Lancaster University Management School Working Paper 2009/024

Forecasting the Real Exchange Rate using a Long Span of Data. A Rematch: Linear vs Nonlinear

E Pavlidis, Ivan Paya and David Peel

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

Forecasting the Real Exchange Rate using a Long Span of Data. A Rematch: Linear vs Nonlinear

Efthymios G. Pavlidis Ivan Paya* David A. Peel

Lancaster University Management School, Lancaster, LA1 4YX, UK

Abstract

This paper deals with the nonlinear modeling and forecasting of the dollar-sterling real exchange rate using a long span of data. Our contribution is threefold. First, we provide significant evidence of smooth transition dynamics in the series by employing a battery of recently developed in-sample statistical tests. Second, we investigate the small sample properties of several evaluation measures for comparing recursive forecasts when one of the competing models is nonlinear. Finally, we run a forecasting race for the post-Bretton Woods era between the nonlinear real exchange rate model, the random walk, and the linear autoregressive model. The winner turns out to be the nonlinear model, against the odds.

KEY WORDS: Real Exchange Rate, Nonlinearity, Robust Linearity Tests, Forecast Evaluation, Bootstrapping.

JEL Classification: C22, C53, F31, F37

*Correspondence to: Ivan Paya, Lancaster University Management School, Lancaster, LA1 4YX, UK. E-mail: i.paya@lancaster.ac.uk

1 Introduction

Despite the overwhelming evidence supporting the presence of nonlinearites in real exchange rates (e.g., Taylor et al., 2001; Pavlidis et al., 2009a), the empirical literature on the out-of-sample performance of Smooth Transition Autoregressive (STAR) models is scarce, and a bet that a nonlinear model beats a linear one would be against the odds. One of the few studies on nonlinear real exchange rate forecasting is that of Sarantis (1999). By employing monthly real effective exchange rates for the G-10 countries from 1980 to 1996, the author provides evidence in favor of the presence of significant smooth-transition nonlinear dynamics for the majority of the processes. Moreover, the estimated STAR models provide more accurate forecasts, in terms of the root mean square error criterion, against the Random Walk (RW) and the Markov Switching model but not the linear autoregressive (AR) model.

A recent study that utilizes more sophisticated forecast evaluation techniques and a longer data set for the post-Bretton Wood era is provided by Rapach and Wohar (2006). The authors replicate the results of Obstfeld and Taylor (1997) and Taylor et al. (2001) by fitting Threshold Autoregressive (TAR) and Exponential STAR (ESTAR) models to four monthly U.S. dollar real exchange rates. On the basis of point, interval and density forecasts comparisons Rapach and Wohar (2006, p. 341) conclude: *"any nonlinearities in monthly real exchange rate data from the post-Bretton Woods period are quite "subtle" for Band-TAR and exponential smooth autoregressive model specifications"*.

These discouraging findings may but do not necessarily imply that the nonlinearity documented in the literature is a spurious artifact. Inoue and Kilian (2005) illustrate that for linear models in-sample tests tend to have, and in many cases substantially, higher power than out-of-sample tests, which contradicts the conventional view that forecasting is the ultimate test of an econometric model. Rossi (2005) also raises concerns regarding the power of out-of-sample predictability tests.

Clark and McCracken (2005) build upon the work of Clark and McCracken (2001) and McCracken (2004) and derive the asymptotic distribution of two F-type tests for the comparison of multi-step forecasts from nested linear models. The tests account for parameter uncertainty and exhibit better power properties than their t-type counterparts, namely the tests of Diebold and Mariano (1995) and Harvey et al. (1998). Although their application in this context is appealing, it is not straightforward due to the fact that their derivation is based on the assumption that the regression models are linear in parameters and the processes are stationary.[1]

Regarding the comparison of nonlinear with linear AR models, numerous studies suggest that in many cases the in-sample superiority of the former is not accompanied by better predictive ability (see, e.g., Lundbergh and Teräsvirta, 2002; Stock and Watson, 1999). In this framework, power issues turn out to be serious.[2] A possible explanation is that nonlinear models perform better only in specific states (regime dependent) so that there are windows of opportunity for substantial reduction in prediction errors (Clements, 2005; Boero and Marrocu, 2004). If these occasions are relatively infrequent, then AR models would provide robust forecasts even if the series under consideration is nonlinear. Hence, the results of Sarantis and Rapach and Wohar may well be attributed to the low power of out-of-sample predictability tests.

[1] We relax these assumptions and examine the finite properties of the tests in Section 4.

[2] The related literature has focused mainly on the comparison of SETAR and AR models. The results presented in Section 4 illustrate that this is also the case for STAR models.

In this paper we give the nonlinear real exchange rate model another chance. We depart from the approach of previous studies and employ a long span of annual data for the dollar-sterling real exchange rate. By doing so, we extend the out-of-sample period to the entire post-Bretton Woods era.

Our modeling cycle consists of a battery of recently developed unit root tests, linearity tests, as well as bootstrap methods, which enable us to obtain a parsimonious specification of the nonlinear real exchange rate model. Subsequently, we employ the chosen specification and use Monte Carlo simulation techniques to examine the empirical size and power properties of several forecast accuracy and encompassing tests.

Our results indicate that most tests have good size properties. This is a particularly important finding since the properties of F-type tests have not been examined when one of the competing models is nonlinear or nonstationary. Furthermore, we show that F-type tests have similar or substantially better power properties than their t-type counterparts. Unfortunately, both appear to exhibit low power for the comparison of nonlinear with linear AR models. Notwithstanding the above, our findings suggest that for the actual data the ESTAR model outperforms both the RW and AR benchmarks at short horizons for the majority of tests.

2 Smooth Transition Models

The basic STAR model representation for a univariate time series $\{y_t\}$ is given by

$$\begin{aligned} y_t &= \pi_{1,0} + \pi_{1,1} y_{t-1} + \cdots + \pi_{1,p} y_{t-p} + (\pi_{2,0} + \\ &\quad + \pi_{2,1} y_{t-1} + \cdots + \pi_{2,p} y_{t-p}) F(y_{t-1}; \gamma, c) + \epsilon_t, \qquad t = 1, \ldots, T, \end{aligned} \tag{1}$$

or, equivalently,

$$y_t = \boldsymbol{\pi}_1' \boldsymbol{x}_t + \boldsymbol{\pi}_2' \boldsymbol{x}_t F(y_{t-1}; \gamma, c) + \epsilon_t, \qquad t = 1, \ldots, T, \tag{2}$$

where $\boldsymbol{x}_t = (1, \tilde{\boldsymbol{x}}_t')'$ with $\tilde{\boldsymbol{x}}_t = (y_{t-1}, \ldots, y_{t-p})'$, and $\boldsymbol{\pi}_j = (\pi_{j,0}, \ldots, \pi_{j,p})'$ for $j = 1, 2$. It is assumed that the error term, ϵ_t, is a martingale difference sequence.

There are two common forms of the STAR model. The one we will discuss here in detail is the ESTAR model, in which transitions between a continuum of regimes are assumed to occur smoothly and symmetrically. The transition function $F(\cdot)$ of the ESTAR model is

$$F(y_{t-1}; \gamma, c) = [1 - \exp(-\gamma (y_{t-1} - c)^2)]. \tag{3}$$

This transition function is symmetric around $(y_{t-1} - c)$ and admits the limits 1 and 0 as $|y_{t-1} - c| \to +\infty$ and $|y_{t-1} - c| \to 0$, respectively. Parameter γ can be seen as the transition speed of the function $F(\cdot)$ towards 1 (0) as the deviation grows larger (smaller). We are particularly interested in the special case that there is a unit root in the linear polynomial, $\sum_{i=1}^{p} \pi_{1,i} = 1$, $\pi_{2,i} = -\pi_{1,i}$ $\forall\ i \geq 1$, $\pi_{1,0} = 0$ and $c = \pi_{2,0}$. Under these restrictions, Equation (1) becomes

$$y_t = \pi_{2,0} + [\pi_{1,1}(y_{t-1} - \pi_{2,0}) + \cdots + \pi_{1,p}(y_{t-p} - \pi_{2,0})] \exp(-\gamma(y_{t-1} - \pi_{2,0})^2) + \epsilon_t. \quad (4)$$

The above formulation is very appealing for modeling real exchange rates (see, e.g., Kilian and Taylor, 2003; Paya et al., 2003). Unlike in a linear model, the process moves between a white noise and a unit root depending on the size of the deviation from PPP, $|y_{t-1} - \pi_{2,0}|$. This type of adjustment is in accordance with the implications of theoretical models, which demonstate how frictions in international trade can induce nonlinear but mean reverting adjustment of the real exchange rate (see, e.g., Dumas, 1992; Berka, 2005). The rational is that small deviations are left uncorrected since they do not to cover transactions costs or the sunk costs of international arbitrage. On the other hand, large deviations are much less persistent. Therefore, the process exhibits strong persistence and near unit root behavior.[3]

We point out that as $\gamma \to 0$ or $\gamma \to \infty$ the exponential transition function approaches a constant and the ESTAR model collapses to a linear AR model. The fact that STAR models nest linear AR models has important implications regarding the asymptotic distribution of commonly used forecast accuracy and encompassing tests (see, e.g., Clements and Galvão, 2004).

[3]The other common form of STAR models is the Logistic, LSTAR. The logistic function is

$$F(y_{t-1}; \gamma, c) = [1 + \exp(-\gamma(y_{t-1} - c))]^{-1}.$$

LSTAR models have also been fitted to real exchange rates (see Sarantis, 1999). Even though the theoretical argument is not as strongly supported as with the case of the ESTAR, there are some attempts to rationalize the asymmetric adjustment in the real exchange rate (e.g., Campa and Goldberg, 2002).

2.1 Linearity and Unit Root Tests

The uncertainty about the exact Data Generating Process (DGP) of real exchange rates motivates the use of data driven methods for the specification of parsimonious empirical models. In this study, we employ several testing procedures so as to examine whether the long-span real exchange rate series exhibits mean reversion and smooth transition dynamics. Namely, we use the unit root tests of Kapetanios et al. (2003) and Kapetanios and Shin (2008), and the linearity tests of Escribano and Jordá (1999) and Harvey and Leybourne (2007). A description of all in-sample tests is provided in Appendix A (see also Pavlidis et al., 2009a).

The appealing feature of the Harvey and Leybourne (2007) test is that it possesses the same properties irrespective of the series being I(0) or I(1).[4] While, the advantages of the linearity test of Escribano and Jordá (1999) are: (i) it enables the selection between ESTAR and LSTAR models, and (ii) it can be easily modified to accommodate for possibly conditional heteroskedastic errors by applying the Wild Bootstrap method (see Pavlidis et al., 2009b).[5] Due to the fact that Escribano and Jordá (1999) test is based on the assumption of a stationary process, pretesting for a unit root is required.

Unit root testing is also useful for the selection of forecasting models. Diebold and Kilian (2000) illustrate that the conventional view of employing models in first-differences when the series under examination is highly persistent can lead to less accurate forecasts. On this ground, the authors advocate the application of unit root tests for choosing between levels and differences.

[4] A major concern in the PPP literature is that real exchange rates exhibits a unit root in which case the asymptotic distribution of most linearity tests changes (Kiliç, 2004).

[5] Time-varying volatility may arise due to changes in exchange rate or monetary regime.

We consider the null hypothesis of a unit root against a globally stationary ESTAR process by using the tests of Kapetanios et al. (2003) and Kapetanios and Shin (2008). The main difference between the two is that the Kapetanios et al. (2003) test uses OLS demeaning/detrending procedures, whilst, the Kapetanios and Shin (2008) test, in the spirit of Elliott et al. (1996), employs GLS demeaning/detrending procedures. Further, we robustify the tests against heteroskedasticity of unknown form by using Heteroskedasticity Consistent Covariance Matrix Estimators. Cook (2006) illustrates that in small samples this practice can lead to moderate oversizing of the Augmented Dickey Fuller (ADF) and the Kapetanios et al. (2003) tests. Pavlidis et al. (2007) draw a similar conclusion for the test of Kapetanios and Shin (2008). We address this issue by constructing exact sample critical values for the heteroskedasticity-robust test statistics via stochastic simulations.

3 Evaluating Forecasts

In this study, we restrict our attention to the comparison of point forecasts on the basis of forecast accuracy and forecast encompassing measures. The former measures include the MSE-t of Diebold and Mariano (1995), the MSE-F of Clark and McCracken (2005) and the Weighted MSE-t (W-MSE-t) proposed by van Dijk and Franses (2003). The latter are the ENC-t of Harvey et al. (1998) and the ENC-F of Clark and McCracken (2005).

Our setting is similar to the one adopted by Clark and McCracken (2005). The number of in-sample and out-of-sample observations is denoted as R and P, respectively, so that the total number of observations is $T = R + P$. We adopt a recursive scheme for forecasting, where as t increases from R to $T - h$ the parameters of the models are re-estimated by employing data up to time t so as to generate forecasts for the following h horizons.[6] In accordance with the notation used in the previous section, y_{t+h} denotes the variable to be predicted at time $t = R, \ldots, T-h$ with the number of forecasts corresponding to horizon h being equal to $P-h+1$. The forecast errors are defined as $\hat{e}_{1,t+h} = y_{t+h} - \hat{y}_{1,t+h|t}$ for the benchmark model and $\hat{e}_{2,t+h} = y_{t+h} - \hat{y}_{2,t+h|t}$ for the competing model.

3.1 Tests of Forecast Accuracy

The first three tests examine forecast accuracy by setting the Mean Square Error (MSE) as the measure of predictive ability. In this setting, the null hypothesis is that the MSEs of the two competing models are equal against the one-sided alternative that the MSE for the second model is smaller. Diebold and Mariano (1995) develop the following widely used t-type test

$$\mathrm{MSE} - t = (P-h+1)^{1/2} \frac{\bar{d}}{\widehat{S}_{dd}^{1/2}}, \tag{5}$$

where $\widehat{d}_{t+h} = \widehat{e}^2_{1,t+h} - \widehat{e}^2_{2,t+h}$, $\bar{d} = (P-h+1)^{-1} \sum_{t=R}^{T-h} \widehat{d}_{t+h} = \mathrm{MSE}_1 - \mathrm{MSE}_2$, $\widehat{\Gamma}_{dd}(j) = (P-h+1)^{-1} \sum_{t=R+j}^{T-h} \widehat{d}_{t+h}\widehat{d}_{t+h-j}$ for $j \geqslant 0$ and $\widehat{\Gamma}_{dd}(j) = \widehat{\Gamma}_{dd}(-j)$, and $\widehat{S}_{dd} = \sum_{j=-\bar{j}}^{\bar{j}} K(j/M)\widehat{\Gamma}_{dd}(j)$ denotes the long-run variance of d_{t+h} estimated using a kernel-based estimator with function $K(\cdot)$, bandwith parameter M and maximum number of lags $\bar{j}$.[7]

[6] In general, closed-form solutions for multi-step forecasts from nonlinear models are not available. We overcome this obstacle by employing bootstrap integration techniques. A discussion regarding methods for constructing multi-step forecasts from nonlinear models is provided in Teräsvirta (2006). An attractive feature of the bootstrap method is that it does not require distributional assumptions. The errors, however, are presumed to be i.i.d.. The results of Clements and Smith (1997) support the use of bootstrap methods in forecasting from nonlinear autoregressive models.

[7] The use of Heteroskedasticity and Autocorrelation Consistent (HAC) estimators for computing the variance of d_{t+h} is based on the fact that h-steps-ahead forecast errors will be serially correlated of order $h-1$.

For non-nested models the long-run variance of $\hat{d}_{t+h}$ is positive and the MSE-t statistic follows asymptotically the standard normal distribution. On the contrary, when the competing models are nested their population errors are identical under the null and, therefore, d_{t+h} and its variance are equal to zero. In this case, the asymptotic distribution of the statistic is non-standard and depends upon nuisance parameters for $h \geq 2$ (McCracken, 2004).[8]

The degeneracy of the long-run variance of d_{t+h} motivates Clark and McCracken (2005) to propose a variant of the above test for nested models. Inspired by the in-sample F-test, the authors suggest replacing $\widehat{S}_{dd}^{1/2}$ with the variance of the forecast error of the "unrestricted" model. The new test statistic is given by

$$\text{MSE} - F = (P - h + 1)^{1/2} \frac{\bar{d}}{\text{MSE}_2}, \tag{6}$$

and has better power properties.[9] The limiting distribution of the MSE-F test statistic, like the MSE-t, is free of nuisance parameters only for $h = 1$ and is non-standard.

The forecast accuracy tests examined so far attach equal importance to all forecasts irrespectively of the available information set at time t. However, a researcher would expect the superiority of the ESTAR model over the RW to become most apparent for large deviations of the process from its equilibrium value. Whilst, for smaller deviations the two models should perform similarly. van Dijk and Franses (2003) propose a forecast

The performance of the MSE-t test using different HAC estimators is examined in Clark (1999).

[8]The asymptotic distributions of all the test statistics for multi-step forecasts from nested models under parameter uncertainty are derived in Clark and McCracken (2005). However, their derivation is based on the sufficient but not necessary assumptions of stationarity and linearity of the parameters, which are clearly not satisfied in our experiment.

[9]Clark and McCracken (2005) and Busetti et al. (2009) provide Monte Carlo evidence illustrating the power advantage of F-type tests over t-type when the models are linear.

evaluation test ($\mathrm{W}-\mathrm{MSE}-t$) that employs a weighted average loss differential and comprises a modification of the MSE-t of Diebold and Mariano (1995). Consequently, more importance is attached to forecasts corresponding to deviations at the tails of the distribution. van Dijk and Franses (2003) show that the modified test statistic follows the same distribution with the MSE-t.

3.2 Forecast Encompassing

In this case, the null hypothesis is that the forecast of the benchmark model incorporates all the relevant information in the forecast of the competing model. Or, equivalently, the covariance between the forecast errors of the first model and the difference of the forecasts errors of the two models is equal to zero (see West, 2006). Under the alternative, the covariance is positive indicating that the second model has additional predictive power. Clearly, the forecast encompassing tests are also one-sided to the right.

Let $\widehat{c}_{t+h}=\widehat{e}_{1,t+h}(\widehat{e}_{1,t+h}-\widehat{e}_{2,t+h})$, $\bar{c}=(P-h+1)^{-1}\sum_{t=R}^{T-h}\widehat{c}_{t+h}$, $\widehat{\Gamma}_{cc}(j)=(P-h+1)^{-1}\sum_{t=R+j}^{T-h}\widehat{c}_{t+h}\widehat{c}_{t+h-j}$ for $j\geqslant 0$, $\widehat{\Gamma}_{cc}(j)=\widehat{\Gamma}_{cc}(-j)$, and let $\widehat{S}_{cc}=\sum_{j=-\bar{j}}^{\bar{j}}K(j/M)\widehat{\Gamma}_{cc}(j)$ denote the long-run variance of c_{t+h}. Harvey et al. (1998), based on the work of of Diebold and Mariano (1995), derive the following forecast-encompassing test statistic[10]

$$\mathrm{ENC}-t=(P-h+1)^{1/2}\frac{\bar{c}}{\hat{S}_{cc}^{1/2}}. \tag{7}$$

Clark and McCracken (2001) illustrate that the distribution of the ENC-t statistic converges to the same type of distribution with the MSE-t statistic when the forecasts are generated from linear nested models. By employing the same reasoning with the one used

[10]The authors employ the small sample correction of Harvey et al. (1997) for the MSE-t statistic.

for the MSE-F test they propose the following F-type test statistic

$$\mathrm{ENC}-F=(P-h+1)^{1/2}\frac{\bar{c}}{\mathrm{MSE}_2}. \tag{8}$$

which again has a non-standard limiting distribution and depends on nuisance parameters for $h \geq 2$. Similarly to forecast accuracy measures, the F-type test has, asymptotically, greater power than its t-type counterpart.

3.3 Bootstrap Inference

Due to the fact that standard distribution theory may not apply in our setting, we conduct statistical inference by employing a parametric bootstrap method similar to Kilian (1999) and Kilian and Taylor (2003). The simulation exercise consists of the following steps.

1. Employ the original real exchange rate series and compute the above forecast evaluation measures for all forecast horizons.

2. Estimate the restricted model for the real exchange rate (the RW or the AR model) using the whole sample, and obtain the fitted residuals and coefficients.

3. Set the estimated model as the Null DGP and randomly draw with replacement from the residuals so as to create an artificial series for the real exchange rate with the same length as the actual series. The process is initialized by employing the observed values of the series.

4. Repeat the forecasting exercise using the artificial data so as to compute h bootstrap test statistics for each forecast evaluation measure.

5. Repeat steps 3 and 4 B times, where B is a large number, so as to obtain the bootstrap distributions of the test statistic under the null.

6. Compute the bootstrap p-value as the percentage of times the simulated statistic is more extreme than the original statistic.

7. Reject the null if the p-value is smaller than the chosen significance level.

Clark and McCracken (2005) illustrate that when forecasts are generated from linear nested models this method performs adequately in terms of size and power even when the bootstrap model is not properly specified. However, the performance of the bootstrap technique and the validity of the F-type tests have not been explored when one of the competing models is nonlinear or the process is nonstationary. We contribute to the literature on nonlinear real exchange rates and forecasting evaluation by examining the finite properties of F-type tests as well as their implications in the following section.

4 Empirical Results

The data set consists of annual observations for the dollar-sterling real exchange rate from 1791 to 2005. For the construction of the series we use the International Financial Statistics database to update the nominal exchange rate and price series analyzed in Lothian and Taylor (1996). The number of in-sample observations, R, is set equal to 183, which corresponds to the pre-Bretton Woods era (1791-1973), and the remaining 32 years, P, comprise the out-of-sample period.

4.1 In-Sample Tests

Starting with the in-sample tests, we present results for both the entire sample period and the subperiod from 1791 to 1973. Table 1 reports the ADF, t_{NL} and $t_{\mathrm{NL}}^{\mathrm{GLS}}$ tests statistics as well as their heteroskedasticity-robust versions, ADF-HC, t_{NL}-HC, $t_{\mathrm{NL}}^{\mathrm{GLS}}$-HC, corresponding to the demean and detrend cases (for a description of the in-sample tests see Appendix A).[11] For the demeaned real exchange rate, the unit root hypothesis is rejected by all tests at the 5% significance level. The only exception is the test proposed by Kapetanios and Shin (2008), which rejects the null at the 10% when data prior to the recent floating period are used. Turning to the detrend case, we observe a small decrease in number of rejections. Specifically, the t_{NL}-HC statistic for the subperiod 1791-1973, and the $t_{\mathrm{NL}}^{\mathrm{GLS}}$ and $t_{\mathrm{NL}}^{\mathrm{GLS}}$-HC tests statistics for the whole period are larger than the corresponding 10% critical values. Overall, the results presented in Table 1 suggest the rejection of the unit root hypothesis in favor of both a linear and a nonlinear stationary process.

[Table 1]

The finding of mean reverting behavior of the long-span real exchange rate is consistent with the empirical literature on PPP (see Frankel, 1990; Lothian and Taylor, 1996). Further, given the stationarity of the series, we follow the recommendation of Diebold and Kilian (2000) and choose to work with levels rather than first differences.

We proceed by examining the presence of STAR-type nonlinearities by applying the Escribano and Jordá (1999) and Harvey and Leybourne (2007) testing procedures. The results are reported in Table 2. First, the wild bootstrap p-values for the Escribano and Jordá

[11] The lag length for the unit root and linearity tests is set to two on the basis of the Akaike Information Criterion.

(1999) tests (top panel) corresponding to the null of linearity is marginally lower than the 5% significance level for the whole sample and slightly higher than the 10% for the sub-sample. Second, the test favors the use of the ESTAR model over the asymmetric LSTAR. The Harvey and Leybourne (2007) test statistic is also greater than the 10% critical value which provides further support for the smooth transition model. The magnitudes of the p-values corresponding to the linearity tests indicate that the nonlinear mean-reverting behavior of the series is more evident for the whole sample period than the pre-Bretton Woods era. This finding can be attributed to the higher power of the tests for larger sample sizes.

[Table 2]

Next, we follow Kilian and Taylor (2003) and model the level of the real exchange rate using the ESTAR parameterization (4).[12] Table 3 shows the estimates of the ESTAR model for the two periods examined, the standard error of the regressions, the corresponding t-statistics, the Ljung-Box Q-statistics for serial correlation in the residuals and the LM test statistic (ARCH) for conditional heteroskedasticity up to lags 1 and 5, and the wild bootstrap p-value for the transition parameter $\widehat{\gamma}$. The Q and ARCH statistics do not indicate the presence of serial correlation or ARCH effects in the regression residuals. Moreover, the p-value is virtually zero in both cases suggesting that the estimated transition parameters are significant at all conventional levels. In line with the linearity tests results, the p-value for the transition parameter is lower for the whole sample illustrating that the degree of nonlinearity is more pronounced when longer spans of data are examined.

[12] Equation (4) imposes that the autoregressive coefficients sum to unity so that the process has a unit root in the inner regime. We test this restriction by running a Wald F-test. The corresponding p-value is substantially larger than 10% implying that the restricted version is also supported by the data.

[Table 3]

4.2 Out-of-Sample Tests

The in-sample test results provide strong support for a nonlinear adjustment mechanism of the real exchange rate. We now turn to the investigation of the performance of the ESTAR model in forecasting. As we highlighted in the previous sections: (i) out-of-sample tests are likely to exhibit lower power than in-sample tests, and (ii) there is uncertainty regarding the behavior of F-type tests when one of the competing model is nonlinear or nonstationary. These motivate us to examine the small sample properties of the forecast evaluation measures by conducting a set of Monte Carlo simulation experiments. The nominal significance level is set equal to 5% for all experiments, the maximum forecast horizon equal to 4 and the number of bootstrap replication, B, equal to 1,000.

4.2.1 Empirical Size of Forecast Evaluation Tests

Initially, we focus on the empirical size of the tests, which is computed by the following procedure

1. Fit the benchmark model (the RW or the linear AR) to the whole sample.

2. Set the the fitted model as the Null DGP and generate 1,000 artificial series of size equal to the size of the actual real exchange rate series.[13]

3. For each series adopt the same setting as for the actual data and generate forecasts from the benchmark and the competing model(s).

[13] Fake series are generated by drawing from the normal distribution with variance equal to variance of the actual residuals. The first observations of the actual data are employed as initial values.

4. Apply the bootstrap methodology outlined in Section 3 so as to compute a vector of bootstrap p-values.

5. The empirical size of the test is defined as the percentage of times the bootstrap p-value is smaller than the 5 % significance level.

The results for the case of the RW against the ESTAR (RW-ESTAR), the RW against the AR (RW-AR) and the AR against the ESTAR (AR-ESTAR) are presented in Table 4. A broad conclusion that emerges is that the empirical size of all tests, but the W-MSE-t, is close to the nominal level with no test consistently outperforming the others. The (absolute) error in rejection probabilities reaches a maximum of just 1.7 percentage points (for the MSE-F at the one year horizon). Most importantly, these results indicate that F-type tests are valid in our nonlinear context.

[Table 4]

As far as the W-MSE-t is concerned, the test exhibits moderate size distortions of up to 5 percentage points. For the RW-ESTAR and the RW-AR cases the test is oversized at short horizons with the empirical size taking values close to 10%. On the other hand, for the AR-ESTAR case the weighted MSE-t statistic becomes undersized with the empirical size reaching a minimum value equal to 0.023 at $h = 2$.

4.2.2 Empirical Power of Forecast Evaluation Tests

We turn to the empirical power of the tests. The procedure adopted is identical to that for the size with the exception that the DGP is given by the estimated ESTAR model. Table 5 shows the results for the RW-ESTAR and AR-ESTAR cases. Overall, we observe that despite the fact that there are major differences across tests and pairs of competing models, the empirical power of all tests tends to decrease with the forecast horizon. Starting with the RW-ESTAR, t-type tests perform substantially worse than F-type tests. Specifically, the MSE-t ranks last with the empirical power ranging from about 15% for $h = 1$ to about 8% for $h = 4$. The W-MSE-t and ENC-t tests follow with the latter being marginally superior than the former but again with very low empirical power.[14] An increase by a factor of two or greater (depending on the horizon) in the frequency of rejecting the null occurs as we move to the MSE-F. The empirical power of the test exceeds 50%. Finally, the ENC-F test exhibits the highest power, which ranges from 68 to about 75%.

[Table 5]

Regarding the AR-ESTAR pair, the performance of the F-type tests deteriorates while t-type tests exhibit similar empirical power to the RW-ESTAR case. The maximum power, which is achieved at $h = 1$ in all cases, ranges from about 16 (W-MSE-t) to about 27% (MSE-F). In other words, there is a small likelihood of identifying the forecasting gains from adopting an ESTAR rather than a linear AR model even though the true DGP process is nonlinear.These results are qualitatively similar to those of Clements and Smith (1999) for SETAR models, and complement the findings of Inoue and Kilian (2005) for linear models. The low power of the tests suggests that superior in-sample but not out-of-sample performance of nonlinear models should not be documented as conclusive evidence against nonlinearity.

[14] The results for the W-MSE-t test should be interpreted with caution due to the poor size properties of the test.

4.2.3 Forecasting the Dollar-Sterling Real Exchange Rate

Table 6 presents the results regarding the comparison of forecasts for the actual real exchange rate series. The first three panels report t-type test statistics, while the last two panels show the F-type tests statistics. The corresponding bootstrap p-values are reported in parentheses.

[Table 6]

A broad conclusion that emerges is that as the forecast horizon increases the p-values for all tests tend to increase indicating that long-horizon predictability depends upon short-horizon predictability. This observation is consistent with the behavior of the empirical power of the tests reported in Table 5. Furthermore, the forecasting gains from using our nonlinear model specification are particularly evident at short forecast horizons.[15] To this end, we mainly focus on one step ahead forecasts.

By examining the RW-ESTAR pair (second column), we observe that all five forecast encompassing and forecast accuracy test statistics are statistically significant at the 10% significance level. By changing the significance level to 5%, the null hypothesis is rejected by the two F-type tests and the MSE-t test (three out of the five cases). We note that for the F-type tests, p-values are close to zero for all forecast horizons, which is not true for the t-type tests. The fact that F-type tests are associated with much lower p-values than their t-type counterparts when the benchmark model is the RW is not surprising given the higher empirical power of the former. Turning to the RW-AR pair (third column), we generally observe higher p-values than for the RW-ESTAR pair. However, the number of

[15] This result is also intuitive given that both the ESTAR and AR models are mean reverting processes, hence the series are expected to approach their conditional mean when projected further ahead in the horizon.

rejections at the 10% level reduces marginally from five to four for $h = 1$.[16] Summarizing the above results, both AR and ESTAR models appear to have predictive ability regarding the behavior of the dollar-sterling real exchange rate.

The final column of Table 6 presents the results for the comparison of the AR-ESTAR models. Despite the low empirical power of the forecast evaluation measures, at $h = 1$ all test statistics are significant at the 5% with the exception of the ENC-F, which has a p-value marginally higher than 10%. The number of rejections substantially reduces with the forecast horizon and at $h = 2$ only the MSE-t test rejects the null hypothesis. This may be due to the fact that both models share the prediction that the series will eventually mean revert to its equilibrium value.

Overall, the out-of-sample results complement those of the in-sample tests and provide strong support for the ESTAR model. In contrast to previous studies, which employ data of higher frequency, our findings illustrate that nonlinear real exchange rate models are useful for forecasting the behavior of the real exchange rate.

5 Conclusion

This paper utilizes long-spans of data in order to investigate the ability of the ESTAR model to forecast the dollar-sterling real exchange rate. We pay special attention to model specification by employing several recently proposed linearity and unit root tests as well as bootstrap techniques. In turn, we investigate the small sample properties of several forecast evaluation measures. Our results, in line with the literature on forecasting from

[16]Lothian and Taylor (1996) and Siddique and Sweeney (1998) also show that AR models provide superior forecasts (in terms of the RMSE criterion) to the RW for the recent float.

nonlinear models, illustrate the difficulty of detecting the superiority of STAR models to AR models. Despite the low power of out-of-sample evaluation tests, we find that recursive ESTAR forecasts for the actual real exchange rate series outperform all rival forecasts. Consequently, researchers and practitioners can extract forecasting gains regarding the behavior of the long-span real exchange rate series by employing nonlinear models.

A Appendix

A.1 Linearity Tests

Testing for the nonlinear part of Equation (2) gives rise to a nuisance parameter problem (Davies, 1977). Consequently, classical Lagrange Multiplier (LM) and Wald statistics may not follow standard distributions. In order to circumvent this problem, Luukkonen et al. (1988) suggest replacing the transition function by a Taylor series approximation around $\gamma = 0$. Escribano and Jordá (1999) build upon the work of Luukkonen et al. (1988) and propose the following auxiliary regression

$$y_t = \boldsymbol{\delta}_0' \boldsymbol{x}_t + \boldsymbol{\delta}_1' \boldsymbol{x}_t y_{t-1} + \boldsymbol{\delta}_2' \boldsymbol{x}_t y_{t-1}^2 + \boldsymbol{\delta}_3' \boldsymbol{x}_t y_{t-1}^3 + \boldsymbol{\delta}_4' \boldsymbol{x}_t y_{t-1}^4 + u_t \tag{9}$$

for testing linearity and distinguishing between ESTAR and LSTAR processes. The null hypothesis of linearity corresponds to $H_0^1 : \boldsymbol{\delta}_1' = \boldsymbol{\delta}_2' = \boldsymbol{\delta}_3' = \boldsymbol{\delta}_4' = 0$ and the selection procedure between ESTAR and LSTAR is

1. Test the null of LSTAR nonlinearity, $H_0^L : \boldsymbol{\delta}_2' = \boldsymbol{\delta}_4' = 0$, with an F test, (F_L).

2. Test the null of ESTAR nonlinearity, $H_0^E : \boldsymbol{\delta}_1' = \boldsymbol{\delta}_3' = 0$, with an F test, (F_E).

3. If the p-value of F_L is lower than F_E then select an ESTAR. Otherwise, select an LSTAR.

The use of the F-test is based on the assumptions that the process under examination is stationary and the error term in Equation (2) is i.i.d. However, a major concern in the PPP literature is that real exchange rates exhibits a unit root in which case the asymptotic distribution of linearity tests changes (Kiliç, 2004). Therefore, in order to avoid false inference one should first test for a unit root in the real exchange rate series. If the unit root hypothesis is rejected, the i.i.d. assumption can be relaxed by employing the wild bootstrap method (see Pavlidis et al., 2009b).

Harvey and Leybourne (2007) derive a more general linearity test statistic which has the same critical values under the null hypotheses of a linear I(0) and a linear I(1) processes. Rejection of the null therefore is indicative of nonlinearity and cannot be attributed to a linear I(1) DGP.

The Harvey and Leybourne test procedure consists of two steps. First is the test of linearity. Second, the order of integration of the linear or nonlinear process is determined. Consider the case of an I(0) process. By setting $p = 1$ and taking a second-order Taylor series expansion of Equation (1) around $\gamma = 0$ we obtain

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-1}^2 + \beta_3 y_{t-1}^3 + u_t. \tag{10}$$

Whilst, in the case of an I(1) variable, the Taylor expansion yields

$$\Delta y_t = \varphi_0 \Delta y_{t-1} + \varphi_1 (\Delta y_{t-1})^2 + \varphi_1 (\Delta y_{t-1})^3 + \varepsilon_t.$$

In order to combine both possibilities, I(0) and I(1), Harvey and Leybourne (2007) propose the following regression model

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-1}^2 + \alpha_3 y_{t-1}^3 + \alpha_4 \Delta y_{t-1} + \alpha_5 (\Delta y_{t-1})^2 + +\alpha_6 (\Delta y_{t-1})^3 + \eta_t. \tag{11}$$

In the presence of serial correlation, Equation (11) is augmented with lags of the first difference of the dependent variable. The null hypothesis of linearity is $H_0 : \alpha_2 = \alpha_3 = \alpha_5 = \alpha_6 = 0$ against the alternative hypothesis (nonlinearity) H_1 : at least one of $\alpha_2, \alpha_3, \alpha_5, \alpha_6$ is different from zero. The corresponding Wald statistic is

$$W_T = \frac{RSS_1 - RSS_0}{RSS_0/T},$$

where the restricted residual sum of squares (RSS_1) comes from an OLS regression of y_t on a constant, y_{t-1}, and Δy_{t-1}. As Harvey and Leybourne point out, the distribution of W_T under the null differs depending on whether the process followed by y_t is I(0) or I(1). In order to make the limiting distribution of W_T homogeneous under the null, they multiply it with a correction that is the exponential of a weighted inverse of the absolute value of the Augmented Dickey Fuller (ADF) statistic,[17]

$$W_T^* = \exp(-b\,|ADF_T|^{-1}) W_T. \tag{12}$$

[17]This approach is suggested by Vogelsang (1998).

An expression for the value of b is provided such that, for a given significance level, the critical value of W_T^* coincides with that from a $\chi^2(4)$. They also prove that, under H_1, W_T^* is consistent at the rate $O_p(T)$. The second step is to test whether the series is an I(0) or an I(1) process.

A.2 Unit Root Tests

Kapetanios et al. (2003) develop a test of a unit root null against the alternative of a globally stationary ESTAR. Their test is also based on a Taylor approximation of the nonlinear autoregressive model. For simplicity, assuming $p = 1$, $d = 1$, $\pi_{1,1} = 1$, $\pi_{2,1} = -\pi_{1,1}$, and $c = 0$, then (1) becomes

$$y_t = y_{t-1} + \left[1 - \exp\left(-\gamma y_{t-1}^2\right)\right](-y_{t-1}) + u_t. \tag{13}$$

Using the first-order Taylor expansion and rearranging yields

$$\Delta y_t = \delta y_{t-1}^3 + u_t. \tag{14}$$

Hence, the null and alternative hypotheses are $H_0 : \delta = 0$ and $H_1 : \delta < 0$, respectively. The corresponding t-statistic is given by

$$t_{\text{NL}} = \frac{\hat{\delta}}{\text{s.e.}(\hat{\delta})}, \tag{15}$$

where s.e.$(\hat{\delta})$ denotes the standard error of $\hat{\delta}$. The asymptotic distribution of t_{NL} converges weakly to a functional of Brownian motions.

The issue of possible residual autocorrelation can be addressed by augmenting Equation (14) with lags of the dependent variable. Further, in the presence of deterministic components, the authors suggest replacing y_t in Equation (14) with the residuals from the regression of $\boldsymbol{y}$ on an intercept (demean case) or an intercept and a time trend (detrend case).

Kapetanios and Shin (2008) proceed in the spirit of Elliott et al. (1996) by employing a GLS procedure in order to increase the power of the nonlinear unit root test. In the case of a mean and a time trend in the data, the first step of the testing procedure includes computing the GLS estimate of $\boldsymbol{\theta}$ in

$$y_t = \tilde{\boldsymbol{\theta}}' \boldsymbol{z}_t + \tilde{y}_t, \tag{16}$$

by regressing $\boldsymbol{y}_{\bar{\rho}} = (y_1, y_2 - \bar{\rho} y_1, \ldots, y_T - \bar{\rho} y_{T-1})'$ on $\boldsymbol{z}_{\bar{\rho}} = (\boldsymbol{z}_1, \boldsymbol{z}_2 - \bar{\rho}\boldsymbol{z}_1, \ldots, \boldsymbol{z}_T - \bar{\rho}\boldsymbol{z}_{T-1})'$ where $\boldsymbol{z}_t = (1, t)'$ and $\bar{\rho} = 1 - \bar{c}/T$ so as to obtain the estimated residuals, $\tilde{y}_t$.[18] For the demean case $\boldsymbol{z}_t$ is replaced by $z_t = 1$. Subsequently, Equation (14) is fitted to the GLS demeaned or detrended series and the t-statistic, $t_{\text{NL}}^{\text{GLS}}$, corresponding to $H_0 : \delta = 0$ is obtained. Kapetanios and Shin (2008) illustrate that the $t_{\text{NL}}^{\text{GLS}}$ statistic, like the t_{NL}, has a non-standard distribution.

[18] Kapetanios and Shin (2008) set $\bar{c}$ equal to -17.5 so that the asymptotic power of the test under the local alternative is 0.5.

References

Berka, Martin, "General Equilibrium Model of Arbitrage Trade and Real Exchange Rate Persistence," MPRA Paper 234, University Library of Munich, Germany 2005.

Boero, Gianna and Emanuela Marrocu, "The performance of SETAR models: a regime conditional evaluation of point, interval and density forecasts," *International Journal of Forecasting*, 2004, *20* (2), 305–320.

Busetti, Fabio, Juri Marcucci, and Giovanni Veronese, "Comparing Forecast Accuracy: A Monte Carlo Investigation," Working Paper, Bank of Italy 2009.

Campa, Jose Manuel and Linda S. Goldberg, "Exchange Rate Pass-Through into Import Prices: A Macro or Micro Phenomenon?," NBER Working Papers 8934, National Bureau of Economic Research, Inc 2002.

Clark, Todd E., "Finite-sample properties of tests for equal forecast accuracy," *Journal of Forecasting*, 1999, *18* (7), 489–450.

_ **and Michael W. McCracken**, "Tests of equal forecast accuracy and encompassing for nested models," *Journal of Econometrics*, 2001, *105* (1), 85–110.

_ **and** _ , "Evaluating Direct Multistep Forecasts," *Econometric Reviews*, 2005, *24* (4), 369–404.

Clements, Michael P., *Evaluating Econometric Forecasts of Economic and Financial Variables*, New York: Palgrave Macmillan, 2005.

_ **and Ana B. Galvão**, "A comparison of tests of nonlinear cointegration with application to the predictability of US interest rates using the term structure," *International Journal of Forecasting*, 2004, *20* (2), 219–236.

_ **and Jeremy Smith**, "The performance of alternative forecasting methods for SETAR models," *International Journal of Forecasting*, 1997, *13* (4), 463–475.

_ **and** _ , "A Monte Carlo Study of the Forecasting Performance of Empirical SETAR Models," *Journal of Applied Econometrics*, 1999, *14* (2), 123–41.

Davies, Robert B., "Hypothesis testing when a nuisance parameter is present only under the alternative," *Biometrika*, 1977, *64*, 179–190.

Diebold, Francis X. and Lutz Kilian, "Unit-Root Tests Are Useful for Selecting Forecasting Models," *Journal of Business & Economic Statistics*, 2000, *18* (3), 265–73.

_ **and Roberto S. Mariano**, "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 1995, *13* (3), 253–63.

Dumas, Bernard, "Dynamic Equilibrium and the Real Exchange Rate in a Spatially Separated World," *Review of Financial Studies*, 1992, *5* (2), 153–80.

Elliott, Graham, Thomas J. Rothenberg, and James H. Stock, "Efficient Tests for an Autoregressive Unit Root," *Econometrica*, 1996, *64* (4), 813–36.

Escribano, Alvaro and Oscar Jordá, "Improved testing and specification of smooth transition regression models," in P. Rothman, ed., *Nonlinear Time Series Analysis of Economic and Financial Data*, Dordrecht: Kluwer Academic Publishers, 1999.

Frankel, Jeffrey A., "Zen and the Art of Modern Macroeconomics: A Commentary," in William S. Haraf and Thomas D. Willett, eds., *Monetary Policy for a Volatile Global Economy*, Washington, D.C.: AEI Press, 1990, pp. 117–23.

Harvey, David I. and Stephen J. Leybourne, "Testing for Time Series Linearity," *Econometrics Journal*, 2007, *10* (1), 149–165.

_ , _ , and Paul Newbold, "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 1997, *13* (2), 281–291.

_ , _ , and _ , "Tests for Forecast Encompassing," *Journal of Business & Economic Statistics*, 1998, *16* (2), 254–59.

Inoue, Atsushi and Lutz Kilian, "In-Sample or Out-of-Sample Tests of Predictability: Which One Should We Use?," *Econometric Reviews*, 2005, *23* (4), 371–402.

Kapetanios, George and Yongcheol Shin, "GLS detrending-based unit root tests in nonlinear STAR and SETAR models," *Economics Letters*, 2008, *100* (3), 377–380.

_ , _ , and Andy Snell, "Testing for a Unit Root in the Nonlinear STAR Framework," *Journal of Econometrics*, 2003, *112* (2), 359–379.

Kilian, Lutz, "Exchange Rates and Monetary Fundamentals: What Do We Learn from Long-Horizon Regressions?," *Journal of Applied Econometrics*, 1999, *14* (5), 491–510.

_ and Mark P. Taylor, "Why Is It So Difficult to Beat the Random Walk Forecast of Exchange Rates?," *Journal of International Economics*, 2003, *60* (1), 85–107.

Kiliç, Rehim, "Linearity tests and stationarity," *Econometrics Journal*, 2004, *7* (1), 55–62.

Lothian, James R. and Mark P. Taylor, "Real Exchange Rate Behavior: The Recent Float from the Perspective of the Past Two Centuries," *Journal of Political Economy*, 1996, *104* (3), 488–509.

Lundbergh, Stefan and Timo Teräsvirta, "Forecasting with Smooth Transition Autoregressive Models," in Michael P. Clements and David F. Hendry, eds., *A companion to economic forecasting*, Oxford: Blackwell, 2002.

Luukkonen, Ritva, Pentti Saikkonen, and Timo Teräsvirta, "Testing linearity against smooth transition autoregressive models," *Biometrika*, 1988, *75*, 491499.

McCracken, Michael W., "Asymptotics for Out-of-Sample Tests of Causality," Manuscript, University of Missouri 2004.

Obstfeld, Maurice and Alan M. Taylor, "Nonlinear Aspects of Goods-Market Arbitrage and Adjustment: Heckscher's Commodity Points Revisited," *Journal of the Japanese and International Economies*, 1997, *11* (4), 441–479.

Pavlidis, Efthymios G, Ivan Paya, and David A Peel, "Linearity Testing in the Presence of Heteroskedasticity," in "ESRC Seminar Series: Nonlinear Economics and Finance Research Community" Brunel 2007.

_ , _ , and _ , "The Econometrics of Exchange Rates," in Terence Mills and Kenneth Patterson, eds., *Palgrave Handbook of Econometrics*, Vol. 2, London: Palgrave, 2009. Part VII.

_ , _ , and _ , “Specifying Smooth Transition Regression Models in the Presence of Conditional Heteroskedasticity of Unknown Form,” Working Papers 005913, Lancaster University Management School, Economics Department 2009.

Paya, Ivan, Ioannis A. Venetis, and David A. Peel, “Further Evidence on PPP Adjustment Speeds: the Case of Effective Real Exchange Rates and the EMS,” *Oxford Bulletin of Economics and Statistics*, 2003, *65* (4), 421–437.

Rapach, David E. and Mark E. Wohar, “The out-of-sample forecasting performance of nonlinear models of real exchange rate behavior,” *International Journal of Forecasting*, 2006, *22* (2), 341–361.

Rossi, Barbara, “Testing Long-Horizon Predictive Ability With High Persistence, And The Meese-Rogoff Puzzle,” *International Economic Review*, 2005, *46* (1), 61–92.

Sarantis, Nicholas, “Modeling non-linearities in real effective exchange rates,” *Journal of International Money and Finance*, 1999, *18* (1), 27–45.

Siddique, Akhtar and Richard J. Sweeney, “Forecasting real exchange rates1,” *Journal of International Money and Finance*, 1998, *17* (1), 63–70.

Stock, James H. and Mark W. Watson, “A Comparison of Linear and Nonlinear Univariate Models for Forecasting Macroeconomic Time Series,” in Robert F. Engle and Halbert White, eds., *Cointegration, Causality and Forecasting: A Festschrift in Honour of Clive Granger*, Oxford: Oxford University Press, 1999.

Taylor, Mark P., David A. Peel, and Lucio Sarno, "Nonlinear Mean-Reversion in Real Exchange Rates: Toward a Solution to the Purchasing Power Parity Puzzles," *International Economic Review*, 2001, *42* (4), 1015–1042.

Teräsvirta, Timo, "Forecasting Economic Variables with Nonlinear Models," in Graham Elliott, Clive W. J. Granger, and Allan Timmermann, eds., *Handbook of Economic Forecasting*, Vol. 1, London: North Holland, 2006.

van Dijk, Dick and Philip Hans Franses, "Selecting a Nonlinear Time Series Model using Weighted Tests of Equal Forecast Accuracy," *Oxford Bulletin of Economics and Statistics*, 2003, *65*, 727–743.

Vogelsang, Timothy J., "Trend function hypothesis testing in the presence of serial correlation," *Econometrica*, 1998, *66* (1), 123–48.

West, Kenneth D., "Forecast Evaluation," in Graham Elliott, Clive W. J. Granger, and Allan Timmermann, eds., *Handbook of Economic Forecasting*, Vol. 1, London: North Holland, 2006.

Table 1: Unit Root Tests

Sample Period: 1791-1973						
Case	ADF	ADF-HC	t_{NL}	t_{NL}-HC	$t_{\text{NL}}^{\text{GLS}}$	$t_{\text{NL}}^{\text{GLS}}$-HC
Demean	-3.082^{**}	-3.321^{**}	-3.488^{**}	-4.687^{***}	-2.211^{*}	-2.866^{**}
Detrend	-4.985^{***}	-5.192^{***}	-3.707^{**}	-3.469	-3.824^{***}	-3.648^{**}
Sample Period: 1791-2005						
Case	ADF	ADF-HC	t_{NL}	t_{NL}-HC	$t_{\text{NL}}^{\text{GLS}}$	$t_{\text{NL}}^{\text{GLS}}$-HC
Demean	-3.794^{***}	-3.991^{**}	-4.522^{***}	-5.314^{***}	-2.873^{**}	-3.258^{**}
Detrend	-4.327^{***}	-4.532^{***}	-4.406^{***}	-5.013^{***}	-2.293	-2.598

Notes: ADF, t_{NL} and t_{NL} are the Augmented Dickey Fuller, the Kapetanios et al. (2003) and the Kapetanios and Shin (2008) unit root tests statistics. HC indicates heteroskedasticity-robust versions. $***$, $**$, $*$ denote significance at the 1%, 5%, and 10% significance level, respectively. Critical values are constructed via Monte Carlo simulations.

Table 2: Linearity Tests

Escribano and Jordá (1999)			
Period	F	F_L	F_E
1791-1973	1.192 (0.114)	0.458 (0.610)	0.368 (0.695)
1791-2005	1.582 (0.043)	1.050 (0.244)	0.886 (0.329)
Harvey and Leybourne (2007)			
Period	W_T^*		
1791-1973	$8.494\ (0.078)$		
1791-2005	$10.478\ (0.033)$		

Notes: p-values are reported in parentheses. For the Escribano and Jordá (1999) test p-values are obtained through the wild bootstrap procedure described in Pavlidis et al. (2009b).

Table 3: Estimated Nonlinear Real Exchange Rate Model

Sample Period: 1791-1973

$$\hat{y}_t - \underset{(63.598)}{1.586} = (\underset{(13.834)}{1.122}(y_{t-1} - \underset{(63.598)}{1.586}) + (1 - \underset{(13.834)}{1.122})(y_{t-2} - \underset{(63.598)}{1.586}))$$
$$\times \exp(-\underset{\substack{(3.508)\\ [0.005]}}{2.076}(y_{t-1} - \underset{(63.598)}{1.586})^2).$$

$s = 0.067$; $Q_1 = 0.005$ $[0.942]$; $Q_5 = 3.941$ $[0.558]$; $\text{ARCH}_1 = 0.059$ $[0.809]$; $\text{ARCH}_5 = 0.220$ $[0.953]$.

Sample Period: 1791-2005

$$\hat{y}_t - \underset{(81.518)}{1.590} = (\underset{(16.053)}{1.185}(y_{t-1} - \underset{(81.518)}{1.590}) + (1 - \underset{(16.053)}{1.185})(y_{t-2} - \underset{(81.518)}{1.590}))$$
$$\times \exp(-\underset{\substack{(4.357)\\ [0.000]}}{2.504}(y_{t-1} - \underset{(81.518)}{1.590})^2).$$

$s = 0.068$; $Q_1 = 0.002$ $[0.963]$; $Q_5 = 4.133$ $[0.530]$; $\text{ARCH}_1 = 0.079$ $[0.778]$; $\text{ARCH}_5 = 0.416$ $[0.837]$.

Notes: Figures in parentheses and square brackets denote absolute t-statistics and p-values, respectively. The p-value for the transition parameter $\hat{\gamma}$ is obtained through a simulation exercise, where the bootstrap DGP is the unit root model. s is the standard error of the regression. Q_1 and Q_5 denote the Ljung-Box Q-statistic for serial correlation up to order 1 and 5, respectively. ARCH_1 and ARCH_5 denote the LM test statistic for conditional heteroskedasticity up to order 1 and 5, respectively.

Table 4: Empirical Size of Forecast Evaluation Tests

RW-ESTAR					
Horizon	MSE-t	W-MSE-t	ENC-t	MSE-F	ENC-F
1	0.056	0.089	0.061	0.058	0.058
2	0.058	0.079	0.053	0.056	0.048
3	0.054	0.072	0.056	0.055	0.047
4	0.038	0.056	0.039	0.058	0.045

RW-AR					
Horizon	MSE-t	W-MSE-t	ENC-t	MSE-F	ENC-F
1	0.055	0.104	0.055	0.052	0.051
2	0.046	0.087	0.044	0.045	0.042
3	0.046	0.071	0.040	0.051	0.044
4	0.041	0.063	0.041	0.053	0.041
AR-ESTAR					
Horizon	MSE-t	W-MSE-t	ENC-t	MSE-F	ENC-F
1	0.043	0.034	0.040	0.067	0.066
2	0.046	0.023	0.047	0.050	0.055
3	0.049	0.032	0.051	0.056	0.054
4	0.052	0.033	0.052	0.059	0.047

Notes: The table shows the empirical size of the MSE-t, W-MSE-t, ENC-t, MSE-F and ENC-F test statistics for the RW-ESTAR, RW-AR and AR-ESTAR pairs. The nominal significance level is 5% and the horizons considered are $h = 1, \ldots, 4$.

Table 5: Empirical Power of Forecast Evaluation Tests

RW-ESTAR					
Horizon	MSE-t	W-MSE-t	ENC-t	MSE-F	ENC-F
1	0.152	0.170	0.239	0.577	0.752
2	0.094	0.124	0.133	0.588	0.730
3	0.078	0.096	0.111	0.566	0.709
4	0.079	0.095	0.114	0.528	0.680
AR-ESTAR					
Horizon	MSE-t	W-MSE-t	ENC-t	MSE-F	ENC-F
1	0.237	0.163	0.203	0.269	0.220
2	0.209	0.121	0.170	0.176	0.124
3	0.163	0.103	0.142	0.106	0.064
4	0.137	0.071	0.108	0.060	0.025

Notes: The table shows the empirical power of the MSE-t, W-MSE-t, ENC-t, MSE-F and ENC-F test statistics for the RW-ESTAR and AR-ESTAR pairs. The nominal significance level is 5% and the horizons considered are $h = 1, \ldots, 4$.

Table 6: Comparing Forecasts for the Dollar-Sterling Real Exchange Rate, 1974-2005

Panel A — MSE-t test			
Horizon	RW-ESTAR	RW-AR	AR-ESTAR
1	1.827 (0.047)	1.441 (0.091)	1.702 (0.019)
2	1.790 (0.067)	1.514 (0.129)	1.281 (0.048)
3	1.664 (0.114)	1.600 (0.134)	0.836 (0.098)
4	1.670 (0.118)	1.702 (0.123)	0.357 (0.203)
Panel B — W-MSE-t test			
Horizon	RW-ESTAR	RW-AR	AR-ESTAR
1	1.718 (0.053)	1.547 (0.066)	1.354 (0.032)
2	1.682 (0.069)	1.654 (0.083)	1.146 (0.062)
3	1.593 (0.095)	1.647 (0.099)	0.815 (0.121)
4	1.617 (0.097)	1.695 (0.104)	0.528 (0.190)
Panel C — ENC-t test			
Horizon	RW-ESTAR	RW-AR	AR-ESTAR
1	2.016 (0.066)	1.794 (0.105)	1.942 (0.033)
2	1.979 (0.105)	1.929 (0.134)	1.607 (0.062)
3	1.942 (0.143)	2.077 (0.131)	1.157 (0.146)
4	2.054 (0.145)	2.276 (0.130)	0.641 (0.295)
Panel D — MSE-F test			
Horizon	RW-ESTAR	RW-AR	AR-ESTAR
1	17.842 (0.000)	11.715 (0.002)	5.369 (0.046)
2	24.793 (0.000)	17.884 (0.008)	5.651 (0.108)
3	29.577 (0.002)	24.756 (0.016)	3.657 (0.181)
4	37.289 (0.007)	34.970 (0.017)	1.583 (0.254)
Panel E — ENC-F test			
Horizon	RW-ESTAR	RW-AR	AR-ESTAR
1	22.449 (0.001)	15.616 (0.003)	6.361 (0.104)
2	30.060 (0.002)	23.754 (0.013)	6.934 (0.181)
3	36.531 (0.013)	32.791 (0.019)	4.729 (0.302)
4	47.439 (0.020)	46.750 (0.023)	2.528 (0.393)

Notes: The table shows the MSE-t, W-MSE-t, ENC-t, MSE-F and ENC-F evaluation measures for the comparison of actual real exchange rate forecasts from the ESTAR, AR and RW models. Bootstrap p-values are reported in parentheses. The horizons considered are $h = 1, \ldots, 4$.