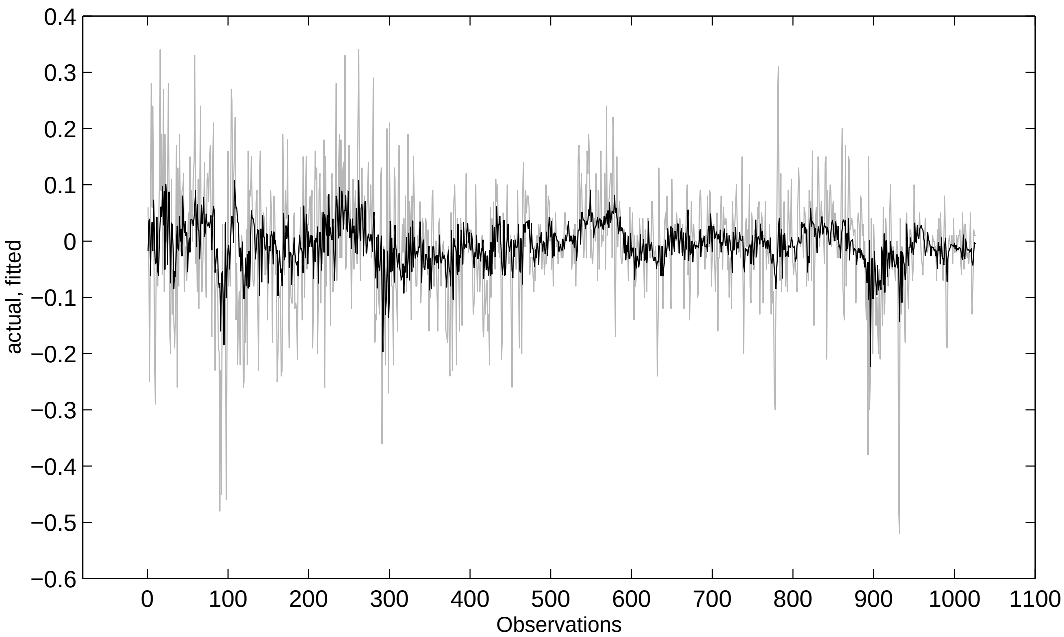
Table 8: Est. Coefficients on BAND-Tar model

	Const	ΔY_{t-1}	ΔY_{t-2}	ΔY_{t-3}	ΔY_{t-4}	ΔX_{t-1}	Z_{t-1}
ΔY_t	0.0027	0.0894	-0.1224	0.0905	0.0231	0.1971	0.1854
(Upper)	(0.406)	(1.042)	(2.797)	(2.067)	(0.535)	(2.252)	(3.386)
ΔY_t	-0.0025	-0.0574	-0.0774	0.0365	0.2107	0.4078	0.0000
(Central)	(0.438)	(0.449)	(1.126)	(0.594)	(3.413)	(3.213)	(-)
ΔY_t	0.0099	-0.0360	0.0952	-0.0020	0.1376	0.1049	0.0390
(Lower)	(0.731)	(0.378)	(1.663)	(0.033)	(2.325)	(1.162)	(0.720)

$\sigma^2 = 0.00903$, $\tau_1 = 0.00312$, $\tau_2 = 0.09314$, %=(54, 27, 19)

Examination of the coefficient estimates obtained from the 3-regime BAND-Tar model allows for an interesting interpretation with particular reference to the individual loadings of the upper and lower regime. First of all, notice that the non-linear least squares procedure has partitioned the data such that more than half of the observations fall into the upper, more than one-quarter fall into the central and just under one-fifth fall into the lower regime. In addition to that, it may be surprising that neither the mean nor the median of the spread is bounded between the two estimated threshold variables, τ_1 and τ_2, but "this might simply reflect an adjustment process whose 'equilibrium' depends on more factors than just the spread between the two bills considered here" (Anderson, 1997). Secondly, the residual variance has again dropped slightly from that of the standard ECM-model estimated above. Thirdly and more interestingly, the estimated loadings for the upper and lower regimes exhibit an economically interpretable asymmetry. While the upper regime's speed of adjustment conditional on the de-meaned spread is positive and statistically significant, the lower regime's speed of adjustment parameter is insignificantly different from zero. This result could arguably be attributed to the often referred-to asymmetric behaviour in the conduct of monetary policy, which postulates that central banks are often more aggressive in loosening monetary policy when faced by an expected economic downturn than tightening it in times of economic overheating. Also, in other empirical work (*see* Anderson, 1997) it has been demonstrated the that short-term bill often adjusts when it is under-priced, but that it is the *long-term* bill that often adjusts when the short-term bill is over-priced. This could also be put forth as an explanation for the encountered asymmetry in the speed of adjustment coefficients.

Figure 7: ΔY and $\widehat{\Delta Y}$ of BAND-Tar model



4.2.3 LSTAR model

Finally, this subsection will report and interpret the coefficient estimates obtained from the LSTAR error-correction model. First of all, notice that just as with the BAND-Tar model the lag-structure has been set to $p_1 = 4$ and $p_2 = 1$ and neither of the two regimes is allowed to contain less than 15% of the total number of observations. Furthermore, the three-dimensional grid-search has been carried out over some pre-chosen trust-regions for λ, δ and c. λ has been searched over the interval $[1; 120]$, δ over $[-35; 35]$ and c over [-2;2]. The trust regions have been chosen such as to maximise the likelihood of the "true" parameters to be enclosed within them. Notice that the trust regions for λ and c must contain the RSS-minimising values by definition, and that the interval for δ has been defined widely enough such as to make values beyond it highly improbable. The increments for λ, δ and c have been set to 200, 200 and 50, respectively and a finer search around the minimising triple $(\lambda, \delta, c)_{min}$ has been carried out using the same specification as was used for the BAND-Tar model.

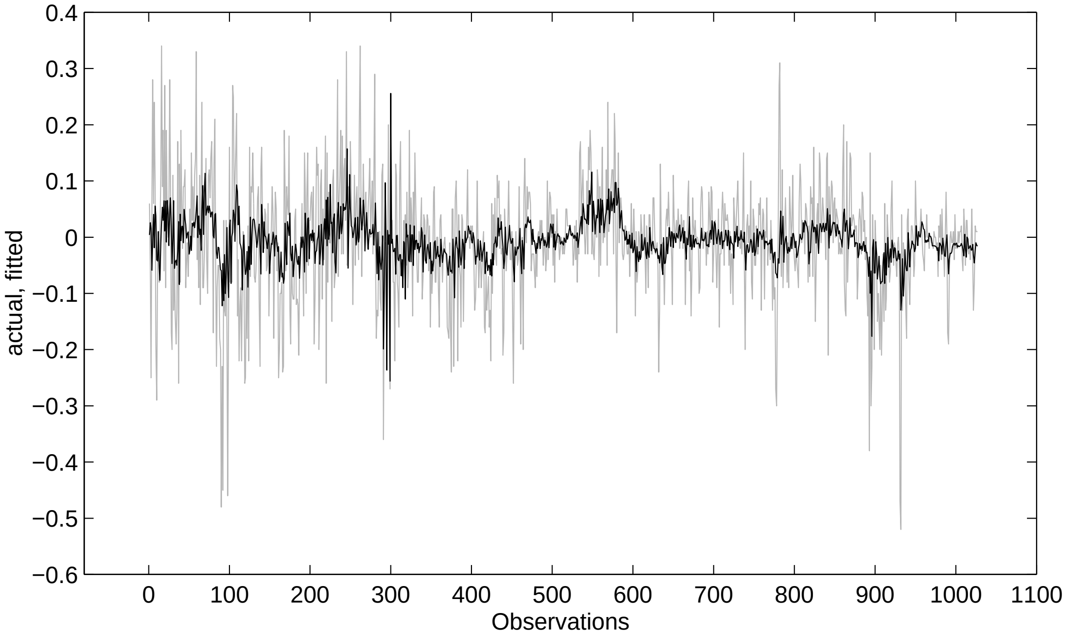
Table 9: Est. Coefficients on LSTAR-model

	Const	ΔY_{t-1}	ΔY_{t-2}	ΔY_{t-3}	ΔY_{t-4}	ΔX_{t-1}	Z_{t-1}
ΔY_t	-0.0734	3.0099	0.1129	-0.0085	0.2565	-2.4931	0.1061
(Outer)	(1.236)	(4.049)	(0.335)	(0.030)	(0.648)	(3.420)	(4.679)
ΔY_t	0.0248	-2.9229	-0.1311	0.0342	-0.1538	2.6548	0.0000
(Central)	(0.432)	(3.970)	(0.393)	(0.123)	(0.393)	(3.676)	(-)

$\sigma^2 = 0.00908$, $\lambda = 95.8668$, $\delta = 4.77400$, $c = -0.31360$, %=$(85, 15)$

First of all, conditional OLS carried out over the grid has partitioned the observations such as to let 15% lie in the central and to let 85% lie in the outer regime modelling the error-correction effect. Secondly, the residual variance σ^2, on the one hand, does exhibit an improvement over the standard ECM model, but on the other hand, has slightly risen compared to the BAND-Tar model. Thirdly, the speed of adjustment coefficient has been estimated at a positively signed 0.1061 with a statistically significant t-ratio of 4.679. Yet again, this accords well with the a-priori expectation of a positive and statistically significant coefficient on the error-correction term, given the expectations theory of the term structure holds.

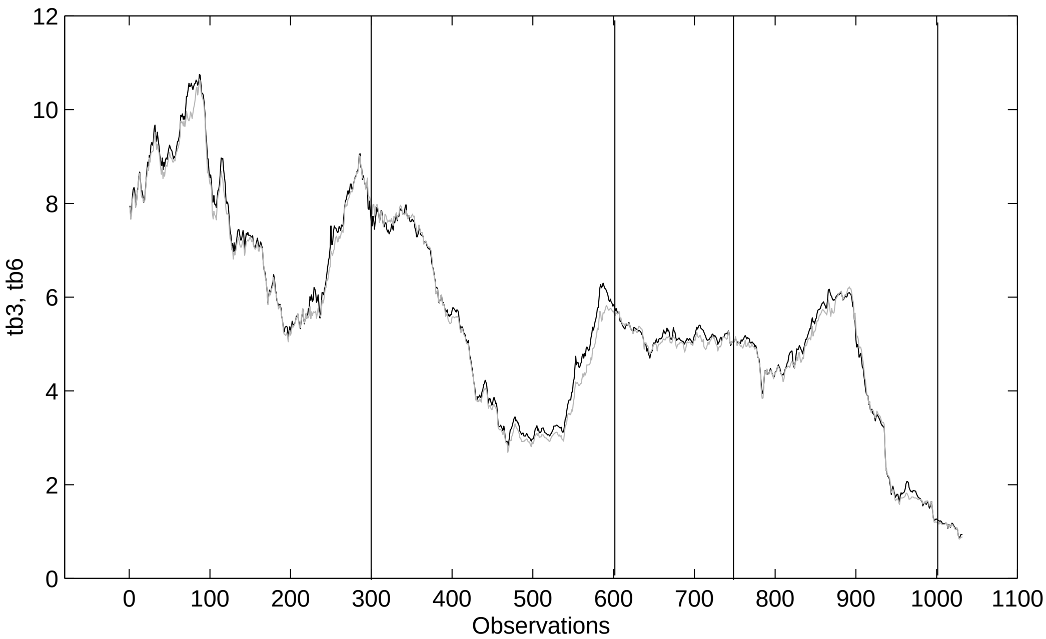
Figure 8: ΔY and $\widehat{\Delta Y}$ of LSTAR-model



Finally, it is interesting to see how the asymmetry parameter δ is estimated at a positively signed 4.774, translating into a relatively stronger adjustment effect for positive values of the spread than for negative values. Although this asymmetry is not very well pronounced at only $\delta = 4.774$, it nevertheless matches quite well the findings on asymmetric responses gathered from the BAND-Tar model.

4.2.4 Forecast comparison

Figure 9: Forecast horizons



In this section 1-step ahead forecast errors for all four models will be computed and then used in the construction of various forecast accuracy criteria. In order to obtain a richer picture, forecast errors will be generated for a variety of different horizons, which are illustrated in the above-given graph. Notice that three different forecast horizons have been marked off such as to define periods ranging from observations 300-600, 600-775 and 775-1000. The reason for choosing these particular forecast horizons is that they exhibit very distinct properties in terms of direction of change and volatility. Loosely speaking, one could refer to the three horizons in sequence as ,,down-and-up”-“tredding-water”-“up-and-down”. The motivation for examining these three horizons in turn is to assess how well or badly all four models forecast over them. One a-priori hypothesis that one might want to test is whether the ECM- and nonlinear-models can ”add much value” to the “tredding-water” period as opposed to the two “rollercoaster-rides”.

Table 10: Summary of forecast criteria (relative) (MSPE, MedAPE, MedSPE)

	300-600			600-775		
	MSPE	MedAPE	MedSPE	MSPE	MedAPE	MedSPE
AR	1.032	1.045	1.093	1.034	1.024	1.048
ECM	1.000	1.000	1.000	1.000	1.000	1.000
TAR	1.075	0.970	0.940	1.027	1.071	1.148
LSTAR	1.024	0.991	0.983	0.987	0.997	0.974

	775-1000			300-775		
	MSPE	MedAPE	MedSPE	MSPE	MedAPE	MedSPE
AR	1.046	1.077	1.160	1.038	1.010	1.060
ECM	1.000	1.000	1.000	1.000	1.000	1.000
TAR	1.005	0.930	0.864	1.037	1.008	0.968
LSTAR	1.129	1.027	1.055	0.998	1.013	0.977

	600-1000			300-1000		
	MSPE	MedAPE	MedSPE	MSPE	MedAPE	MedSPE
AR	1.043	1.051	1.104	1.033	1.030	1.021
ECM	1.000	1.000	1.000	1.000	1.000	1.000
TAR	1.011	0.991	0.982	1.062	0.984	1.015
LSTAR	1.096	0.983	0.967	0.989	1.006	0.995

In addition to that, forecast criteria will also be computed for the first and second (300-775), the second and third (600-1000), and all three horizons (300-1000) jointly. The criteria computed are the MSPE, the MedSPE, the MedAPE, the test-statistics using the loss differential and the test-statistics using the success-ratio.

Table 10 provides a summary of the statistics computed using the prediction errors of the individual models. Since the main purpose of this paper is to compare the "linear" ECM-model with its non-linear counterparts, all statistics have been normalised using the statistics obtained for the ECM-model as the benchmark-case.

First of all, mention should be given to the fact that for each and every forecast horizon, the "linear" ECM-model *always* outperforms the standard AR-model. This result justifies the decision to define the ECM-model as the benchmark-case, against which all other models (and especially the non-linear specifications) should be compared with. Secondly, comparison of the first three forecast horizons reveals, except for the LSTAR model predicting 600-775, that none of the non-linear models exhibit a lower (and therefore relatively "better") MSPE-value than the benchmark ECM-model's normalised value of 1.00. On the other hand though, if judged by the MedSPE and MedAPE, both non-linear models predict slightly better for 300-600, only the LSTAR-model predicts better for 600-775 and only the BAND-Tar model predicts better for 775-1000. Notice that for the "tredding-water"-period of 600-775, the LSTAR-model is capable of "adding some value" in terms of prediction accuracy (judged by all three criteria) to the ECM-model's performance, whereas the BAND-Tar model performs poorly and even worsens the AR-model's record if judged by the MedAPE and the MedSPE. One interesting interpretation for this result could be cast if one was willing to hypothesise that the (de-meaned) spread never becomes too pronounced during the "tredding-water"-period. In that case, the BAND-Tar model would mostly "activate" either one of its three regimes, whereas the LSTAR-model could benefit from its "smooth activation" of the outer regime modelling the adjustment effect for even small deviations of the de-meaned spread from its centrality parameter. It is difficult to explain why the LSTAR-model does less well than the BAND-Tar model (which does quite well if judged against the median forecast criteria) for the 775-1000 period; notice that this particular period contains the monetary expansion episode following the relatively recent downturn in U.S. economic activity.

Finally, it remains to analyse the final three forecast horizons, given by 300-775, 600-1000 and 300-1000. The results for the first and last of these three horizons are rather mixed and inconclusive regarding the forecast accuracy of the non-linear models. This may of course be a result of using only less than 300 observations in an information set to estimate models with the end of forecasting 475 and 700 observations, respectively. Therefore, less attention will be paid to analysing these horizons. The most interesting forecast horizon is probably the 600-1000, since it constitutes a relatively representative (that is, usual) horizon containing both the "tredding-water" and the "up-and-down" episodes. If one were to take this horizon as a typical out-of-sample representation facing the forecaster, then in terms of the MedAPE and MedSPE, both non-linear model specifications provide slightly better forecasts than the "linear" ECM-model.

However, bearing in mind the higher MSPE for both non-linear models over the same period (and for most of the other periods as well), it appears as if the non-linear counterparts to the ordinary ECM-model were better at minimising the median distance of absolute and squared prediction errors from their conditional mean prediction, than the average distance, for which they almost consistently predict worse than the "linear" ECM-model.

Table 11: Summary of forecast criteria (DM- and S-statistics)

	300-600		600-775		775-1000	
	S-stat	DM-stat	S-stat	DM-stat	S-stat	DM-stat
AR vs ECM	2.425	0.115	1.134	0.144	1.400	0.200
ECM vs LSTAR	0.231	-0.049	0.983	0.026	-0.333	-0.057
ECM vs TAR	-1.270	-0.136	0.076	-0.066	0.333	-0.012

	300-775		600-1000		300-1000	
	S-stat	DM-stat	S-stat	DM-stat	S-stat	DM-stat
AR vs ECM	2.615	0.118	1.800	0.171	2.948	0.147
ECM vs LSTAR	0.780	-0.037	0.400	-0.043	0.454	-0.037
ECM vs TAR	-0.964	-0.115	0.300	-0.023	-0.605	-0.074

Table 11 reports the statistics based on the loss-differential between two rivalling models as described in Diebold and Mariano (1995). First of all, the S-statistic comparing the predictive errors of the simple AR-model with those of the "linear" ECM-model evaluates the median loss differential to be statistically significantly different from zero in one-half of the forecast horizons (i.e. for 300-600, 300-775 and 300-1000). Over the chosen "representative" forecast horizon of 600-1000 the S-statistic is equal to 1.8, which just falls short of the 5% critical value of 1.96. The S-statistics for the two other remaining forecast horizons of 600-775 and 775-1000 clearly fail to attain the 5% threshold reporting values of 1.34 and 1.4, respectively. Disregarding statistical insignificance of the S-statistic in one-half of the forecast horizons, it is important to emphasise the fact that the S-statistics, comparing the predictive accuracy of the simple AR- with that of the 'linear" ECM-model, are *all* positively signed, attesting the ECM-model consistently smaller prediction errors (in median!) than those obtained from the AR-model. The DM-statistics, on the other hand, all lie within the area of positive 0.1-0.2, revealing the average loss differential between the AR- and the ECM-model's predictive errors to be insignificantly different from zero. However, yet again sweeping aside statistical insignificance for a moment, they are again consistently positively signed, attesting the ECM-model consistently smaller prediction errors (in mean!) than those obtained from the AR-model.

Moving on to the S-statistics comparing the predictive errors of the non-linear models with those of the "linear" ECM-model, it is clear to see that none of them permit the rejection of the null of predictive equality. Regardless of statistical insignificance, it is noteworthy that the LSTAR-model predicts worse (in median) than the ECM-model only once (for 775-1000), whereas the BAND-Tar model does so three times (for 300-600, 300-775 and 300-1000). Over the chosen "representative" horizon of 600-1000, both non-linear models predict statistically insignificantly better than the "linear" ECM-model (thus the positively signed, but very small S-statistics). It is interesting to see how the non-linear models, in many cases, predict statistically insignificantly better than the 'linear" ECM-model in median (thus the more positively signed, but small S-statistics), but how they unanimously predict worse in mean (hence the consistently negatively signed DM-statistics). This conflicting picture between the S- and the DM-statistics arises because in mean, all sequences of loss differentials may be negative but very close to zero, however it happens to be the case for many of the computed sequences of loss differentials that slightly more elements in $\sum_{j=1}^{m}\left(I[d_j > 0] - \frac{1}{2}\right)$ are positively signed than negatively. To give one example, if one were to compute the loss differential between the ECM- and the LSTAR-model's predictive errors for the 300-1000 horizon, then one would find a negative mean statistically indifferent from zero, but a positive-to-negative ratio of 51:49. For the same horizon, the loss differential between the AR- and ECM-model would have a positive but close to zero mean, but in contrast a 55:45 positive-to-negative ratio, hence the higher S-statistic for this comparison than that for the latter (*see* Figures 12 and 13, section A.2).

Table 12: Summary of forecast criteria (SR- and DA-statistics)

	300-600		600-775		775-1000	
	SR-stat	DA-stat	SR-stat	DA-stat	SR-stat	DA-stat
AR	0.560	0.053	0.474	-0.055	0.604	0.120
ECM	0.563	0.062	0.526	0.005	0.591	0.087
TAR	0.583	0.103	0.491	-0.036	0.587	0.085
LSTAR	0.580	0.089	0.514	-0.016	0.600	0.105

	300-775		600-1000		300-1000	
	SR-stat	DA-stat	SR-stat	DA-stat	SR-stat	DA-stat
AR	0.528	0.013	0.548	0.043	0.553	0.048
ECM	0.549	0.040	0.563	0.051	0.563	0.055
TAR	0.549	0.050	0.545	0.032	0.561	0.061
LSTAR	0.556	0.047	0.563	0.050	0.570	0.066

Finally, Table 13 provides a summary of the direction of change statistics discussed above. The results are sobering, to say the least, if one were to have hoped for obtaining models useful for predicting the change in the 3-month U.S. treasury bill. Notice that since all of the SR-statistics lie in the region of 50%, it is not surprising that none of the corresponding DA-statistics permit rejecting the hypothesis of the success ratio (SR) being insignificantly different from the success ratio in case of independence (SRI). There are, however, a few discernable patterns worth mention. First of all, except for the "tredding-water" period of 600-775, the ECM-model predicts changes always slightly better than the standard AR-model. Furthermore, for the periods of 300-600, 300-775 and 300-1000, both non-linear models predict changes better than their linear counterpart. However, the same chosen "representative" period as above, which was 600-1000, reveals no qualitative improvement in the change-of-direction predictions obtained from the non-linear models over the ordinary ECM-model.

5 Conclusion

All in all, the evidence does not speak much in favour of using the above-described non-linear error-correction models instead of their linear counterpart in modelling the change in the 3-month U.S. treasury bill. Although they manage to bring about somewhat of a forecast improvement in terms of reducing the median-squared prediction error, they seem to be doing so at the cost of increasing the mean-squared prediction error, which should ultimately be used as the main criterion for comparing the respective forecast performances of the various models considered above, since it measures the average (or expected) squared prediction error, and therefore the SPE carrying on average the highest probability of occurring. However, the all important message has to be that, disregarding considerations of good or poor prediction aspects depending on which forecast criteria being applied, the S- and DM-statistics clearly and firmly place the non-linear models into the region of statistical insignificance regarding their difference in prediction to the "linear" ECM-model. This evidence questions the robustness of the results found by Anderson (1997), who concludes her paper with a much more optimistic verdict on the forecast accuracy of non-linear threshold error-correction models. Going even one step further, although the ordinary ECM-model clearly appears to distinguish itself qualitatively from the simple AR-model in terms of MSPE, MedSPE and MedAPE, as illustrated in Table 10, the corresponding S- and DM-statistics however question the overall significance of difference in prediction partially for the null of a zero-median and completely for the null of a zero-mean loss differential. However, bearing in mind the merely asymptotic validity of the DM-statistic, the S-statistic, which has better power in small samples, attests the ECM-model to exhibit some difference in prediction to the simple AR-model specification, and as such recommends it for forecasting the 3-month U.S. treasury bill.

There exist a variety of ways of extending the above forecast exercise. First of all, minimisation of the RSS-function could theoretically be carried out over more parameters than those considered above. The optimisation problem could be increased in complexity by allowing the algorithm to flexibly choose the delay parameter d as well as the lag order of the autoregressive structure for each of the regimes. However, this would no longer be feasible using a simple grid-search, and therefore other optimisation methods would have to be employed.

Secondly, the above exercise could be carried out using a non-linear VAR framework, modelling and forecasting both the 3- and 6-month treasury bills. The most obvious advantage of such an approach would be the possibility of computing *ex-ante* forecasts of the spread, thus permitting computation of h-step ahead forecasts of either one of the variables modelled in the VAR (Clements and Galvao, 2003). Another possibly fruitful extension to this approach, in turn, could then be to consider non-linear threshold as well as non-linear smooth transition vector error-correction models modelling not only two T-bills or government bonds with their respective spread, but rather three or more, including the corresponding possibility of using more than one type of spread.

A Appendices

A.1 Table with absolute forecast criteria

Table 13: Summary of forecast criteria (absolute) (MSPE, MedAPE, MedSPE)

	300-600			600-775		
	MSPE	MedAPE	MedSPE	MSPE	MedAPE	MedSPE
AR	0.00627	0.04588	0.00210	0.00388	0.03469	0.00120
ECM	0.00608	0.04389	0.00193	0.00375	0.03389	0.00115
TAR	0.00653	0.04256	0.00181	0.00385	0.03631	0.00131
LSTAR	0.00622	0.04351	0.00189	0.00374	0.03344	0.00112

	775-1000			300-775		
	MSPE	MedAPE	MedSPE	MSPE	MedAPE	MedSPE
AR	0.00909	0.04490	0.00202	0.00539	0.04035	0.00163
ECM	0.00870	0.04169	0.00174	0.00522	0.03994	0.00160
TAR	0.00874	0.03876	0.00150	0.00554	0.04024	0.00162
LSTAR	0.00982	0.04282	0.00183	0.00531	0.03985	0.00159

	600-1000			300-1000		
	MSPE	MedAPE	MedSPE	MSPE	MedAPE	MedSPE
AR	0.00681	0.04040	0.00163	0.00658	0.04174	0.00174
ECM	0.00653	0.03844	0.00148	0.00634	0.04054	0.00164
TAR	0.00660	0.03810	0.00145	0.00657	0.03989	0.00159
LSTAR	0.00716	0.03781	0.00143	0.00676	0.04008	0.00161

A.2 Distribution histograms of squared prediction errors

Figure 10: Distribution of SPEs for ECM over 600-1000

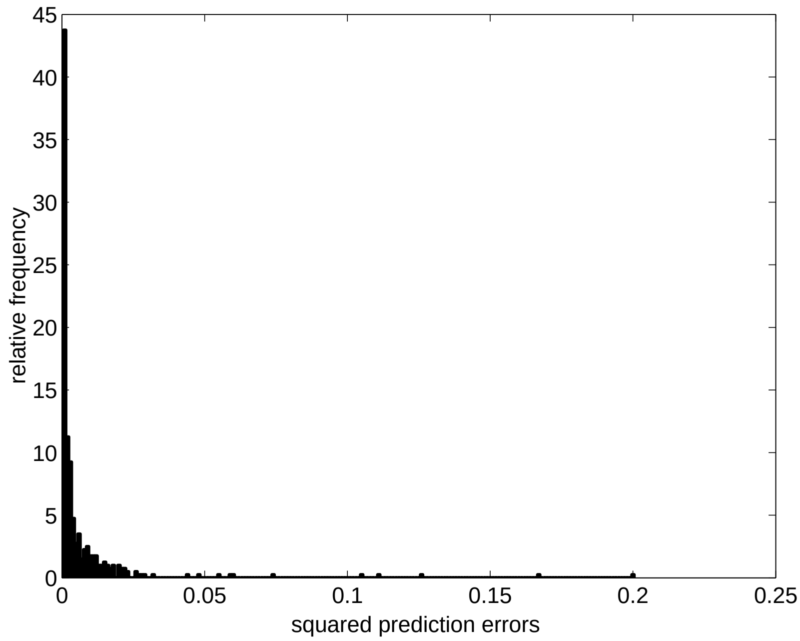
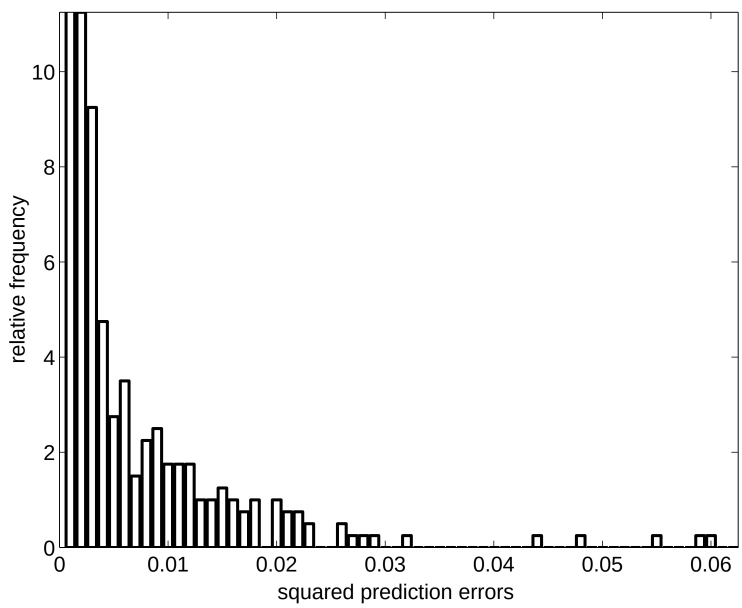


Figure 11: Distribution of SPEs for ECM over 600-1000 (magnified)



A.3 Graphs of loss differentials

Figure 12: Loss differential AR vs ECM (600-100)

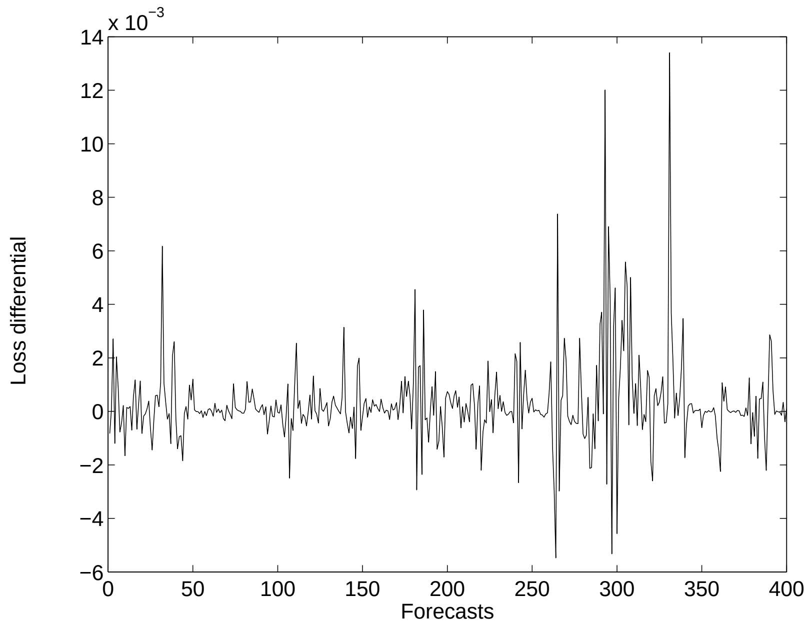
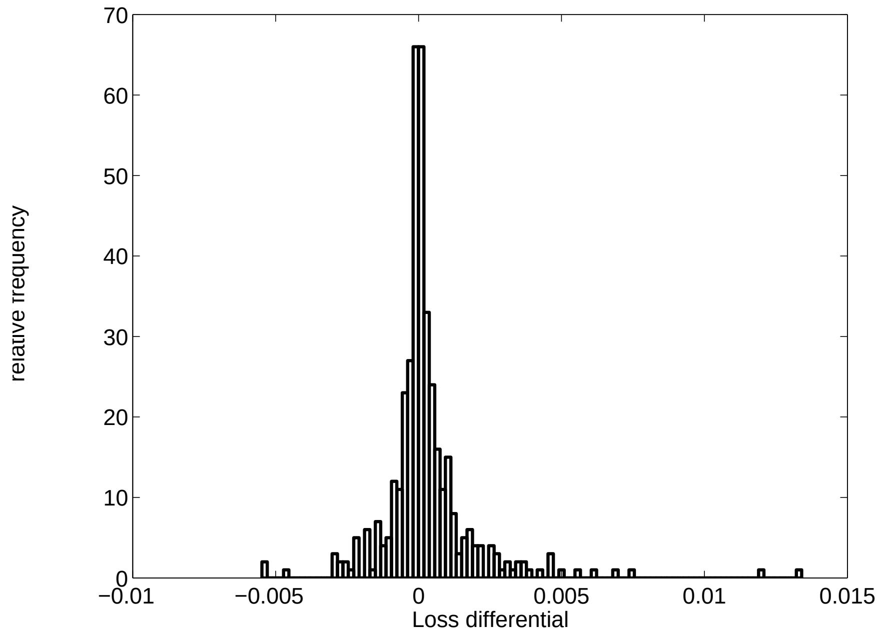


Figure 13: Distribution of loss differential AR vs ECM (600-100)



References

ANDERSON, H. (1997): "Transaction costs and non-linear adjustment towards equilibrium in the U.S. treasury bill market," *Oxford Bulletin of Economics and Statistics*, 59(4), 465–484.

ANDERSON, H., AND F. VAHID (2000): "Predicting the probability of a recession with non-linear autoregressive leading indicator models," Working Paper 3/2000, Monash University, Department of Econometrics and Business Studies.

BALKE, N., AND T. FOMBY (1997): "Threshold cointegration," *Internationl Economic Review*, 38, 627–645.

CHANG, K., AND H. TONG (1986): "On estimating thresholds in autoregressive models," *Journal of Time Series Analysis*, 7, 179–190.

CLEMENTS, M., AND A. GALVAO (2003): "A comparison of non-linear cointegration with an application to the predictability of U.S. interest rates using the term structure," *Review of Economics and Statistics*, Forthcoming.

CLEMENTS, M., AND D. HENDRY (1998): *Forecasting Economic Time Series*. Cambridge University Press, Cambridge.

DIEBOLD, F., AND R. MARIANO (1995): "Comparing predictive accuracy," *Journal of Business and Economic Statistics*, 13, 253–263.

ENDERS, W., AND C. GRANGER (1998): "Unit-root tests and asymmetric adjustment with an example using the term structure of interest rates," *Journal of Business and Economics Statistics*, 19, 304–311.

ENGLE, R., AND C. GRANGER (1987): "Co-integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, 55, 251–276.

ESTRELLA, A., AND F. MISHKIN (1998): "Predicting U.S. recessions: Financial variables as leading indicators," *Review of Economics and Statistics*, 80, 45–61.

FRANSES, P., AND D. VAN DIJK (2000a): *Non-linear time series models in empirical finance*. Cambridge University Press, Cambridge.

——— (2000b): *Nonlinear Econometric Modelling in Time Series: Proceedings of the Eleventh International Symposium in Economic Theory*. Cambridge University Press, Cambridge.

HALL, A., H. ANDERSON, AND C. GRANGER (1992): "A cointegration analysis of treasury bill yields," *The Review of Economics and Statistics*, 74, 116–126.

HAMILTON, J., AND D. KIM (2000): "A re-examination of the predictability of economic activity using the yield spread," Working Paper 7954, National Bureau of Economic Research.

HANSEN, B. (1997): "Inference when a nuisance parameter is not identified under the null hypothesis," *Econometrica*, 64, 413–430.

——— (2000): "Sample splitting and threshold estimation," *Studies in Nonlinear Dynamics and Econometrics*, 2, 1–14.

Johansen, S. (1988): "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control*, 12, 231–254.

Leybourne, S., P. Newbold, and D. Vougas (1998): "Unit roots and smooth transitions," *Journal of Time Series Analysis*, 19, 83–97.

Lo, M., and E. Zivot (2001): "Threshold Cointegration and Nonlinear Adjustment to The Law of One Price," *Macroeconomic Dynamics*, 5(4), 533–576.

Luukkonen, R., P. Saikkonen, and T. Teräsvirta (1988): "Testing linearity against transition autoregressive models," *Biometrika*, 75, 491–499.

Marshall, D. (1994): "Asset return volatility with extremely small costs of consumption adjustment," Working Paper Series, Macroeconomic Issues 94-23, Federal Reserve Bank of Chicago.

Obstfeld, M., and A. Taylor (1997): "Non-linear aspects of goods market aspects arbitrage and adjustment: Heckscher's commodity points revisited," *Journal of the Japanese and International Economies*, 11, 441–479.

O'Connell, P., and S.-J. Wei (1997): "The Bigger They Are, the Harder They Fall: How Price Differences Across US Cities Are Arbitraged," Working Paper 6089, National Bureau of Economic Research.

Pesaran, M., and A. Timmermann (1992): "A simple nonparametric test of predictive performance," *Journal of Business and Economic Statistics*, 10, 461–465.

Ramsey, J. (1969): "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society*, 31, 350–371.

Stock, J., and M. Watson (1999): *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*. Oxford University Press, Oxford.

Teräsvirta, T. (1994): "Specification, estimation, and evaluation of smooth transition autoregressive models," *Journal of American Statistical Association*, 72, 208–218.

Tong, H. (1978): *On a Threshold Model in Pattern Recognition and Signal Processing*. Sijhoff, Noordhoff, Amsterdam.

Tong, H., and K. Lim (1980): "Threshold Autoregression, Limit Cycles and Cyclical Data," *Journal of the Royal Statistical Society*, 42, 245–292.

Trinkler, C., and N. Wolf (2003): "Economic Integration in Interwar Poland - A Threshold Cointegration Analysis of the Law of One Price for Poland 1924-1937," Economics Working Papers eco2003-05, European University Institute.

Tsay, R. (1989): "Testing and Modeling Threshold Autoregressive Processes," *Journal of American Statistical Association*, 84, 231–240.

——— (1998): "Testing and modeling multivariate threshold models," *Journal of American Statistical Association*, 93, 1188–1202.