A Predictive Comparison of Some Simple Long Memory and Short Memory Models of Daily U.S. Stock Returns, With Emphasis on Business Cycle Effects*

Geetesh Bhardwaj and Norman R. Swanson
Rutgers University

Current Version May 2005

Abstract

This chapter builds on previous work by Bhardwaj and Swanson (2004) who address the notion that many fractional *I(d)* processes may fall into the "empty box" category, as discussed in Granger (1999). However, rather than focusing primarily on linear models, as do Bhardwaj and Swanson, we analyze the business cycle effects on the forecasting performance of these ARFIMA, AR, MA, ARMA, GARCH, and STAR models. This is done via examination of ex ante forecasting evidence based on an updated version of the absolute returns series examined by Ding, Granger and Engle (1993); and via the use of Diebold and Mariano (1995) and Clark and McCracken (2001) predictive accuracy tests. Results are presented for a variety of forecast horizons and for recursive and rolling estimation schemes. We find that the business cycle does not seem to have an effect on the relative forecasting performance of ARFIMA models.

JEL classification: C15; C22; C53
Keywords: fractional integration; long memory; parameter estimation error; stock returns; long horizon prediction

----

*Geetesh Bhardwaj, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, bhardwaj@econ.rutgers.edu. Norman R. Swanson, Department of Economics, Rutgers University, 75 Hamilton Street, New Brunswick, NJ 08901, USA, nswanson@econ.rutgers.edu. The authors are grateful to the editors, Valentina Corradi and Clive W.J. Granger for stimulating discussions, and Zhuanxin Ding, Steve Leybourne, and Mark Watson for providing the financial and macroeconomic datasets used in the empirical section of the paper. Swanson has benefited from the support of Rutgers University in the form of a Research Council grant.

1 Introduction

As previously discussed in Bhardwaj and Swanson (2004), the last two decades of macro and financial economic research has resulted in a vast array of important contributions in the area of long memory modelling, both from a theoretical and an empirical perspective. From a theoretical perspective, much effort has focussed on issues of testing and estimation, and a very few important contributions include Granger (1980), Granger and Joyeux (1980), Hosking (1981), Geweke and Porter-Hudak (1983), Lo (1991), Sowell (1992a,b), Ding, Granger and Engle (1993), Cheung and Diebold (1994), Robinson (1995), Engle and Smith (1999), Diebold and Inoue (2001), Breitung and Hassler (2002), and Dittman and Granger (2002). The empirical analysis of long memory models has seen equally impressive treatment, including studies by Diebold and Rudebusch (1989, 1991a,b), Hassler and Wolters (1995), Hyung and Franses (2001), Bos, Franses and Ooms (2002), Chio and Zivot (2002), Bhansali and Kokoszka (2002), and van Dijk, Franses and Paap (2002), to name but a few.[1] The impressive array of papers on the subject is perhaps not surprising, given that long memory models in economics is one of the many important areas of research that has stemmed from seminal contributions made by Clive W.J. Granger (see e.g. Granger (1980) and Granger and Joyeux (1980)). Indeed, in the write-up disseminated by the Royal Swedish Academy of Sciences upon announcement that Clive W.J. Granger and Robert F. Engle had won the 2003 Nobel Prize in Economics, it was stated that:[2]

> *Granger has left his mark in a number of areas. [other than in the development of the concept of cointegration] His development of a testable definition of causality (Granger (1969)) has spawned a vast literature. He has also contributed to the theory of so-called long-memory models that have become popular in the econometric literature (Granger and Joyeux (1980)). Furthermore, Granger was among the first to consider the use of spectral analysis (Granger and Hatanaka (1964)) as well as nonlinear models (Granger and Andersen (1978)) in research on economic time series.*

Overall, there has been relatively little evidence in the literature supporting the usefulness of long memory models for prediction. In a discussion of this and related issues, for example, Granger (1999) acknowledges the importance of outliers, breaks, and undesirable distributional properties in the context of long memory models, and concludes that there is a good case to be made for $I(d)$ processes falling into the "empty box" category (i.e. ARFIMA models have stochastic properties that essentially do not mimic the properties of the data).

[1] Many other empirical and theoretical studies are referenced in the entensive survery paper by Baillie (1996).

[2] see list of references under "Bank of Sweden (2003)" for a reference to the document.

In this chapter we offer new evidence on the usefulness of ARFIMA models by building on earlier work by Bhardwaj and Swanson (2004). In particular, while Bhardwaj and Swanson focus on primarily linear models, and do not take business cycle effects into account in their empirical analysis, we also consider a small group of nonlinear models, including STAR models, and we examine the effect that the business cycle has on predictive performance of a variety of models including ARFIMA, AR, MA, ARMA, GARCH, and STAR models. This is done via examination of ex ante forecasting evidence based on an updated version of the absolute returns series examined by Ding, Granger and Engle (DGE: 1993); and via the use of Diebold and Mariano (1995) and Clark and McCracken (2001) predictive accuracy tests. Results are presented for a variety of forecast horizons and for recursive and rolling estimation schemes.

Our approach is to divide the data that we examine into recession and non recession periods and analyze them separately. The data is further divided on the basis of pre and post World War II period. We also analyze the effect of oils shocks, and the recent post 1982 data; thus covering many important macroeconomic and financial developments. Interestingly, we find that the business cycle does not seem to have an effect on the relative forecasting performance of ARFIMA and non ARFIMA models. The single most important factor affecting the forecasting performance of the different models appears to be the sample size, with forecast horizon also an important determinant of forecast performance! In particular, we present evidence suggesting that long memory models may be particularly useful at longer forecast horizons, and that samples of 5000 or more observations yield very stable rolling and recursive estimates of d, while samples of 2500 or fewer observations lead to substantial increases in estimator standard errors, and affect the forecasting performance.

The rest of the chapter is organized as follows. In Section 2 we briefly review ARFIMA processes, and outline the empirical estimation methodology used in the rest of the chapter. In Section 3 predictive accuracy testing techniques, while section 4 outlines the predictive model selection procedures used. Of note certain parts of Sections 2, 3, and 4 which summarize estimators and tests also used in Bhardwaj and Swanson (2004) are taken from that paper. Finally, Section 5 presents the results of our empirical investigation, and Section 6 concludes.

2 Empirical Methods

The prototypical ARFIMA model examined in the literature is

$$\Phi\left(L\right)\left(1-L\right)^{d}y_{t}=\Theta\left(L\right)\epsilon_{t}, \tag{1}$$

where d is the fractional differencing parameter, ϵ_t is white noise, and the process is covariance stationary for $-0/5 < d < 0.5$, with mean reversion when $d < 1$. This model is a generalization of the fractional white noise process described in Granger (1980), Granger and Joyeux (1980), and Hosking (1981), where, for the purpose of analyzing the properties of the process, $\Theta\left(L\right)$ is set equal to unity (Baillie (1996) surveys numerous papers that have analyzed the properties of the ARFIMA process). Given that many time series exhibit very slowly decaying autocorrelations, the potential advantage of using ARFIMA models with hyperbolic autocorrelation decay patterns when modelling economic and financial times series seems clear (as opposed to models such as ARMA processes that have exponential or geometric decay). The potential importance of the hyperbolic decay property can be easily seen by noting that

$$\left(1-L\right)^{d}=\sum_{j=0}^{\infty}\left(-1\right)^{j}\left(\begin{array}{c} d \\ j \end{array}\right)\left(L\right)^{j}=1-dL+\frac{d(d-1)}{2!}L^{2}-\frac{d\left(d-1\right)\left(d-2\right)}{3!}L^{3}+\cdots=\sum_{j=0}^{\infty}b_{j}\left(d\right), \tag{2}$$

for any $d > -1$.[3]

There are currently dozens of estimation methods for and tests of long memory models. Perhaps one of the reasons for the wide array of tools for estimation and testing is that the current consensus suggests that good estimation and testing techniques remain elusive. Much of this evidence has been reported in the context of comparing one or two classes of estimators, such as rescaled range (RR) type estimators (as introduced by Hurst (1951) and modified by Lo (1991), for example) and log periodogram regression estimators due to Geweke and Porter-Hudak (GPH: 1983). We consider 4 quite widely used estimation methods. Perhaps the most glaring omission from our list of estimators is the full information maximum likelihood estimator of Sowell (1992a). While his estimator is theoretically appealing, it suffers from numerical instability upon actual implementation. However Doornik and Ooms (2003) provide an interesting algorithm, whereby maximum

[3] For $d > 0$, the differencing filter can also be expanded using hypergeometric functions, as follows: $(1-L)^d = \Gamma(-d)\sum_{j=0}^{\infty} L^k\Gamma(j-d)/\Gamma(j+1) = F(-d,1,1,L)$, where $F(a,b,c,z) = \Gamma(c)/[\Gamma(a)\Gamma(b)]\sum_{j=0}^{\infty} z^j\Gamma(a+j)\Gamma(b+j)/[\Gamma(c+j)\Gamma(j+1)]$

likelihood estimation becomes feasible. Their approach involves taking advantage of the Teoplitz structure of the covariance matrix using the Levinson-Durbin algorithm, avoiding the inversion of a TxT matrix.

2.1 Long Memory Model Estimation

2.1.1 GPH Estimator

The GPH estimation procedure is a two-step procedure, which begins with the estimation of d, and is based on the following log-periodogram regression[4]:

$$\ln\left[I\left(\omega_j\right)\right] = \beta_0 + \beta_1 \ln\left[4\sin^2\left(\frac{\omega_j}{2}\right)\right] + \nu_j, \tag{3}$$

where

$$\omega_j = \frac{2\pi j}{T}, j = 1, 2 \ldots, m.$$

The estimate of d, say $\widehat{d}_{GPH}$, is $-\widehat{\beta}_1$, ω_j represents the $m = \sqrt{T}$ Fourier frequencies, and $I\left(\omega_j\right)$ denotes the sample periodogram defined as

$$I\left(\omega_j\right) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} y_t e^{-\omega_j t}\right|^2. \tag{4}$$

The critical assumption for this estimator is that the spectrum of the ARFIMA(p,d,q) process is the same as that of an ARFIMA(0,d,0) process (the spectrum of the ARFIMA(p,d,q) process in (1), under some regularity conditions, is given by $I\left(\omega_j\right) = z\left(\omega_j\right)\left(2\sin\left(\frac{\omega_j}{2}\right)\right)^{-2d}$, where $z\left(\omega_j\right)$ is the spectrum of an ARMA process). We use $m = \sqrt{T}$, as is done in Diebold and Rudebusch (1989), although the choice of m when ϵ_t is autocorrelated can heavily impact empirical results (see Sowell (1992b) for discussion). Robinson (1995a) shows that $(\frac{\pi^2}{24m})^{-1/2}\left(\widehat{d}_{GPH} - d\right) \rightarrow N(0,1)$, for $-1/2 < d < 1/2$, and for $j = l, ..., m$ in the equation for ω above, where l is analogous to the usual lag truncation parameter. As is also the case with the next two estimators, the second step of the GPH estimation procedure involves fitting an ARMA model to the filtered data, given the estimate of d. Agiakloglou, Newbold and Wohar (1992) show that the GPH estimator has substantial finite sample bias, and is inefficient when ϵ_t is a persistent AR or MA process. Many authors assume normality of the filtered data in order to use standard estimation and inference procedures in the analysis of the final ARFIMA model (see e.g. Diebold and Rudebusch (1989,1991a)). Numerous variants of this estimator continue to be widely used in the empirical literature.[5]

[4] The regression model is usually estimated using least squares.

[5] For a recent overview of frequency domain estimators, see Robinson (2003, chapter 1).

2.1.2 WHI Estimator

Another seminparametric estimator, the Whittle estimator, is also often used to estimate d. Perhaps one of the more promising of these is the local Whittle estimator proposed by Künsch (1987) and modified by Robinson (1995b). This is another periodogram based estimator, and the crucial assumption is that for fractionally integrated series, the autocorrelation (ρ) at lag l is proportional to l^{2d-1}. This implies that the spectral density which is the Fourier transform of the autocovariance γ is proportional to $(\omega_j)^{-2d}$. The local Whittle estimator of d, say $\widehat{d}_{WHI}$, is obtained by maximizing the local Whittle log likelihood at Fourier frequencies close to zero, given by

$$\Gamma(d) = -\frac{1}{2\pi m}\sum_{j=1}^{m}\frac{I(\omega_j)}{f(\omega_j;d)} - \frac{1}{2\pi m}\sum_{j=1}^{m} f(\omega_j;d), \tag{5}$$

where $f(\omega_j;d)$ is the spectral density (which is proportional to $(\omega_j)^{-2d}$). As frequencies close to zero are used, we require that $m \to \infty$ and $\frac{1}{m}+\frac{m}{T} \to 0$, as $T \to \infty$. Taqqu and Teverovsky (1997) show that $\widehat{d}_{WHI}$ can be obtained by minimizing the following function:

$$\widehat{\Gamma}(d) = \ln\left(\frac{1}{m}\sum_{j=1}^{m}\frac{I(\omega_j)}{\omega_j^{-2d}}\right) - 2d\frac{1}{m}\sum_{j=1}^{m}\ln(\omega_j). \tag{6}$$

Robinson (1995b) shows that for estimates of d obtained in this way, $(4m)^{1/2}\left(\widehat{d}_{WHI} - d\right) \to N(0,1)$, for $-1/2 < d < 1/2$. Taqqu and Teverovsky (1997) study the robustness of standard, local, and aggregated Whittle estimators to non-normal innovations, and find that the local Whittle estimator performs well in finite samples. Shimotsu and Phillips (2002) develop an exact local Whittle estimator that applies throughout the stationary and nonstationary regions of d, while Andrews and Sun (2002) develop an adaptive local polynomial Whittle estimator in order to address the slow rate of convergence and associated large finite sample bias associated with the local Whittle estimator. In this paper, we use the local Whittle estimator discussed in Taqqu and Teverovsky (1997).

2.1.3 Data Driven Bandwidth Selection for GPH and WHI Estimators

The choice of bandwidth (m) is a crucial determinant of the rate of convergence for both of the semiparametric estimators listed above. Several Monte Carlo studies have documented the effect of bandwidth selection on the bias and variance of the long memory parameter for both the estimators (see e.g. Henry and Robinson (1996), Hurvich *et al.* (1998), and Taqqu and Teverovsky (1995)).

Hurvich *et al.* (1998) give a mean square minimizing bandwidth selection method for the GPH estimator and Henry and Robinson (1996) provide the same for the local Whittle estimator discussed above. Henry (2001) studies the robustness of the proposed bandwidth selection to the presence of conditional heteroscedasticity in the errors, and suggests the following optimal bandwidth formulae:

$$m_{GPH} = \left(\frac{27}{512\pi^2}\right)^{1/5} |\tau^*|^{-2/5} T^{4/5} \tag{7}$$

and

$$m_{WHI} = \left(\frac{3}{4\pi}\right)^{4/5} \left|\tau^* + \frac{d_x}{12}\right|^{-2/5} T^{4/5}, \tag{8}$$

where $d_x = \underset{d\in(-0.5,0.5)}{\arg\min} \widehat{\Gamma}(d)$ (for GPH, $d_x = -\widehat{\beta}_1$, where $\widehat{\beta}_1$ is the OLS estimate of slope in equation (3)). Thus it is an iterative procedure where in the first step the ad hoc value of m is chosen to be $T^{4/5}$. Further, $\tau^* = \frac{f^{*''}(0)}{2f^*(0)}$, where $f^*(0)$ and $f^{*''}(0)$ are the first and last coefficients in the OLS regression of the periodogram $I(\omega_j)$ against $|1 - \exp(i\omega_j)|^{-2d_x} \left(1, \omega_j, \frac{\omega_j^2}{2}\right)$, for $j = 1, 2 \ldots, m$. Henry (2001) reports that this iterative procedure often fails to converge in Monte Carlo experiments if τ^* is updated at every iteration, while keeping τ^* fixed at the first iteration value of $m \left(= T^{4/5}\right)$ results in convergence at the second iteration.

2.1.4 RR Estimator

The rescaled range estimator was originally proposed as a test for long-term dependence in the time series. The statistics is calculated by dividing range with standard deviation. In particular, define:

$$\widehat{Q}_T = \frac{\widehat{R}_T}{\widehat{\sigma}_T}, \tag{9}$$

where $\widehat{\sigma}_T^2$ is the usual maximum likelihood variance estimator of y_t , and $\widehat{R}_T = \max_{0<i\leq T} \sum_{t=1}^{i} (y_t - \overline{y}) - \min_{0<i\leq T} \sum_{t=1}^{i} (y_t - \overline{y})$. The estimate of d, say $\widehat{d}_{RR}$, is obtained using the result that $plim_{T\to\infty}(T^{-d-\frac{1}{2}} \frac{\widehat{R}_T}{\widehat{\sigma}_T})$ $=$ *constant* (see Hurst (1951), Lo (1991), and the references cited therein), and is:

$$\widehat{d}_{RR} = \frac{\ln\left(\widehat{Q}_T\right)}{\ln(T)} - \frac{1}{2}. \tag{10}$$

Lo (1991) shows that $T^{-1/2}\widehat{Q}_T$ is asymptotically distributed as the range of a standard Brownian bridge. With regard to testing in this context, note that there are extensively documented deficiencies associated with long memory tests based on $T^{-1/2}\widehat{Q}_T$, particularly in the presence of data generated by a short memory processes combined with a long memory component (see e.g. Cheung

(1993)). For this reason, Lo (1991) suggests the modified RR test, whereby $\widehat{\sigma}_T^2$ is replaced by a heteroskedasticity and autocorrelation consistent variance estimator, namely:

$$\widehat{\sigma}_T^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \overline{y})^2 + \frac{2}{T}\sum_{j=1}^{q} w_j(q)\left\{\sum_{t=j+1}^{T}(y_t - \overline{y})(y_{t-j} - \overline{y})\right\}, \tag{11}$$

where

$$w_j(q) = 1 - \frac{j}{q+1}, q < T,$$

It is known from Phillips (1987) that $\widehat{\sigma}_T^2$ is consistent when $q = O(T^{1/4})$, at least in the context of unit root tests, although choosing q in the current context is a major difficulty. This statistic still weakly converges to the range of a Brownian bridge.

2.1.5 AML Estimator

The fourth estimator that we use is the approximate maximum likelihood estimator of Beran (1995). For any ARFIMA model given by equation (1), $d = m + \delta$, where $\delta \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, and m is an integer (assumed known) denoting the number of times the series must be differenced in order to attain stationarity, say:

$$x_t = (1 - L)^m y_t. \tag{12}$$

To form the estimator, a value of δ is fixed, and an ARMA model is fitted to the filtered x_t data, yielding a sequence of residuals. This is repeated over a fine grid of $d = m + \delta$, and $\widehat{d}_{AML}$ is the value which minimizes the sum of squared residuals. The choice of m is critical, given that the method only yields asymptotically normal estimates of the parameters of the ARFIMA model if $\delta \in \left(-\frac{1}{2}, \frac{1}{2}\right)$, for example (see Robinson (2003, chapter 1) for a critical discussion of the AML estimator).

In summary, three of the estimation methods described in the preceding paragraphs for ARFIMA models require first estimating d. When fitting ARFIMA models, we used an arbitrary cut-off of $1.0e - 004$. Terms in the polynomial expansion with coefficients smaller in absolute value than this cut-off were truncated. Thereafter, an ARMA model is fitted to the filtered data by using maximum likelihood to estimate parameters, and via the use of the Schwarz Information Criterion for lag selection. The maximum number of lags was picked for each of the datasets examined in our empirical section by initially examining the first half of the sample to ascertain what sorts of lag structures were usually chosen using the SIC. The exception to the above approach is the AML estimator, for which a grid of d values is searched across, with a new ARMA model fitted for each values of d in the grid, and resulting models compared using mean square error.

2.2 Short Memory Models

In our empirical investigation, the following short memory models are used (apart from STAR models):

1) *Random Walk:* $y_t = y_{t-1} + \epsilon_t$;

2) *AR(p):* $\Phi(L)\, y_t = \alpha + \epsilon_t$;

3) *MA(q):* $y_t = \alpha + \Theta(L)\, \epsilon_t$;

4) *ARMA(p,q):* $\Phi(L)\, y_t = \alpha + \Theta(L)\, \epsilon_t$;

5) *ARIMA(p,d,q):* $\Phi(L)\,(1-L)^d\, y_t = \alpha + \Theta(L)\, \epsilon_t$, where d can take integer values;

6) *GARCH:* $\Phi(L)\, y_t = \alpha + \epsilon_t$, where $\epsilon_t = h_t^{1/2}\nu_t$ with $E(\epsilon_t^2|\Im_{t-1}) = h_t = \varpi + \alpha_1\epsilon_{t-1}^2 + \cdots + \alpha_q\epsilon_{t-q}^2 + \beta_1 h_{t-1} + \cdots + \beta_p h_{t-p}$, and where $\Im_{t-1}$ is the usual filtration of the data; and

In these models, ϵ_t is the disturbance term, $\Phi(L) = 1 - \phi_1 L - \phi_2 L^2 - \cdots - \phi_p L^p$, and $\Theta(L) = 1 - \theta_1 L - \theta_2 L^2 - \cdots - \theta_q L^q$, where L is the lag operator. All models are estimated using (quasi) maximum likelihood, with values of p and q chosen via use of the Schwarz Information Criterion (SIC), and integer values of d in ARIMA models selected via application of the augmented Dickey Fuller test at a 5% level. Errors in the GARCH models are assumed to be normally distributed.

2.3 Non-linear STAR Models

Over the past two decades there has been a growing interest in time series models that allow for regime switching behavior. In particular, starting with Hamilton (1989), regime switching models have been widely used in various macroeconomic and financial applications.[6] One of the most popular regime switching type models is the smooth transition autoregressive (STAR) model. In this paper we consider the following two variants of the STAR model.

1) *Logistic STAR (LSTAR)*:
$$y_t = (\alpha_1 + \phi_{1,1}y_{t-1} + \phi_{1,2}y_{t-2} + \cdots + \phi_{1,p}y_{t-p})\,(1 - G(s_t, \gamma, c)) + (\alpha_2 + \phi_{2,1}y_{t-1} + \phi_{2,2}y_{t-2} + \cdots + \phi_{2,p}y_{t-p})\, G(s_t, \gamma, c) + \epsilon_t,$$
where $G(s_t, \gamma, c) = (1 + \exp\{-\gamma(s_t - c)\})^{-1}, \gamma > 0$;

2) *Exponential STAR (ESTAR)*: $y_t = (\alpha_1 + \phi_{1,1}y_{t-1} + \phi_{1,2}y_{t-2} + \cdots + \phi_{1,p}y_{t-p})\,(1 - G(s_t, \gamma, c))$

[6] For a survey of regime switching methods and applications, see Kim and Nelson (1999).

$$+\left(\alpha_2+\phi_{2,1}y_{t-1}+\phi_{2,2}y_{t-2}+\cdots+\phi_{2,p}y_{t-p}\right)G\left(s_t,\gamma,c\right)+\epsilon_t,$$

where $G\left(s_t,\gamma,c\right)=1-\exp\left\{-\gamma\left(s_t-c\right)^2\right\},\gamma>0$;

In these models, ϵ_t is the disturbance term, $G\left(s_t,\gamma,c\right)$ is a continuous transition function bounded between 0 and 1, and s_t is transition variable which is assumed to be a lagged endogenous variable. STAR models can be interpreted as regime switching models with two regimes that are associated with the two extreme values of the transition function i.e. 0 and 1. However, given the continuous transition function, the transition from one regime to the other is smooth. STAR models can be estimated using nonlinear least squares (NLS) which can be interpreted as (quasi) maximum likelihood estimation.[7]

3 Predictive Accuracy Testing

If, as is often the case, the ultimate goal of an empirical investigation is the specification of predictive models, then a natural tool for testing for the presence of long memory is the predictive accuracy test. In this case, if an ARFIMA model can be shown to yield predictions that are superior to those from a variety of alternative linear (and nonlinear) models, then one has direct evidence of long memory, at least in the sense that the long memory model is the best available "approximation" to the true underlying DGP. There is a rich recent literature on predictive accuracy testing, most of which draws in one way or another on Granger and Newbold (1986), where simple tests comparing mean square forecast errors (MSFEs) of pairs of alternative models under assumptions of normality are outlined. Perhaps the most important of the predictive accuracy tests that have been developed over the last 20 years is the Diebold and Mariano (1995: DM) test. The statistic is:

$$\widehat{d}_P=P^{-1/2}\frac{\sum_{t=R-h+1}^{T-1}(f(\widehat{v}_{0,t+h})-f(\widehat{v}_{1,t+h}))}{\widehat{\sigma}_P}, \tag{13}$$

where R denotes the estimation period, P is the prediction period, f is some generic loss function, $h\geq 1$ is the forecast horizon, $\widehat{v}_{0,t+h}$ and $\widehat{v}_{1,t+h}$ are h-step ahead prediction errors for models 0 and 1 (where model 0 is assumed to be the ARFIMA model), constructed using consistent estimators, and $\widehat{\sigma}_P^2$ is defined as

$$\widehat{\sigma}_P^2=\frac{1}{P}\sum_{t=R-h+1}^{T-1}(f(\widehat{v}_{0,t+h})-f(\widehat{v}_{1,t+h}))^2+\frac{2}{P}\sum_{j=1}^{l_P}w_j\sum_{t=R-h+1+j}^{T-1}(f(\widehat{v}_{0,t+h})-f(\widehat{v}_{1,t+h}))(f(\widehat{v}_{0,t+h-j})-f(\widehat{v}_{1,t+h-j})) \tag{14}$$

[7] For a detailed discussion of estimation, see van Dijk, Teräsvirta and Franses (2002), and the references cited therein.

where $w_j = 1 - \frac{j}{l_P+1}$, $l_P = o(P^{1/4})$. The hypotheses of interest are

$$H_0 : E(f(v_{0,t+h}) - f(v_{1,t+h})) = 0,$$

and

$$H_A : E(f(v_{0,t+h}) - f(v_{1,t+h})) \neq 0.$$

The DM test, when constructed as outlined above for nonnested models, has a standard normal limiting distribution under the null hypothesis.[8] West (1996) shows that when the out-of-sample period grows at a rate not slower than the rate at which the estimation period grows (i.e. $\frac{P}{R} \to \pi$, with $0 < \pi \leq \infty$), parameter estimation error generally affects the limiting distribution of the DM test in stationary contexts. On the other hand, if $\pi = 0$, then parameter estimation error has no effect. Additionally, Clark and McCracken (2001) point out the importance of addressing the issue of nestedness when applying DM and related tests.[9] Other recent papers in this area include Christoffersen (1998), Christoffersen and Diebold (1997), Clements and Smith (2000,2002), Corradi and Swanson (2002), Diebold, Gunther and Tay (1998), Diebold, Hahn and Tay (1999), Harvey, Leybourne and Newbold (1998), and the references contained therein, to name but a few. Although the DM test does not have a normal limiting distribution under the null of non causality when nested models are compared, the statistic can still be used as an important diagnostic in predictive accuracy analyses. Furthermore, the nonstandard limit distribution is reasonably approximated by a standard normal in many contexts (see McCracken (1999) for tabulated critical values). For this reason, and as a rough guide, we use critical values gotten from the $N(0, 1)$ distribution when carrying out DM tests. A final caveat that should be mentioned is that the work of McCracken (and that of Clark and McCracken discussed below) assumes stationarity, assumes correct specification under the null hypothesis, and often assumes that estimation is via least squares. Of course, if we are willing to make the strong assumption of correct specification under the null, then the ARFIMA model and the non-ARFIMA models are the same, implying for example that $d = 0$, so that only the common ARMA components in the models remain, and hence errors are short-memory. Nevertheless, it is true that in general some if not many of the assumptions may be

[8] We assume quadratic loss in our applications, so that $f(v_{0,t+h}) = v_{0,t+h}^2$, for example.

[9] Chao, Corradi, and Swanson (2001) address not only nestedness, by using a consistent specification testing approach to predictive accuracy testing, but also allow for misspecification amongst competing models; an important feature if one is to presume that all models are approximations, and hence all models may be (dynamically) misspecified. White (2000) further extends the Diebold and Mariano framework by allowing for the joint comparison of multiple models, while Corradi and Swanson (2003,2004a,b) extend White (2000) to predictive density evaluation with parameter estimation error.

broken in our context, and extensions of their tests and related tests to more general contexts is the subject of ongoing research by a number of authors.[10] This is another reason why the critical values used in this chapter should be viewed only as rough approximations.

We also report results based on the application of the Clark and McCracken (CM: 2001) encompassing test, which is designed for comparing nested models. The test statistic is

$$ENC-t = (P-1)^{1/2}\frac{\overline{c}}{(P^{-1}\sum_{t=R}^{T-1}(c_{t+h}-\overline{c}))^{1/2}},$$

where $c_{t+h} = \widehat{v}_{0,t+h}(\widehat{v}_{0,t+h} - \widehat{v}_{1,t+h})$ and $\overline{c} = P^{-1}\sum_{t=R}^{T-1} c_{t+1}$. This test has the same hypotheses as the DM test, except that the alternative is $H_A : E(f(v_{0,t+h}) - f(v_{k,t+h})) > 0$. If $\pi = 0$, the limiting distribution is $N(0,1)$ for $h = 1$. The limiting distribution for $h > 1$ is non-standard, as discussed in CM. However, as long as a Newey-West (1987) type estimator (of the generic form given above for the DM test) is used when $h > 1$, then the tabulated critical values are quite close to the $N(0,1)$ values, and hence we use the standard normal distribution as a rough guide for all horizons (see CM for further discussion).

Following Granger and Hyung (2004), we also run encompassing regressions, where we regress the observed time series on τ-step ahead forecasts from ARFIMA and other models. In particular, we run the two following regressions:

$$y_{t+\tau} = \beta_0 + \beta_{ARFIMA}y_{t+\tau,t}^{ARFIMA} + \beta_{non-ARFIMA}y_{t+\tau,t}^{non-ARFIMA} + \varepsilon_{t+\tau}$$

and

$$y_{t+\tau} = \beta_0 + \beta_{ARFIMA}y_{t+\tau,t}^{ARFIMA} + \beta_{STAR}y_{t+\tau,t}^{STAR} + \varepsilon_{t+\tau}.$$

If the forecasts from non-ARFIMA (STAR) models have no additional forecasting power over those from the ARFIMA model we would expect that the coefficients $\beta_{non-ARFIMA}$ (β_{STAR}) to be statistically insignificant. h-step

4 Predictive Model Selection

In the sequel, forecasts are 1-step, 5-steps and 20-steps ahead, when daily stock market data are examined, corresponding to 1-day, 1-week and 1-month ahead predictions. Estimation is carried out as discussed above for ARFIMA models, and using maximum likelihood for non-ARFIMA models.

[10] For example, for further discussion of the ramifications of using nonstationary variables when constructing tests of predictive ability, see Corradi, Swanson and Olivetti (2001) and Rossi (2003).

More precisely, each sample of T observations is first split in half. The first half of the sample is then used to produce $0.25T$ rolling (and recursive) predictions (the other $0.25T$ observations are used as the initial sample for model estimation) based on rolling (and recursively) estimated models (i.e. parameters are updated before each new prediction is constructed).[11] These predictions are then used to select a "best" ARFIMA, a "best" non-ARFIMA (chosen from the short memory models excluding the STAR models), and a "best" STAR model, based on point out-of-sample mean square forecast error comparison. At this juncture, the specifications of the ARFIMA, non-ARFIMA, and STAR models to be used in later predictive evaluation are fixed. Parameters in the models may be updated, however. In particular, recursive and rolling ex ante predictions of the observations in the second half of the sample are then constructed, with parameters in the "best" models updated before each new forecast is constructed. Additionally, different models are constructed for each forecast horizon, as opposed to estimating a single model and iterating forward when constructing multiple step ahead forecasts. Reported DM and encompassing t-tests are thus based on the second half of the sample, and involve comparing only two models. We report the comparison of "best" ARFIMA model with the "best" non-ARFIMA model, and that of "best" ARFIMA model with the "best" STAR model.

It should be stressed that the methodology presented above is often used in 'horse-races' of the type that we are carrying out, so as not to "cherry-pick" the forecast-best models (see e.g. Swanson and White (1995,1997) and the references cited therein). However, there are many other ways to avoid issues that arise when comparing many models, such as the prevalence of sequential test-bias and overfitting. For recent important papers that address these and related issues, the reader is referred to White (2000), Inoue and Kilian(IK: 2003), Corradi and Swanson (2004a), and Hansen, Lunde and Nason (HLN: 2004). IK suggest the use of information criterion (such as the Schwarz Information Criterion - SIC) for choosing the best forecasting model, while HLN propose a model confidence set approach to the same problem. Of note, though, is that the SIC based approach of IK is not applicable under near stationarity and non-linearity, and is not consistent when non-nested models are being compared. HLN takes a different approach, as they are concerned with narrowing down from a larger set of models to a smaller set that includes the best forecasting model. When their approach is used, for example, it is found that ARFIMA volatility models do not outperform

[11] An interesting and potentially very useful alternative to the h-step ahead recursive prediction used here involves implementing the Levinson-Durbin algorithm, as outlined in Brockwell and Dahlhaus (2004). Implementation of this algorithm, however is left to future research.

simpler non-ARFIMA volatility models.

In the proceeding sections, we carry out our empirical investigation by examining the long memory and ARFIMA predictive properties of the S&P500 series used by DGE and Granger and Ding (1996). Our dataset is an updated version of the long historical S&P500 returns dataset of DGE. The period covered is January 4, 1928 - September 30, 2003 (20,105 observations), so that our dataset is somewhat longer than the 17,054 observations (ending on August 30, 1990) examined by DGE.

5 Empirical Results

Table 1.a summarizes results based on analysis of our long returns dataset. Before discussing these results, however, it is first worth noting that the four alternative estimators of d yield quite similar estimates (except the RR estimator). In particular, note that if one were to use the first half of the sample for estimation, one would find values of d equal to 0.49 (GPH), 0.41 (AML), 0.31 (RR) and 0.43 (WHI).[12] Furthermore, all methods find one AR lag, and all but one method finds 1 MA lag. This is as expected for large samples. Bhardwaj and Swanson (2004) show that 4 estimators yield radically different values even when the in-sample period used is moderately large, with approximately 2500 observations, so that the convergence of the estimators is extremely slow, although they do eventually converge. The same is observed below when we analyze series of smaller length. This yields credence to Granger's (1999) observation that estimates of d can vary greatly across different sample periods and sample sizes, and are generally not robust at all.

In the table, the "best" ARFIMA, non-ARFIMA and STAR models are first chosen as discussed above. As d is re-estimated prior to the construction of each new forecast, means and standard errors of the sequence of d values are reported in the table. As might be expected, different d mean values, which are calculated for each estimation scheme (i.e. recursive or rolling) and each forecast horizon, are all quite close to one another, with the exception of the RR estimator. Additionally, all standard errors are extremely small. Interestingly, though, the means are always above 0.5 except in the case of RR estimator. This is in contrast to the usual finding that $d < 0.5$. Although various explanations for these seemingly large values of d are possible, a leading explanation might be as follows. If, as suggested by Clive Granger and others, long memory arises in part due to various sorts

[12] These estimates of d are very close to those obtained by Ding, Granger and Engle (1993) and by Granger and Ding (1996) using their fractionally integrated ARCH model.

of misspecification, then it may be the case that greater accumulation of misspecification problems leads to greater "spurious" long memory. In the sense that our multiple step ahead prediction models may be more poorly specified than our 1-step ahead models (given that we construct a new prediction model for each horizon, and that greater horizons involve using more distant lags of the dependent variable on the RHS of the forecasting model), we have indirect evidence that more severe misspecification, in the form of missing dynamic information, may lead to larger estimated values for d. This finding, if true, has implications for empirical research, as it may help us to better understand the relative merits of using different approaches for constructing multiple-step ahead forecasting models. Finally, it should be stressed that the best ARFIMA/non-ARFIMA models yield significantly better forecasts, when compared to a 'naive' forecasts based on a random walk (or no-change) model.

Turning next to the DM and encompassing test results reported in the table, notice that the DM statistics are negative in all but one case. As the ARFIMA model is taken as model 0 (see discussion in Section 3), this means that the point MSFEs are lower for the ARFIMA model than the non-ARFIMA/STAR model. The exception is the case where the rolling estimation scheme is used and $h = 1$ (this is the case where the RR estimator is used, and where the average d value across the out-of-sample period is 0.25). In the other cases, use of the rolling estimation scheme results in significantly superior multiple-step ahead predictions for the ARFIMA model when compared with non-ARFIMA models, at standard significance levels. This finding is relevant, given that the MSFEs are quite similar when comparing recursive and rolling estimation schemes. The encompassing t-test yields somewhat similar results. In particular, the null hypothesis is most clearly rejected in favor of the alternative that the non-ARFIMA model is the more precise predictive model for the rolling estimation scheme with $h = 1$. In contrast with the results based on the DM test, the null is also be rejected for $h = 20$ when recursive estimation is used (the statistic value is 2.91), although for $h = 20$, using critical values from the $N(0, 1)$ is only a rough approximation, as the distribution is nonstandard, and contains nuisance parameters (so that, in principle, bootstrap methods need to be should to be valid and need to be used in order to obtain valid critical values, for example).

It should also be noted that DM statistics are always negative when comparing ARFIMA and STAR models, and use of ARFIMA model results in significantly better prediction than the STAR models. The encompassing t-test yields similar results.

Table 1.b extends the set of ARFIMA models to include GPH and WHI estimation schemes where the bandwidth is selected by the data driven procedures noted in Henry (2001), as discussed above. The two cases are labeled GPH-OPT and WHI-OPT, respectively. It should be noted that in the horse race for the best ARFIMA models, the models with data driven bandwidth selection are chosen four out of six times. Although there thus seems to be a clear case for including these models when comparing ARFIMA models with non-ARFIMA and STAR models there is no significant difference in the forecasting performance results of Table 1.a when the bandwidth selection is data driven. It should also be noted that the average of the estimated values of d across the entire ex ante sample with the data driven bandwidth selection is very similar to that with fixed bandwidth, and the standard errors are much smaller when the bandwidth is fixed. Given these considerations, in the rest of the paper we discuss the case of fixed bandwidth only. Further results for the data driven bandwidth case are available upon request.

Table 1.c reports the results for the encompassing regressions discussed above. The results support the findings from the forecast accuracy test reported in Table 1.a. However for two cases under the rolling estimation scheme (1 and 20 steps ahead), β_{STAR} is highly significant, indicating that structural break models do provide additional useful information for modeling the absolute return series, as documented by Granger and Hyung (2004). This result is strengthed when we compare the "best" ARFIMA model with the pooled forecasts constructed by taking the median values across the point forecasts from the linear non-ARFIMA and STAR models. As reported in Table1.c, even for the largest sample size considered, there appears little to choose between ARFIMA and pooled median forecasts from non-ARFIMA and STAR models.

While the results discussed above and summarized in Table 1 are somewhat mixed, they do constitute evidence that long memory models may actually be useful in certain cases (i.e. large samples and multiple step-ahead prediction), when constructing forecasting models. Correspondingly, as long as the in-sample period is very large, then all of our differencing operator estimators perform adequately (with the possible exception of the RR estimator), and any one of them can be successfully used to estimate "winning" prediction models. Put differently, no model from amongst those considered performs better than our simple ARFIMA models, at least based on point MSFE (with the one exception that is noted above). It should, however, be stressed that structural breaks, regime switching, etc. have not been accounted for thus far (except when STAR models are used), and it remains to see whether the types of results obtained here will also hold when structural breaks and regime switching are allowed for in both our short memory and long memory models.

Some results in this regard are given in the next subsection, where different periods of data are examined. In particular, we explore the forecasting performance of the models during recession and non-recession periods, and we also analyze the period of oil shocks from 1973 to 1982, and the post 1982 period separately.

5.1 S&P500 Returns: Business Cycle Effects

To capture the effect of business cycles on the forecasting performance of the models analyzed in this chapter, we have considered six different divisions of the long return series considered above. The first and the most obvious division is into recession versus non recession periods. In choosing the dates of business-cycle turning points we follow the chronology of the U.S. business cycle as reported by National Bureau of Economic Research.[13] To analyze the effect of business cycles on the forecasting performance of these models, we wanted to club together the data for all the recession (expansions) periods. In order to justify doing so we carried out a small experiment, where we fitted simple linear autoregressive models with dummy variables for all the recession (expansion) periods. The dummy variables for pre World War II periods turned out to be significant, especially for the 1929 great depression and the subsequent recovery. Given these results we decided to also divide the data into pre and post World War II periods.

In an attempt to classify major recent global developments two further periods were also considered. The first period starts with the world oil shock of 1973 that began on October 17, 1973, when Arab members of the Organization of Petroleum Exporting Countries (OPEC), in the midst of the Yom Kippur War, announced that they would no longer ship petroleum to nations that had supported Israel in its conflict with Egypt and Syria; i.e. to the United States and its allies in Western Europe. At around the same time, OPEC-member states agreed to use their leverage over the world price-setting mechanism for oil to quadruple world oil prices. This period ended with a rapid decline in oil prices early in 1982 when OPEC appeared to lose control over world oil prices. Finally OPEC agrees to individual output quotas and cuts prices by $5. The other period considered is the most recent period starting in 1982. This period covers the two very long episodes of expansion in the 80's and 90's, the stock market crash of 1987, and the recent recession of 2001. Result are reported in Tables 2 to 7.

It turns out that the single most important factor that affects the performance of these models,

[13] For further details on actual dates and methodology see, http://www.nber.org/cycles.html/

especially the ARFIMA models, is the sample size. Based on the performance of long memory models we can divide our six data groupings into two categories i.e. small sample size, and moderate or large sample size. In the small sample size category we have the pre World War II recession (1,387 data points), the pre World War II expansion (1,707 data points), the post World War II recession (2,401 data points), and the 1973-82 period of oil shocks (2,273 data points). In the other category we have the post World War II expansion (12,141 data points), and the most recent data in the post 82 period (5,369 data points). The first thing to note is that while for the longer data sets, four estimates of d yield quite similar estimates, we have huge variations when the sample size is small.

For the smaller sample sizes, and based on the use of DM and encompassing tests, there is little to choose between ARFIMA and non ARFIMA models, most of the time. For example, for the pre World War II expansion and the pre World War II recession, in all but two cases DM test statistics fails to significantly distinguish between ARFIMA and non AFIMA models. However, based on point mean square forecast error, ARFIMA models do seem to out perform non-ARFIMA models more than half of the time. Note, however, that for one of the smaller samples (i.e. the pre World War II expansion), the non-ARFIMA model outperforms the ARFIMA model based on point mean square forecast error for all the cases expect the longest horizon forecasting (i.e. 20 days ahead). With regard to the STAR models, they are clearly out performed by the ARFIMA models for the larger sample sizes, although based on the smaller sample sizes there is little difference between their respective forecasting performances. It should be further noted that as reported in Table 7, for the post 1982 period, and for one day ahead forecasts, non-ARFIMA models clearly have lower point MSFE compared to ARFIMA models. Since one day ahead forecasts are important to practitioners this is a notewothy observation.[14]

Of further note is that it is clear that when the sample size increases, ARFIMA models significantly outperform the STAR model. Finally, though the motivation for the data groupings was to capture the possible effects of the business cycle on model performance, what we have found is that the most important factor seems to be the sample size and that ARFIMA models clearly improve their performance as the sample size becomes large; suggesting, at least in part, the importance of estimating d as precisely as possible when constructing ARFIMA based prediction models.

[14] We thank the editors for pointing this out.

6 Concluding Remarks

We have presented the results of an empirical study of the usefulness of ARFIMA models in a practical prediction based application where returns data are the object of interest, and find evidence that such models may yield reasonable approximations to unknown underlying DGPs, in the sense that the models often significantly outperform a fairly wide class of the benchmark non-ARFIMA models, including AR, ARMA, ARIMA, random walk, GARCH, and STAR models. This finding is particularly apparent with longer samples of data, underscoring the importance of estimating d as precisely as possible when constructing ARFIMA type forecasting models. Interestingly, there appears little to choose between various estimators of d when samples are as large as often encountered in financial economics. Overall, and in support of the finding of Bhardwaj and Swanson (2004), we conclude that long memory processes, and in particular ARFIMA processes, might not fall into the "empty box" category after all, although much further research is needed before overwhelmingly conclusive evidence in either direction can be given.

7 References

Agiakloglou, C., P. Newbold and M. Wohar (1992), Bias in an Estimator of the Fractional Difference Parameter, Journal of Time Series Analysis 14:235-246.

Andrews, D.W.K. and Y. Sun (2002), Adaptive Local Whittle Estimation of Long-range Dependence, Working Paper, Yale University.

Baillie, R.T. (1996), Long Memory Processes and Fractional Integration in Econometrics, Journal of Econometrics 73:5-59.

Bank of Sweden (2003), Time-Series Econometrics: Cointegration and Autoregressive Conditional Heteroskedasticity, Advanced Information on the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, The Royal Swedish Academy of Sciences.

Beran, J. (1995), Maximum Likelihood Estimation of the Differencing Parameter for Invertible Short and Long Memory Autoregressive Integrated Moving Average Models, Journal of the Royal Statistical Society: Series B 57:659-672.

Bhansali R.J. and P.S. Kokaszka (2002), Computation of the forecast coefficients for multistep prediction of long-range dependent time series, International Journal of Forecasting 18:181-206.

Bhardwaj, G. and N.R. Swanson (2004), An Empirical Investigation of the Usefulness of ARFIMA Models For Predicting Macroeconomic and Financial Time Series, *Journal of Econometrics,* forthcoming.

Bos, C.S., P.H. Franses and M. Ooms (2002), Inflation, Forecast Intervals and Long Memory Regression Models, International Journal of Forecasting 18:243-264.

Breitung, Jörg and U. Hassler (2002), Inference on the Cointegration Rank in Fractionally Integrated Processes, Journal of Econometrics 110:167-185.

Brockwell, P.J. and R. Dahlhaus (2004), Generalized Levinson-Durbin and Burg Algorithms, Journal of Econometrics 118:129-149.

Chao, J.C., V. Corradi and N.R. Swanson (2001), An Out of Sample Test for Granger Causality, Macroeconomic Dynamics 5:598-620.

Cheung, Y.-W. (1993), Tests for Fractional Integration: A Monte Carlo Investigation, Journal of Time Series Analysis 14:331-345.

Cheung, Y.-W. and F.X. Diebold (1994), On Maximum Likelihood Estimation of the Difference Parameter of Fractionally Integrated Noise with Unknown Mean, Journal of Econometrics 62:301-316.

Chio, K. and E. Zivot (2002), Long Memory and Structural Changes in the Forward Discount: An Empirical Investigation, Working Paper, University of Washington.

Christoffersen, P.F. (1998), Evaluating Interval Forecasts, International Economic Review 39:841-862.

Christoffersen, P. and F.X. Diebold (1997), Optimal Prediction Under Asymmetric Loss, Econometric Theory 13:808-817.

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

Clements, M.P. and J. Smith (2000), Evaluating the Forecast Densities of Linear and Nonlinear Models: Applications to Output Growth and Unemployment, Journal of Forecasting 19:255-276.

Clements, M.P. and J. Smith (2002), Evaluating Multivariate Forecast Densities: A Comparison of Two Approaches, International Journal of Forecasting 18:397-407.

Corradi, V. and N.R. Swanson (2002), A Consistent Test for Out of Sample Nonlinear Predictive Ability, Journal of Econometrics 110:353-381.

Corradi, V. and N.R. Swanson (2003), The Block Bootstrap for Parameter Estimation Error in Recursive Estimation Schemes, With Applications to Predictive Evaluation, Working Paper, Rutgers University.

Corradi, V. and N.R. Swanson (2004a), Predictive Density Accuracy Tests, Working Paper, Rutgers University.

Corradi, V. and N.R. Swanson (2004b), Predictive Density Evaluation, forthcoming in: Graham Elliott, Clive W.J. Granger and Allan Timmerman, eds., Handbook of Economic Forecasting (Elsevier, Amsterdam).

Corradi, V., N.R. Swanson and C. Olivetti (2001), Predictive Ability with Cointegrated Variables, Journal of Econometrics 104:315-358.

Diebold, F.X., T. Gunther and A.S. Tay (1998), Evaluating Density Forecasts with Applications to Finance and Management, International Economic Review 39:863-883.

Diebold, F.X., J. Hahn and A.S. Tay (1999), Multivariate Density Forecast Evaluation and Calibration in Financial Risk Management: High Frequency Returns on Foreign Exchange, Review of Economics and Statistics 81:661-673.

Diebold, F. and A Inoue (2001), Long Memory and Regime Switching, Journal of Econometrics 105:131-159.

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

Diebold, F.X. and G.D. Rudebusch (1989), Long Memory and Persistence in Aggregate Output, Journal of Monetary Economics 24:189-209.

Diebold, F.X. and G.D. Rudebusch (1991a), Is Consumption Too Smooth? Long Memory and the Deaton Paradox, Review of Economics and Statistics 73:1-9.

Diebold, F.X. and G.D. Rudebusch (1991b), On the Power of the Dickey-Fuller Test Against Fractional Alternatives, Economics Letters 35:155-160.

Ding, Z, C.W.J. Granger and R.F. Engle (1993), A Long Memory Property of Stock Returns and a New Model, Journal of Empirical Finance 1:83-106.

Dittman, I. and C.W.J. Granger (2002), Properties of Nonlinear Transformations of Fractionally Integrated Processes, Journal of Econometrics 110:113-133.

Doornik, J.A. and M. Ooms (2003), Computational Aspects of Maximum Likelihood Estimation of Autoregressive Fractionally Integrated Moving Average Models, Computational Statistics and Data Analysis 42:333-348.

Engle, R.F. and A.D. Smith (1999), Stochastic Permanent Breaks, Review of Economics and Statistics 81:553-574.

Geweke, J. and S. Porter-Hudak (1983), The estimation and application of long memory time series models, Journal of Time Series Analysis 4:221-238.

Granger, C.W.J. (1969), Investigating Causal Relations by Econometric Models and Cross-Spectral Methods, Econometrica 37:424-438.

Granger, C.W.J. (1980), Long Memory Relationships and the Aggregation of Dynamic Models, Journal of Econometrics 14:227-238.

Granger, C.W.J. (1999), Aspects of Research Strategies for Time Series Analysis, Presentation to the conference on New Developments in Time Series Economics, Yale University.

Granger, C.W.J. and A.P. Andersen (1978), Introduction to Bilinear Time Series Models (Vandenhoeck and Ruprecht: Göttingen).

Granger, C.W.J., and Z. Ding (1996), Varieties of Long Memory Models, Journal of Econometrics 73:61-77.

Granger, C.W.J. and M. Hatanaka (1964), Spectral Analysis of Economic Time Series (Princeton University Press: Princeton).

Granger, C.W.J. and N. Hyung (2004), Occasional Structural Breaks and Long Memory with application to the S&P 500 absolute stock returns, Journal of Emperical Finance, 11:399-421.

Granger, C.W.J. and R. Joyeux (1980), An Introduction to Long Memory Time Series Models and Fractional Differencing, Journal of Time Series Analysis 1:15-30.

Granger, C.W.J. and P. Newbold (1986), Forecasting Economic Time Series (Academic Press, San Diego).

Hamilton, J.D. (1989), A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle, Econometrica, 57:357-384.

Henry, M, (2001), Robust Automatic Bandwidth for Long Memory, Journal of Time Series Analysis 22:293-316.

Hansen, P.R., A. Lunde and J. M. Nason (2004), Model Confidence Sets for Forecasting Models, Working Paper, Brown University.

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

Hassler, U. and J. Wolters (1995), Long Memory in Inflation Rates: International Evidence, Journal of Business and Economic Statistics 13:37-45.

Henry, M. and P. M. Robinson (1996), Bandwidth Choice in Gaussian Semiparametric Estimation of Long Range Dependence. In *Athens Conference on Applied Probability and Time Series Analysis*, Vol. II: *Time Series Analysis, In Memory of E. J. Hannan* (eds. P.M. Robinson and M. Rosenblatt), New York: Springer, 220-32.

Hosking, J. (1981), Fractional Differencing, Biometrica 68:165-76.

Hurst, H.E. (1951), Long-term Storage Capacity of Reservoirs, Transactions of the American Society of Civil Engineers 116:770-799.

Hurvich, H., R. Deo and J. Brodsky (1998), The Mean Squared Error of Geweke and Porter-Hudak's Estimator of Memory Parameter of a Long Memory Time Series, Journal of Time Seires Analysis 19:19-46.

Hyung, N. and P.H. Franses (2001), Structural Breaks and Long Memory in US Inflation Rates: Do They Matter for Forecasting?, Working Paper, Erasmus University.

Inoue, A. and Kilian, L. (2003), On the Selection of Forecasting Models, Working Paper, University of Michigan.

Kim, C. and C. Nelson (1999), State space models with regime switching: classical and Gibbs-sampling approaches with applications (MIT press, Cambridge).

Künsch, H.R. (1987), Statistical Aspects of Self-similar Processes, in: Y. Prohorov and V.V. Sasanov, eds., Proceedings of the first World Congress of the Bernoulli Society (VNU Science Press, Utrecht).

Lee, D. and P. Schmidt (1996), On the Power of the KPSS Test of Stationarity Against Fractionally-Integrated Alternatives, Journal of Econometrics 73:285-302.

Leybourne, S., D. Harris and B. McCabe (2003), A Robust Test for Short Memory, Working Paper, University of Nottingham.

Lo, A. (1991), Long-Term Memory in Stock Market Prices, Econometrica 59:1279-1313.

McCracken, M.W. (1999), Asymptotics for Out of Sample Tests of Causality, Working Paper, Louisiana State University.

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

Phillips, P.C.B. (1987), Time Series Regression with a Unit Root, Econometrica 55:277-301.

Robinson, P. (1995a), Log-Periodogram Regression of Time Series with Long Range Dependence, The Annals of Statistics 23:048- 1072.

Robinson, P. (1995b), Gaussian Semiparametric Estimation of Long Range Dependence, The Annals of Statistics 23:1630- 1661.

Robinson, P. (2003), Time Series With Long Memory (Oxford University Press, Oxford).

Rossi, B. (2003), Testing Long-Horizon Predictive Ability with High Persistence, and the Meese-Rogoff Puzzle, Working Paper, Duke University.

Shimotsu, K. and P.C.B. Phillips (2002), Exact Local Whittle Estimation of Fractional Integration, Working Paper, University of Essex.

Sowell, F.B. (1992a), Maximum Likelihood Estimation of Stationary Univariate Fractionally Integrated Time Series Models, Journal of Econometrics 53:165-188.

Sowell, F.B. (1992b), Modelling Long-Run Behavior with the Fractional ARIMA Model, Journal of Monetary Economics 29:277-302.

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

Swanson, N.R. and H. White (1995), A Model Selection Approach to Assessing the Information in the Term Structure Using Linear Models and Artificial Neural Networks, Journal of Business and Economic Statistics 13:265-279.

Swanson, N.R. and H. White (1997), A Model Selection Approach to Real-Time Macroeconomic Forecasting Using Linear Models and Artificial Neural Networks, Review of Economics and Statistics 79:540-550.

Taqqu, M. and V. Teverovsky (1995), Estimators for Long Range Dependence: An Emperical Study, Fractals 3:785-789.

Taqqu, M. and V. Teverovsky (1997), Robustness of Whittle-type Estimators for Time Series with Long-range Dependence, Stochastic Models 13:723-757.

van Dijk, D. and P.H. Franses (1999), Modeling Multiple Regimes in the Business Cycle, Macroeconomic Dynamics 3:311-340.

van Dijk, D., P.H. Franses and A. Lucas (1999), Testing for Smooth Transition Nonlinearity in the Presence of Additive Outliers, Journal of Business & Economic Statistics 17:217-235.

van Dijk, D., P.H. Franses and R. Paap (2002), A Nonlinear Long Memory Model, with an Application to US Unemployment, Journal of Econometrics 110:135-165.

van Dijk, D., T. Teräsvirta and P.H. Franses (2002), Smooth Transition Autoregressive Models a Survey of Recent Developments, Econometric Reviews 21:1-47.

West, K. (1996), Asymptotic Inference About Predictive Ability, Econometrica 64:1067-1084.

White, H. (2000), A Reality Check for Data Snooping, Econometrica 68:1097-1126.

Table 1.a: Analysis of U.S. S&P500 Daily Absolute Returns $^{(*)}$

Estimation Scheme and Forecast Horizon	ARFIMA Model	d	non-ARFIMA Model	DM$_1$	ENC-t$_1$	STAR Model	DM$_2$	ENC-t$_2$
1 day ahead, recursive	WHI (1,1)	0.41 (0.0001)	ARMA (4,2)	-1.18	0.47	LSTAR (8)	-5.14	-0.84
5 day ahead, recursive	GPH (1,2)	0.57 (0.0011)	ARMA (4,2)	-0.71	1.75	LSTAR (8)	-4.16	0.07
20 day ahead, recursive	GPH (1,2)	0.57 (0.0011)	ARMA (4,2)	-0.68	2.91	LSTAR (8)	-6.00	0.52
1 day ahead, rolling	RR (1,1)	0.25 (0.0009)	ARMA (4,2)	2.02	4.56	LSTAR (8)	-4.17	1.01
5 day ahead, rolling	GPH (1,2)	0.55 (0.0044)	ARMA (4,2)	-2.28	0.26	LSTAR (8)	-6.50	1.18
20 day ahead, rolling	GPH (1,2)	0.55 (0.0044)	ARMA (4,2)	-2.44	0.79	LSTAR (8)	-7.30	0.48

$^{(*)}$ Notes: Models are estimated as discussed above, and model acronyms used are as outlined in Section 3. Data used in this table correspond to the extended series of those used in Ding, Granger, and Engle (1993). Reported results are based on predictive evaluation using the second half of the sample. The 'ARFIMA Model', 'non-ARFIMA Model', and 'STAR Model' are the models chosen using MSFEs associated with ex ante recursive (rolling) estimation and 1, 5 and 20 step ahead prediction of the different model/lag combinations using the first 50% of sample. The remaining 50% of sample is used for subsequent ex ante prediction, the results of which are reported in the table. Further details are given in Section 4. In the second column, entries in brackets indicate the number of AR and MA lags chosen for the ARFIMA model. The third column lists the average (and standard error) of the estimated values of d across the entire ex ante sample, thus these entries are conditional on the selected ARFIMA model. The fourth column reports the chosen "best" non-ARFIMA model, entries in brackets indicate the number of AR and MA lags thus chosen. The seventh column reports the chosen "best" STAR model, and entry in brackets indicate the number of AR lags chosen. Diebold and Mariano (DM) test statistics are based on MSFE loss, and application of the test assumes that parameter estimation error vanishes and that the standard normal limiting distribution is asymptotically valid, as discussed in Section 3. Negative statistic values for DM statistics indicate that the point MSFE associated with the ARFIMA model is lower than that for the non-ARFIMA model, and the null hypothesis of the test is that of equal predictive accuracy. ENC-t statistics reported in the sixth column of the table, are normally distributed for $h = 1$, and correspond to the null hypothesis that the ARFIMA model encompasses the non ARFIMA model. DM$_1$ and ENC-t$_1$ compare the "best" ARFIMA model with the "best" non-ARFIMA model, while DM$_2$ and ENC-t$_2$ compare the "best" ARFIMA model with the "best" STAR model.

Table 1.b: Analysis of U.S. S&P500 Daily Absolute Returns $^{(*)}$

Estimation Scheme and Forecast Horizon	ARFIMA Model	d	non-ARFIMA Model	DM$_1$	ENC-t$_1$	STAR Model	DM$_2$	ENC-t$_2$
1 day ahead, recursive	GPH-OPT (1,2)	0.55 (0.0227)	ARMA (4,2)	-0.51	1.08	LSTAR (8)	-5.46	-0.77
5 day ahead, recursive	WHI-OPT (1,2)	0.44 (0.0113)	ARMA (4,2)	-0.15	2.38	LSTAR (8)	-5.11	-0.05
20 day ahead, recursive	GPH-OPT (1,2)	0.55 (0.0227)	ARMA (4,2)	-1.41	1.83	LSTAR (8)	-6.89	-2.27
1 day ahead, rolling	RR (1,1)	0.25 (0.0009)	ARMA (4,2)	2.02	4.56	LSTAR (8)	-4.17	1.01
5 day ahead, rolling	WHI-OPT (1,2)	0.42 (0.0196)	ARMA (4,2)	-1.61	1.06	LSTAR (8)	-6.42	1.21
20 day ahead, rolling	GPH (1,2)	0.55 (0.0044)	ARMA (4,2)	-2.44	0.79	LSTAR (8)	-7.30	0.48

$^{(*)}$ Notes: See notes to Tables 1.a. The set of ARFIMA models now also includes those where the GPH and WHI estimation schemes have bandwidths selected using the data driven procedure of Henry (2001), as discussed above. GPH-OPT and WHI-OPT refer to these two additional cases.

Table 1.c: Analysis of U.S. S&P500 Daily Absolute Returns: Encompassing Regressions and Pooled Median Forecasts $^{(*)}$

Estimation Scheme	Encompassing Regression				DM$_3$	ENC-t$_3$
and Forecast Horizon	with non-ARFIMA		with STAR			
	β_{ARFIMA}	$\beta_{non-ARFIMA}$	β_{ARFIMA}	β_{STAR}		
1 day ahead, recursive	0.78 (0.30)	0.17 (0.29)	1.08 (0.08)	-0.13 (0.07)	-0.35	0.42
5 day ahead, recursive	0.65 (0.16)	0.31 (0.16)	0.96 (0.06)	0.01 (0.14)	-0.52	1.06
20 day ahead, recursive	0.62 (0.10)	0.32 (0.11)	1.00 (0.06)	-0.13 (0.08)	0.41	2.11
1 day ahead, rolling	0.79 (0.12)	O.17 (0.13)	1.13 (0.07)	-0.21 (0.09)	-0.40	0.66
5 day ahead, rolling	1.02 (0.16)	-0.05 (0.18)	0.98 (0.05)	0.04 (0.06)	-0.71	1.66
20 day ahead, rolling	0.89 (0.13)	0.04 (0.15)	0.98 (0.03)	-0.35 (0.13)	-1.64	1.88

$^{(*)}$ Notes: See notes to Tables 1.a. β_{ARFIMA}, $\beta_{non-ARFIMA}$ and β_{STAR} refer to coefficients in the encompassing regressions, as discussed above. Standard errors of the coefficients are in parentheses. DM$_3$ and ENC-t$_3$ compare the "best" ARFIMA model with pooled forecasts constructed by taking the median value across the point forecasts from the linear non-ARFIMA and STAR models.

Table 2: Analysis of U.S. S&P500 Daily Absolute Returns, Pre WWII Recession $^{(*)}$

Estimation Scheme	ARFIMA	d	non-ARFIMA	DM_1	ENC-t_1	STAR	DM_2	ENC-t_2
and Forecast Horizon	Model		Model			Model		
1 day ahead, recursive	GPH (1,1)	0.62 (0.0028)	ARMA (1,1)	-0.24	1.00	LSTAR (4)	-1.36	0.46
5 day ahead, recursive	RR (1,1)	0.21 (0.0001)	ARMA (1,1)	-0.55	1.34	ESTAR (4)	-0.86	1.12
20 day ahead, recursive	RR (1,1)	0.21 (0.0001)	ARMA (1,1)	-0.26	1.40	LSTAR (4)	-0.64	0.81
1 day ahead, rolling	RR (1,1)	0.21 (0.0002)	ARMA (1,1)	-0.50	1.33	LSTAR (4)	-1.54	0.05
5 day ahead, rolling	RR (1,1)	0.21 (0.0002)	ARMA (1,1)	-0.01	1.96	LSTAR (4)	-1.18	0.80
20 day ahead, rolling	WHI (1,1)	0.62 (0.0397)	MA (4)	-1.40	1.46	LSTAR (4)	-1.27	1.47

$^{(*)}$ Notes: See notes to Tables 1.a. Data for this table correspond to the two pre World War II recessions from August 1929 till March 1933, and from May 1937 till June 1938. We have a total of 1,387 data points.

Table 3: Analysis of U.S. S&P500 Daily Absolute Returns, Pre WWII Expansion $^{(*)}$

Estimation Scheme	ARFIMA	d	non-ARFIMA	DM_1	ENC-t_1	STAR	DM_2	ENC-t_2
and Forecast Horizon	Model		Model			Model		
1 day ahead, recursive	RR (1,1)	0.25 (0.0001)	ARMA (1,1)	1.10	1.91	LSTAR (2)	-4.28	-2.24
5 day ahead, recursive	RR (1,1)	0.25 (0.0001)	ARMA (1,1)	0.56	1.86	LSTAR (2)	-4.12	-2.00
20 day ahead, recursive	WHI (1,1)	0.58 (0.0005)	ARMA (1,1)	-0.89	1.70	LSTAR (2)	-3.86	-0.38
1 day ahead, rolling	RR (1,1)	0.21 (0.0021)	ARMA (1,1)	0.57	1.83	LSTAR (2)	-4.86	-2.62
5 day ahead, rolling	RR (1,1)	0.21 (0.0021)	ARMA (1,1)	0.95	2.43	LSTAR (2)	-4.85	-3.54
20 day ahead, rolling	WHI (1,1)	0.62 (0.0004)	ARMA (1,1)	-3.63	-0.81	LSTAR (2)	-4.05	-0.81

$^{(*)}$ Notes: See notes to Tables 1.a. Data for this table correspond to the pre World War II expansion period.We have a total of 1,707 data points.

Table 4: Analysis of U.S. S&P500 Daily Absolute Returns, Post WWII Recession $^{(*)}$

Estimation Scheme	ARFIMA	d	non-ARFIMA	DM_1	ENC-t_1	STAR	DM_2	ENC-t_2
and Forecast Horizon	Model		Model			Model		
1 day ahead, recursive	WHI (2,2)	0.49 (0.0001)	ARMA (1,1)	-0.63	0.65	LSTAR (2)	-4.90	-0.04
5 day ahead, recursive	WHI (2,2)	0.49 (0.0001)	ARMA (1,1)	-2.07	-0.82	LSTAR (2)	-3.17	1.68
20 day ahead, recursive	WHI (2,2)	0.49 (0.0001)	ARMA (1,1)	-3.27	-1.46	LSTAR (2)	-3.20	0.76
1 day ahead, rolling	WHI (2,2)	0.54 (0.0008)	ARMA (1,1)	1.11	2.07	LSTAR (2)	-4.37	0.86
5 day ahead, rolling	WHI (2,2)	0.54 (0.0008)	ARMA (1,1)	-0.09	1.25	LSTAR (2)	-3.98	1.04
20 day ahead, rolling	RR (1,1)	0.21 (0.0015)	ARMA (1,1)	-0.60	1.27	LSTAR (2)	-3.05	-1.89

$^{(*)}$ Notes: See notes to Tables 1.a. Data for this table correspond to the post World War II recessions. We have a total of 2,401 data points.

Table 5: Analysis of U.S. S&P500 Daily Absolute Returns, Post WWII Expansion $^{(*)}$

Estimation Scheme	ARFIMA	d	non-ARFIMA	DM_1	ENC-t_1	STAR	DM_2	ENC-t_2
and Forecast Horizon	Model		Model			Model		
1 day ahead, recursive	WHI (1,1)	0.38 (0.0002)	ARMA (1,1)	-1.04	0.50	ESTAR (2)	-7.69	-0.29
5 day ahead, recursive	WHI (1,1)	0.38 (0.0002)	ARMA (1,1)	-3.09	-1.09	ESTAR (2)	-5.19	-0.38
20 day ahead, recursive	WHI (1,1)	0.38 (0.0002)	ARMA (1,1)	-5.49	-1.21	LSTAR (2)	-5.87	-1.28
1 day ahead, rolling	WHI (1,1)	0.40 (0.0009)	ARMA (1,1)	-0.34	1.01	LSTAR (2)	-7.45	-0.23
5 day ahead, rolling	WHI (1,1)	0.40 (0.0009)	ARMA (1,1)	-2.31	-0.48	ESTAR (2)	-4.69	0.35
20 day ahead, rolling	WHI (1,1)	0.40 (0.0009)	ARMA (1,1)	-4.57	-0.62	ESTAR (2)	-5.44	-0.40

$^{(*)}$ Notes: See notes to Tables 1.a. Data for this table correspond to the post World War II expansion. We have a total of 12,141 data points.

Table 6: Analysis of U.S. S&P500 Daily Absolute Returns, The period of oil shocks 1973-1982 (*)

Estimation Scheme	ARFIMA	d	non-ARFIMA	DM_1	ENC-t_1	STAR	DM_2	ENC-t_2
and Forecast Horizon	Model		Model			Model		
1 day ahead, recursive	WHI (1,1)	0.51 (0.0021)	ARMA (2,2)	-0.92	0.84	ESTAR (2)	-2.82	1.13
5 day ahead, recursive	WHI (1,1)	0.51 (0.0021)	ARMA (2,2)	-0.74	0.94	ESTAR (2)	-2.90	1.28
20 day ahead, recursive	WHI (1,1)	0.51 (0.0021)	ARMA (2,2)	-0.52	1.26	LSTAR (2)	-1.38	1.40
1 day ahead, rolling	WHI (1,1)	0.50 (0.0030)	ARMA (2,2)	-1.71	0.12	LSTAR (2)	-3.60	0.61
5 day ahead, rolling	WHI (1,1)	0.50 (0.0030)	ARMA (2,2)	-0.77	0.97	ESTAR (2)	-3.78	0.47
20 day ahead, rolling	WHI (1,1)	0.50 (0.0030)	ARMA (2,2)	-1.38	0.07	ESTAR (2)	-2.37	0.91

(*) Notes: See notes to Tables 1.a. Data for this table correspond to the period of oil shocks. The starting date is October 17, 1973 when Arab members of OPEC restricted shipment of petroleum. While the ending data for this period corresponds to the 1982 reduction in oil prices by OPEC. We have a total of 2,273 data points.

Table 7: Analysis of post 1982 S&P500 Daily Absolute Returns (*)

Estimation Scheme	ARFIMA	d	non-ARFIMA	DM_1	ENC-t_1	STAR	DM_2	ENC-t_2
and Forecast Horizon	Model		Model			Model		
1 day ahead, recursive	RR (2,2)	0.21 (0.0004)	ARMA (1,1)	1.77	3.76	LSTAR (3)	-4.23	-0.01
5 day ahead, recursive	GPH (2,3)	0.64 (0.0009)	ARMA (1,1)	-1.29	1.67	LSTAR (3)	-4.29	-0.35
20 day ahead, recursive	GPH (2,3)	0.64 (0.0009)	ARMA (1,1)	-2.80	0.32	LSTAR (3)	-4.20	0.44
1 day ahead, rolling	RR (2,2)	0.24 (0.0001)	ARMA (1,1)	0.58	3.18	ESTAR (3)	-6.41	-2.30
5 day ahead, rolling	RR (2,2)	0.24 (0.0001)	ARMA (1,1)	-0.51	2.18	LSTAR (3)	-5.41	-3.02
20 day ahead, rolling	WHI (1,2)	0.46 (0.0003)	ARMA (1,1)	-4.37	-2.51	ESTAR (3)	-5.01	0.40

(*) Notes: See notes to Tables 1.a. Data for this table correspond to the post 1982 period. We have a total of 5,369 data points.