Journal of Forecasting
J. Forecast. **26**, 95–111 (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/for.1012

Optimal Prediction with Nonstationary ARFIMA Model

MOHAMED BOUTAHAR*
GREQAM, and Department of Mathematics, Luminy Faculty of Sciences, Marseille, France

ABSTRACT

We propose two methods to predict nonstationary long-memory time series. In the first one we estimate the long-range dependent parameter d by using tapered data; we then take the nonstationary fractional filter to obtain stationary and short-memory time series. In the second method, we take successive differences to obtain a stationary but possibly long-memory time series. For the two methods the forecasts are based on those obtained from the stationary components. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS ARFIMA model; long memory; nonstationary processes; optimal prediction

INTRODUCTION

Prediction of time series can be achieved by using the Wiener–Kolmogorov approach (see Bhansali and Kokoszka, 2001). Earlier, Box and Jenkins (1976) applied this theory to the well-known ARIMA(p, d, q) process (d is an integer) to obtain optimal predictions for nonstationary processes. In this paper, we extend this approach to the ARFIMA(p, d, q) where d is any real number with $d > -\frac{1}{2}$. We give optimum linear predictors by making use of two methods. In the first one, called Method 1 and presented in the next section, we use the raw data (we do not take differences) to estimate a nonstationary fractional filter. We then apply this filter to obtain stationary and short-memory time series. This method belongs to the philosophy modelling of Parzen (1982), who called it ARARMA modelling. However, Parzen (1982) did not use the fractionally integrated approach. In the second method, called Method 2 and presented in the third section, we take successive differences to obtain a stationary but possibly long-memory time series; this is, of course, the classical approach of Box and Jenkins (1976).

For the two methods we use the predictors, obtained from the stationary process by using any classical method, to compute those of the initial nonstationary process. We give also the mean squared errors of the h-step predictors. Moreover, we show how these two methods can make use of the innovation algorithm applied to the stationary short-memory component which is assumed to be an ARMA process. In the fourth section we perform a Monte Carlo study to compare the two

* Correspondence to: Mohamed Boutahar, Department of Mathematics, Luminy Faculty of Sciences, 163 Av. de Luminy, 13288 Marseille Cedex 9, France. E-mail: boutahar@univmed.fr

methods. Our conclusion is that Method 1 is slightly superior to Method 2. This conclusion is also confirmed in the fifth section when we apply them to real data.

METHOD 1: PREDICTION WITHOUT DIFFERENCING

Consider the model

$$\phi(B)(1-B)^d y_t = \theta(B)u_t, \quad u_t \sim \text{i.i.d.}(0, \sigma_u^2) \tag{1}$$

where

$$\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p, \quad \theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$$

are stationary AR and invertible MA operators, $(1 - B)^d$ is the fractional filter given by the binomial expansion

$$(1-B)^d = \sum_{j=0}^{\infty} \delta_j(d) B^j, \quad \delta_j(d) = \frac{j-d-1}{j}\delta_{j-1}(d), \quad \delta_0(d) = 1 \tag{2}$$

$d \in \mathbb{R}$ with $d > -1/2$, and B is the backward shift operator, i.e. $By_t = y_{t-1}$.

The estimation of d can be made by the log-periodogram method of Geweke and Porter-Hudak (1983), but if $d \geqslant 1/2$, then we apply a taper of order $p_1 \geqslant s + 1$ to the data, where $s = [d + 1/2]$, ([.] is the integer part of the argument) (see Velasco, 1999a, theorem 7). The values of p and q can be unknown. Also Beran *et al.* (1998) consider a similar model with $q = 0$, and give a version of Akaike information criterion (AIC) for determining an appropriate autoregressive order when d and the autoregressive parameters are estimated simultaneously by a maximum likelihood procedure (Beran, 1995). In this paper we will consider the ARFIMA(p, d, q) model with possibly moving average component, $q \neq 0$. Moreover we do not use the same model selection procedure as Beran *et al.* (1998), but we first filter out the long-memory stationary or nonstationary component and afterward, we use standard criteria to select the orders (p, q) of the short-memory component.

Recall that a sequence of data taper (h_t, $1 \leq t \leq T$) is of order p_1 if the following two conditions are satisfied:

1. $\sum_{t=1}^{T} h_t^2 = b_T T, \quad 0 < b_T < \infty.$
2. For $N = \dfrac{T}{p_1}$ (which is assumed to be an integer), the Dirichlet Kernel $D_{p_1}(\lambda)$ satisfies

$$D_{p_1}(\lambda) = \sum_{t=1}^{T} h_t e^{i\lambda t} = \frac{a(\lambda)}{T^{p_1-1}} \left(\frac{\sin \dfrac{T\lambda}{2p_1}}{\sin \dfrac{\lambda}{2}} \right)^{p_1}$$

where $a(\lambda)$ is a complex function, whose modulus is bounded and bounded away from zero, with $p_1 - 1$ derivatives, all bounded in modulus as T increases for $\lambda \in [-\pi, \pi]$.

DOI: 10.1002/for

Assuming d is known, we will define an h-step prediction of y_{t+h}. Denote by $\overline{Sp}\ \{y_t, t \in I\}$ the closed span of the subset $\{y_t, t \in I\}$ in the Hilbert space H; it is the smallest closed subspace of H which contains each element y_t, $t \in I$. Let

$$x_t = (1-B)^d y_t \tag{3}$$

then x_t is a stationary and invertible ARMA (p, q) process

$$\phi(B)x_t = \theta(B)u_t, \quad u_t \sim \text{i.i.d.}\,(0, \sigma_u^2) \tag{4}$$

Assume that $(y_j, j \le 0)$ is uncorrelated with x_t, $t \geqslant 1$. The optimum linear predictor $\hat{y}_{t+h}$ of y_{t+h}, $h > 0$, in terms of $(y_s, s \le t)$ is given by

$$\hat{y}_{t+h} = P_{S_t}(y_{t+h}), \text{ where } S_t = \overline{Sp}\{y_j, j \le t\}$$

and P_{S_t} is the projection of H onto S_t.

The mean squared error of the h-step predictor is

$$\sigma^2(h) = E(y_{t+h} - \hat{y}_{t+h})^2.$$

Let y^*_{t+1} denote the one-step predictor of y_{t+1}, i.e. $y^*_{t+1} = P_{S_t}(y_{t+1})$,

$$\phi^*(z) = (1-z)^d \phi(z) = 1 - \sum_{j=1}^{\infty} \phi_j^*(d) z^j \tag{5}$$

$$\left(1 - \sum_{j=1}^{p+h-1} \psi_j z^j\right)^{-1} = 1 + \sum_{j=1}^{\infty} \delta_j^*(d) z^j \tag{6}$$

where

$$\psi(z) = \delta(z)\phi(z) = 1 - \sum_{j=1}^{p+h-1} \psi_j z^j, \quad \delta(z) = 1 + \sum_{j=1}^{h-1} \delta_j(d) z^j \tag{7}$$

and $\delta_j(d)$ are given by (2).

Remark: Since $\phi_k = 0$ for all $k > p$, it follows that

$$\phi_j^*(d) = \sum_{k=0}^{\min(j,p)} \phi_k \delta_{j-k}(d) \tag{8}$$

Theorem 1: *Assume that* $t > m = \max(p, q)$. The predictors $\hat{y}_{t+h}$ *and their mean squared error are given by*

$$\hat{y}_{t+h} = \sum_{j=1}^{\infty} \phi_j^*(d)\hat{y}_{t+h-j} + \sum_{j=h}^{q} \theta_{t+h-1,j}\left(y_{t+h-j} - y^*_{t+h-j}\right) \tag{9}$$

DOI: 10.1002/for

and

$$\sigma^2(h) = \sum_{j=0}^{h-1} \left(\sum_{r=0}^{j} \delta_j^*(d) \theta_{t+h-r-1,\, j-r} \right)^2 v_{t+h-j-1} \tag{10}$$

where $\theta_{i,j}$ and v_t *are obtained from the innovation algorithm applied to the ARMA(p, q) process* x_t *defined in (4) (see Brockwell and Davis, 1991, equations 5.2.16 and 5.3.5).*

Proof: From (3) we have

$$y_t = x_t - \sum_{j=1}^{\infty} \delta_j(d) y_{t-j} \tag{11}$$

Since $(y_j, j \leq 0) \perp x_t$, $\forall\ t \geqslant 1$, and $\overline{Sp}\ \{y_j, j \leq t\} = \overline{Sp}\ \{y_j, j \leq 0, x_1, \ldots, x_t\}$ we have

$$\hat{y}_{t+h} = P_{S_t}(y_{t+h}) = P_t(x_{t+h}) - \sum_{j=1}^{\infty} \delta_j(d) P_{S_t}(y_{t+h-j}) \tag{12}$$

where

$$P_t(x_{t+h}) = P_{\overline{Sp}(x_1, \ldots, x_t)}(x_{t+h})$$

Moreover

$$\begin{aligned} y_{t+1}^* = P_{S_t}(y_{t+1}) &= P_t(x_{t+1}) - \sum_{j=1}^{\infty} \delta_j(d) P_{S_t}(y_{t+1-j}) \\ &= x_{t+1}^* - \sum_{j=1}^{\infty} \delta_j(d) y_{t+1-j} \qquad (13) \\ &= x_{t+1}^* + y_{t+1} - x_{t+1}, \ \text{from (11)} \qquad (14) \end{aligned}$$

then

$$y_{t+1} - y_{t+1}^* = x_{t+1} - x_{t+1}^* \quad \forall\ t > m \tag{15}$$

This seems to be an important result, which means that the one-step prediction error of the non-stationary and long-memory ARFIMA process y_t is the same as that of the stationary and short-memory ARMA process x_t.

Now, the optimum linear predictors $\hat{x}_{t+h}$ can be obtained from the stationary theory. We have from Brockwell and Davis (1991) that

$$x_{t+1}^* = \sum_{i=1}^{p} \phi_i x_{t+1-i} + \sum_{j=1}^{q} \theta_{t,j} (x_{t+1-j} - x_{t+1-j}^*) \quad \forall\ t > m \tag{16}$$

DOI: 10.1002/for

$$\hat{x}_{t+h} = \sum_{i=1}^{p} \phi_i \hat{x}_{t+h-i} + \sum_{j=h}^{q} \theta_{t+h-1,j}(x_{t+h-j} - x^*_{t+h-j}) \quad \forall\, t > m \tag{17}$$

Denote by $\hat{x}_{t-j} = B^j \hat{x}_t$, then combining (15) and (17) we obtain

$$\phi(B)\hat{x}_{t+h} = \sum_{j=h}^{q} \theta_{t+h-1,j}(y_{t+h-j} - y^*_{t+h-j}) \tag{18}$$

The equality (12) can be written as

$$(1-B)^d \hat{y}_{t+h} = \hat{x}_{t+h} \tag{19}$$

Equations (18) and (19) imply that

$$(1-B)^d \phi(B)\hat{y}_{t+h} = \sum_{j=h}^{q} \theta_{t+h-1,j}(y_{t+h-j} - y^*_{t+h-j})$$

then (9) holds.

To prove (10) we have from (12)

$$\hat{y}_{t+h} = \hat{x}_{t+h} - \sum_{j=1}^{h-1} \delta_j(d)\hat{y}_{t+h-j} - \sum_{j=h}^{\infty} \delta_j(d) y_{t+h-j} \tag{20}$$

and from (11)

$$y_{t+h} = x_{t+h} - \sum_{j=1}^{h-1} \delta_j(d) y_{t+h-j} - \sum_{j=h}^{\infty} \delta_j(d) y_{t+h-j} \tag{21}$$

Subtracting (20) from (21) gives

$$y_{t+h} - \hat{y}_{t+h} = x_{t+h} - \hat{x}_{t+h} - \sum_{j=1}^{h-1} \delta_j(d)(y_{t+h-j} - \hat{y}_{t+h-j})$$

or

$$\delta(B)(y_{t+h} - \hat{y}_{t+h}) = x_{t+h} - \hat{x}_{t+h}$$

where $\delta(z)$ is the polynomial defined in (7). The equalities (16)–(17) imply that

$$\phi(B)(x_{t+h} - \hat{x}_{t+h}) = \sum_{j=0}^{h-1} \theta_{t+h-1,j}(x_{t+h-j} - x^*_{t+h-j})$$

DOI: 10.1002/for

hence we obtain from the last two equalities

$$\delta(B)\phi(B)(y_{t+h}-\hat{y}_{t+h})=\psi(B)(y_{t+h}-\hat{y}_{t+h})=\sum_{j=0}^{h-1}\theta_{t+h-1,j}(x_{t+h-j}-x^{*}_{t+h-j})$$

which implies that

$$\mathbf{\Psi}\begin{pmatrix} y_{t+1}-\hat{y}_{t+1} \\ \vdots \\ y_{t+h}-\hat{y}_{t+h} \end{pmatrix}=\mathbf{\Theta}\begin{pmatrix} x_{t+1}-x^{*}_{t+1} \\ \vdots \\ x_{t+h}-x^{*}_{t+h} \end{pmatrix} \tag{22}$$

where $\mathbf{\Psi}$ and $\mathbf{\Theta}$ are (h, h)-matrices such that

$$\mathbf{\Psi}=-(\psi_{i-j}),\psi_0=-1,\psi_j=0 \quad \text{if} \quad j>p+h-1 \quad \text{or} \quad j<0 \tag{23}$$

and

$$\mathbf{\Theta}=(\theta_{t+i-1,i-j}),\theta_{t,0}=1,\theta_{t,j}=0 \quad \text{if} \quad j>q \quad \text{or} \quad j<0 \tag{24}$$

Define

$$\mathbf{V}=\operatorname{diag}(\upsilon_t,\ldots,\upsilon_{t+h-1}),e_h=(0,\ldots,0,1)' \text{ an h - dimensional vector} \tag{25}$$

where $v_t = E(x_{t+1} - x^{*}_{t+1})^2$ is the mean squared error of the one-step predictor of the ARMA(p, q) process x_t. We then obtain from (22) that

$$\sigma^2(h) = e_h'\mathbf{\Psi}^{-1}\mathbf{\Theta}\mathbf{V}\mathbf{\Theta}'\mathbf{\Psi}'^{-1}e_h.$$

Straightforward calculation leads to (10). □

METHOD 2: PREDICTION USING DIFFERENTIATED DATA

In this section, we show how to obtain predictors for the nonstationary ARFIMA process from those computed for the stationary ARFIMA process. There are many methods to obtain forecasts for a stationary series; see Bhansali and Kokoszka (2003) for some of them. In this paper, we make use of the results obtained by Peiris and Perera (1988).

Let

$$s=\left[d+\frac{1}{2}\right],\quad d_1=d-s$$

and assume that s is known.

Define $z_t = (1 - B)^s y_t$, then from (1) z_t is a stationary and invertible ARFIMA(p, d_1, q) process

$$\phi(B)(1-B)^{d_1}z_t=\theta(B)u_t,\quad u_t\sim \text{i.i.d.}\,(0,\sigma_u^2) \tag{26}$$

DOI: 10.1002/for

Let

$$x_t = (1-B)^{d_1} z_t \tag{27}$$

then x_t is a stationary and invertible ARMA(p, q) process

$$\phi(B)x_t = \theta(B)u_t, \quad u_t \sim \text{i.i.d.}\,(0, \sigma_u^2) \tag{28}$$

We assume that the vector $(y_{1-s}, \ldots, y_0)$ is uncorrelated with z_t, $t \geqslant 1$, and let

$$\alpha_j = \frac{s!}{j!(s-j)!}(-1)^j, \quad 1 \leq j \leq s,$$

$$\chi_j = \sum_{k=1}^{\min(p,j)} \phi_k \chi_{j-k}, \quad \chi_0 = 1,$$

$$\gamma^*(z) = (1-z)^s \gamma(z) = 1 - \sum_{k=1}^{s+h-1} \gamma_k^*(d_1) z^k \tag{29}$$

$$\gamma(z) = 1 + \sum_{j=1}^{h-1} \gamma_j(d_1) z^j, \quad \gamma_j(d) = \sum_{i=0}^{j} \delta_i(d_1) \tag{30}$$

where $\delta_i(d_1)$ are given by (2):

$$\xi_i(d_1) = \sum_{\mathrm{k}=1}^{\min(s+h-1,i)} \gamma_k^*(d_1)\xi_{i-k}(d_1), \quad 1 \leq i \leq h-1, \quad \xi_0(d_1) = 1$$

Theorem 2: *Assume that $t > m = max(p, q)$. The predictors $\hat{y}_{t+h}$ and their mean squared error are given by*

$$\hat{y}_{t+h} = -\sum_{j=1}^{s} \alpha_j \hat{y}_{t+h-j} - \sum_{j=1}^{h-1} \sum_{i=0}^{j} \delta_i(d_1) \hat{z}_{t+h-j} - \sum_{j=0}^{\infty} \sum_{i=j+1}^{h+j} \delta_i(d_1) z_{t-j} + \sum_{j=1}^{h} \hat{x}_{t+j} \tag{31}$$

and

$$\sigma^2(h) = \sum_{j=1}^{h} \left(\sum_{l=0}^{h} \theta_{t+l-1,l-j} \sum_{k=0}^{h-l} \chi_k \left(\sum_{i=0}^{h-k-1} \xi_i(d_1) \right) \right)^2 \upsilon_{t+j-1} \tag{32}$$

where $\theta_{i,j}$ and υ_t are obtained from the innovation algorithm applied to the ARMA(p, q) process x_t defined in (28) (see Brockwell and Davis, 1991, equations 5.2.16 and 5.3.5).

Proof: Since $z_t = (1-B)^s y_t$ we have

$$y_{t+h} = z_{t+h} - \sum_{j=1}^{s} \alpha_j y_{t+h-j} \tag{33}$$

DOI: 10.1002/for

then the same argument as used above (equalities (11) and (12)) leads to

$$\hat{y}_{t+h} = \hat{z}_{t+h} - \sum_{j=1}^{s} \alpha_j \hat{y}_{t+h-j} \tag{34}$$

Theorem 3.1 of Peiris and Perera (1988) gives the predictors $\hat{z}_{t+h}$

$$\hat{z}_{t+h} = -\sum_{j=1}^{h-1}\sum_{i=0}^{j} \delta_i(d_1)\hat{z}_{t+h-j} - \sum_{j=0}^{\infty}\sum_{i=j+1}^{h+j} \delta_i(d_1) z_{t-j} + \sum_{j=1}^{h} \hat{x}_{t+j} \tag{35}$$

and consequently the equality (31) holds.
To prove (32), we have from (26) and (27)

$$z_{t+h} = -\sum_{j=1}^{h-1}\sum_{i=0}^{j} \delta_i(d_1) z_{t+h-j} - \sum_{j=0}^{\infty}\sum_{i=j+1}^{h+j} \delta_i(d_1) z_{t-j} + \sum_{j=1}^{h} x_{t+j} \tag{36}$$

Subtracting (34) from (33) and (35) from (36) we obtain

$$(1-B)^s (y_{t+h} - \hat{y}_{t+h}) = z_{t+h} - \hat{z}_{t+h} \tag{37}$$

$$\gamma(B)(z_{t+h} - \hat{z}_{t+h}) = \sum_{j=1}^{h} (x_{t+j} - \hat{x}_{t+j}) \tag{38}$$

Therefore

$$\gamma^*(B) = (1-B)^s \gamma(B)(y_{t+h} - \hat{y}_{t+h}) = \sum_{j=1}^{h} (x_{t+j} - \hat{x}_{t+j}) \tag{39}$$

This equality implies that

$$\mathbf{\Gamma}(d_1) \begin{pmatrix} y_{t+1} - \hat{y}_{t+1} \\ \vdots \\ y_{t+h} - \hat{y}_{t+h} \end{pmatrix} = \mathbf{T} \begin{pmatrix} x_{t+1} - \hat{x}_{t+1} \\ \vdots \\ x_{t+h} - \hat{x}_{t+h} \end{pmatrix} \tag{40}$$

where

$$\mathbf{\Gamma}(d_1) = -(\gamma^*_{i-j}(d_1)),\ \gamma^*_0(d_1) = 1,\ \gamma^*_j(d_1) = 0 \text{ if } j > s + h - 1 \text{ or } j < 0$$

$$\mathbf{T} = (t_{i,j}),\ t_{i,j} = 1,\ t_{i,j} = 0 \text{ if } i < j$$

Since x_t is an ARMA(p, q), it follows that

$$\mathbf{\Phi} \begin{pmatrix} x_{t+1} - \hat{x}_{t+1} \\ \vdots \\ x_{t+h} - \hat{x}_{t+h} \end{pmatrix} = \mathbf{\Theta} \begin{pmatrix} x_{t+1} - x^*_{t+1} \\ \vdots \\ x_{t+h} - x^*_{t+h} \end{pmatrix} \tag{41}$$

DOI: 10.1002/for

where

$$\mathbf{\Phi} = -(\phi_{i-j}),\ \phi_0 = -1,\ \phi_j = 0 \text{ if } j > p \text{ or } j < 0$$

and $\mathbf{\Theta}$ is given by (24), where the $\theta_{i,j}$ are obtained from the innovation algorithm applied to the ARMA(p, q) process x_t defined in (28).

The equalities (40) and (41) imply that

$$\begin{pmatrix} y_{t+1} - \hat{y}_{t+1} \\ \vdots \\ y_{t+h} - \hat{y}_{t+h} \end{pmatrix} = \mathbf{\Gamma}^{-1}(d_1)\mathbf{T}\mathbf{\Phi}^{-1}\Theta \begin{pmatrix} x_{t+1} - x^*_{t+1} \\ \vdots \\ x_{t+h} - x^*_{t+h} \end{pmatrix}$$

which implies that

$$\sigma^2(h) = e_h'\mathbf{\Gamma}^{-1}(d_1)\mathbf{T}\mathbf{\Phi}^{-1}\mathbf{\Theta}\mathbf{V}\mathbf{\Theta}'\mathbf{\Phi}'^{-1}\mathbf{T}'\mathbf{\Gamma}'^{-1}(d_1)e_h$$

where $\mathbf{V}$ and e_h are the same as (25). Straightforward calculation leads to (32). □

MONTE CARLO STUDY

We have generated times series driven by five different models as follows; in all cases the simulated generating noise was standard Gaussian.

Model 1: ARFIMA(0,1.4,0), a simple persistent nonstationary long-memory process with all correlation coefficients positive.

Model 2: ARFIMA(1,1.4,0), with $\phi_1 = 0.5$, a process with persistent nonstationary long-memory component and a short-memory stationary autoregressive component.

Model 3: ARFIMA(0,1.4,1), with $\theta_1 = -0.7$, a process with persistent nonstationary long-memory component and a short-memory stationary moving average component.

Model 4: ARFIMA(1,1.4,1), with $\phi_1 = 0.5$, $\theta_1 = -0.8$, a process with persistent nonstationary long-memory component and an ARMA(1,1) short-memory stationary component.

Model 5: Nonstationary F-EXP model defined by

$$(1 - B)^{1.4} y_t = x_t, \quad \log f_x(\lambda) = \frac{1}{2\pi} - \cos\lambda$$

The number of time series NR say, generated from each model was 1000. For a given simulated model and each $h = 1, \ldots, H(=20)$, let $\hat{y}_{jT,1}(h)$ denote the h-step forecast provided by Method 1 for the jth simulated time series from that model, $j = 1, \ldots,$ NR, and let $\hat{y}_{jT,2}(h)$ denote the corresponding h-step forecast according to Method 2. The simulated h-step mean square error of prediction SMSE_1 for Method 1 is given by

$$\mathrm{SMSE}_1(h) = \frac{1}{\mathrm{NR}} \sum_{j=1}^{\mathrm{NR}} (\hat{y}_{jT,1}(h) - y_{j,T+h})^2$$

DOI: 10.1002/for

where $y_{j,T+h}$ represents the hth out-of-sample observation after the last observation used for estimation at the jth simulation.

The proportionate change in the simulated h-step mean squared error of prediction for Method 1 relative to Method 2 is then given by

$$\mathrm{PSMSE}(h) = \frac{\mathrm{SMSE}_1(h) - \mathrm{SMSE}_2(h)}{\mathrm{SMSE}_2(h)}$$

The graph of PSMSE(h), $h = 1, \ldots, H$ measures the change in the simulated mean squared error of prediction when using Method 1 in preference to Method 2. If for a given value of h PSMSE(h) is positive then this implies that Method 2 has an advantage over Method 1 for the h-step forecasts. Bhansali and Kokoszka (2001) used a similar criterion.

In the stationary framework, forecasting with ARFIMA models can be made by many other methods; there are two kinds:

1. The one-stage method, such as the HR procedure—the time domain maximum likelihood method—of Haslett and Raftery (1989) and the FT procedure—the spectral domain Whittle likelihood method—suggested by Fox and Taqqu (1986).
2. The two-stage method in which at the first stage, a non-parametric estimator of d, is obtained by existing methods (the GPH method of Geweke and Porter-Hudak, 1983, the Gaussian semi-parametric method of Robinson, 1995b, etc.). At the second stage the long-memory component is filtered out and a standard ARMA model is fitted to the filtered series and forecasts computed.

A comparative study of the two kinds of methods was used by Crato and Ray (1996), who also compared different selection criteria such as the AIC, AICc and the SIC. For forecasting purpose, the conclusion of these authors is that no one method systematically dominated. Moreover, there is no theoretical results about the behaviour of the HR and FT procedures in nonstationary models. For this reason we will only use the two-stage method. To simplify the study at the second stage the short-memory component will be modelled by an AR instead of an ARMA. For the two methods, the following procedures are used:

Method 1

(i) We compute the GPH estimator $\hat{d}$ by using the Kolmogorov taper of order $p_1 > [d + 1/2]$ (here $p_1 = 3$).
(ii) We filter out the long-memory nonstationary component $\tilde{y}_t = (1-B)^{\hat{d}} y_t$.
(iii) We use the AIC criterion to fit an AR(p) model to the short-memory component $\phi(B)\,\tilde{y}_t = u_t$.
(iv) We compute forecasts using Theorem 1.

Method 2

(i) We consider the differences $\Delta^s y_t$ (here $s = 1$).
(ii) We compute the GPH estimator $\hat{d}_1$ by using raw data (without tapering).
(iii) We filter out the long-memory stationary component $\tilde{y}_t = (1-B)^{\hat{d}_1 \Delta^s y_t}$.
(iv) We use the AIC criterion to fit an AR(p) model to the short-memory component $\phi(B)\,\tilde{y}_t = u_t$.
(v) We compute forecasts using Theorem 2.

DOI: 10.1002/for

For each model, we generated $T = 500$ data points; we used 20 additional data points for out-of-sample forecast.

Remarks: (i) To filter out the long-memory stationary or nonstationary component we may apply to the data the fractional filter $(1-B)^d = \sum_{j=0}^{\infty} \delta_j(d)B^j$. Practically we have only a finite sample and we must truncate the filter $(1-B)^d$ to $\sum_{j=0}^{m} \delta_j(d)B^j$, where m is such that $|\delta_m(d)| < 0.0001$. The resulting model will be an AR(m) − AR(p) for Method 1 and an ARI(s,m) − AR(p) for Method 2; this modelling coincides with the AR-ARMA approach proposed by Parzen (1982); however, here we estimate the long-memory filter by using a non-parametric approach.

(ii) The choice of m is not unique and hence $\tilde{y}_t$, which depends on m, will not be unique, as in Parzen (1982). However, the whitening filters (AR(m) − AR(p) for Method 1 and ARI(s,m) − AR(p) for Method 2) are unique. One filters out the long-memory component and the arbitrariness of the transformed $\tilde{y}_t$ will compensated by the transformation to u_t. Note that this arbitrariness of transformation (y_t to $\tilde{y}_t$) can have other sources. In addition to the choice of m, for example, one source is the method used to estimate the parameter d. The existing methods (the GPH method, 1983, Robinson method, 1995b, Whittle method, etc.) do not give the same estimate for the parameter d.

The graphs of PSMSE(h), $h = 1, \ldots, H$, for Model i, $1 \le i \le 5$, are given in Figure 1. These graphs imply that Method 1 has somewhat an advantage over Method 2 for all models excepting Model 3, where Method 2 has somewhat an advantage over Method 1 for lead times $h = 6$, 7, 8, and $h > 13$.

DATA EXAMPLES

In this section we analyse two empirical series, plotted in Figure 2. To compare the performance of the two methods we use the MSEP measure, i.e. the mean square error of prediction given by

$$\text{MSEP} = \frac{1}{H}\sum_{h=1}^{H}(\hat{y}_{T+h} - y_{T+h})^2$$

Arizona tree data

The first example is the annual tree-ring widths in Arizona (548–1983) which can be founded in the web page of R. Hyndman: www-personal.buseco.monash.edu.au. By considering only the first 500 observations, Velasco and Robinson (2000) found that the memory of the series is equal to $\hat{d} = 0.556$ by using Whittle estimates. Here we consider the whole sample; we use the first 1426 observations for estimation and the last 10 observations for the out-of-sample forecast.

In the first method, the order of taper is equal to $p_1 = 2$ in the estimation of d; the forecasts are presented in Table I, row 2. In the second method we take the first differences ($s = 1$) and we use raw data, i.e. the order of taper is equal to $p_1 = 1$ to estimate d_1; the forecasts are given in Table I, row 3.

We obtain $\hat{d} = 0.5631$ for the undifferentiated series and $\hat{d}_1 = -0.4891$ (anti-persistent memory) for the differentiated series.

The chemical process temperature readings (series C from Box and Jenkins, 1976)

We use the first 216 observations for estimation and the last 10 observations for the out-of-sample forecast. This series was also studied by Velasco and Robinson (2000), and all the proposed esti-

DOI: 10.1002/for

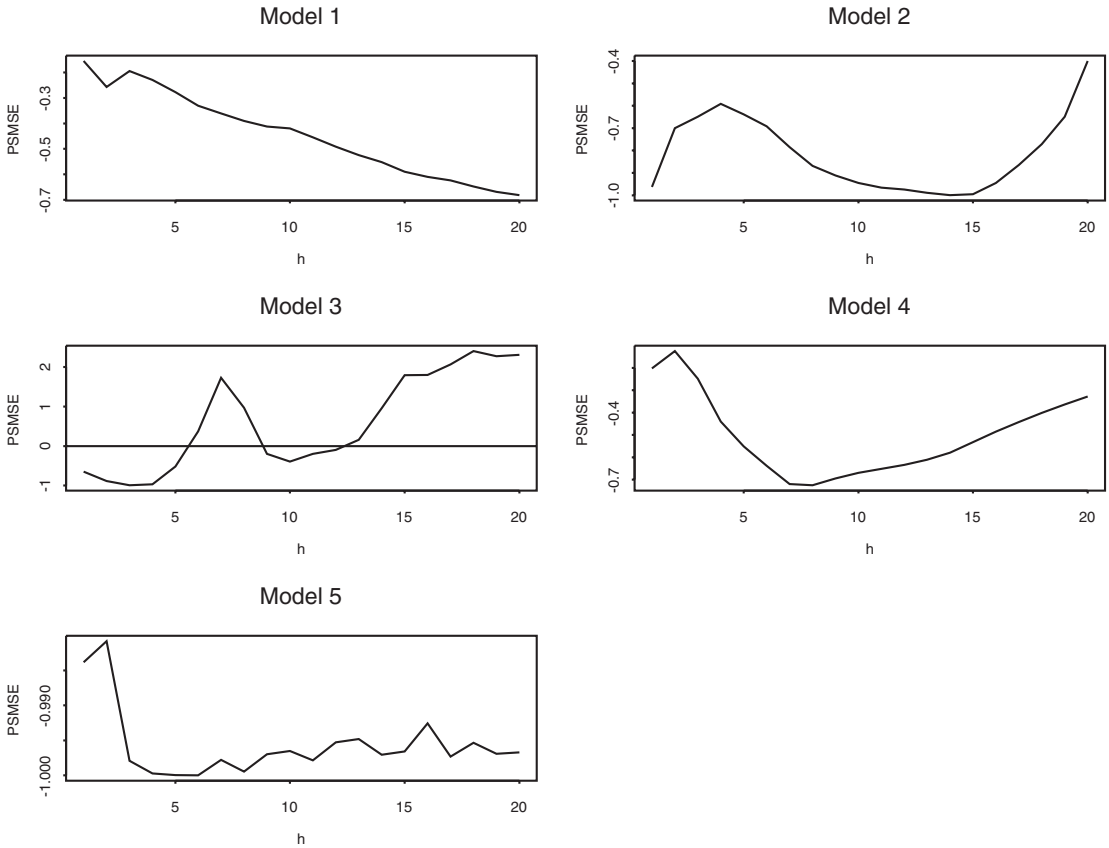

Figure 1. The graph of *PSMSE*(*h*) for Model *i*, $i = 1, \ldots, 5$

Table I. Forecasts of tree-ring widths in Arizona

True data (Arizona)	0.812	1.129	1.128	1.243	1.174	1.134	1.376	1.455	1.638	1.436	MSEP
Prediction 1	0.928	0.904	0.948	0.955	0.960	0.964	0.967	0.970	0.973	0.975	0.1297
Prediction 2	0.945	0.927	0.945	0.948	0.948	0.946	0.951	0.954	0.956	0.961	0.1386

mators of the memory of the series indicated that this memory is greater than 0.8676, and hence we need to apply a taper of order $p_1 = 2$. However, when we take $p_1 = 2$, the estimate of the memory is equal to $\hat{d} = 2.0$; i.e., the estimator converges towards p_1; this indicates that the order of taper is less than the memory of the process (see Velasco, 1999a, p. 349); in this case d will be greater than 1.5 (i.e., the process is nonstationary and non-mean-reverting). To apply Method 1, we need in fact a taper of order $p_1 = 3$; the forecasts by this method are presented in Table II, row 2.

In the second method we take the second differences ($s = 2$) and we use raw data, and the order of taper is equal to $p_1 = 1$; the forecasts are given in Table II, row 3. We obtain $\hat{d} = 2.3434$ for the undifferentiated series and $\hat{d}_1 = -0.2567$ for the double differentiated series.

DOI: 10.1002/for

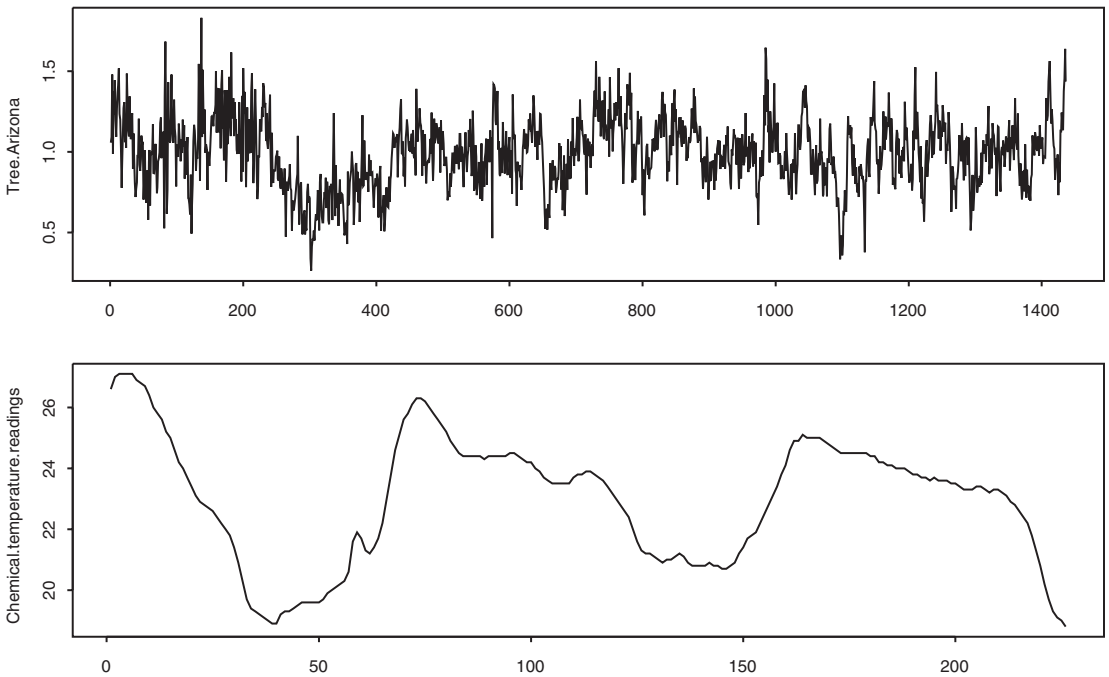

Figure 2. Arizona tree-ring widths (548–1983) (top) and chemical process temperature readings (bottom)

Table II. Forecasts of the chemical process temperature readings

True data (chemical)	22.2	21.8	21.3	20.8	20.2	19.7	19.3	19.1	19.0	18.8	MSEP
Prediction 1	22.196	21.994	21.830	21.625	21.442	21.261	21.051	20.865	20.636	20.425	1.6475
Prediction 2	22.221	22.049	21.882	21.728	21.578	21.429	21.280	21.127	20.977	20.823	2.1855

Remark: The estimators of d and d_1 are obtained by choosing the number of periodogram ordinates used in the GPH regression $g(T) = T^{0.74}$. If we choose $g(T) = T^{0.5}$, as suggested by Geweke and Porter-Hudak (1983), then we obtain $\hat{d} = 1.6833$ and $\hat{d}_1 = -0.1053$. However, there is no optimal choice of $g(T)$ in the GPH regression for nonstationary time series. This is also another source of arbitrariness of the transformed $\tilde{y}_t$ in our AR(m) − AR(p) methodology.

MEAN CORRECTION

When we forecast stationary time series with ARMA processes, generally we estimate the mean of the model by the sample mean $\bar{y}$, then we fit a (zero-mean) ARMA model to the 'mean-corrected' data. To obtain forecasts for the original data, we then add $\bar{y}$ to the forecasts obtained from the ARMA model. Indeed, we forecast the deviation from the mean. For ARFIMA(p, d, q) model, the 'mean correction' is not necessary, since if we assume that $E(y_t) = \mu$, then by applying equation (9)

DOI: 10.1002/for

to $y_t - \mu$, we obtain $\hat{y}_{t+h} = \phi^*(1)\mu + \sum_{j=1}^{\infty} \phi_j^*(d)\hat{y}_{t+h-j} + \sum_{j=h}^{q} \theta_{t+h-1,j}(y_{t+h-j} - y_{t+h-j}^*)$. But $\phi^*(1) = 0$, and hence the equation forecast for y_t is the same as for $y_t - \mu$. The same arguments can also be used for equation (31).

We have computed forecasts without 'mean correction' and obtained the results shown in Table III.

From Tables I and III, and Figure 3, we can see that for Method 2 the forecasts are identical, but for Method 1 the non 'mean correction' gives the worst forecasts.

From Tables II and IV, and Figure 4, we can see that for Method 2 the forecasts remain identical, but for Method 1 the non-'mean correction' gives the best forecasts.

The memory of the annual tree-ring widths is near the boundary of stationarity ($\hat{d} = 0.5631$) and is mean reverting, whereas the memory of the chemical process temperature readings is greater than

Table III. Forecasts of the tree-ring widths in Arizona without subtracting the mean

True data (Arizona)	0.812	1.129	1.128	1.243	1.174	1.134	1.376	1.455	1.638	1.436	MSEP
Prediction 1 (without)	0.900	0.897	0.893	0.889	0.885	0.881	0.876	0.872	0.867	0.863	0.1900
Prediction 2 (without)	0.945	0.927	0.945	0.948	0.948	0.946	0.951	0.954	0.956	0.961	0.1386

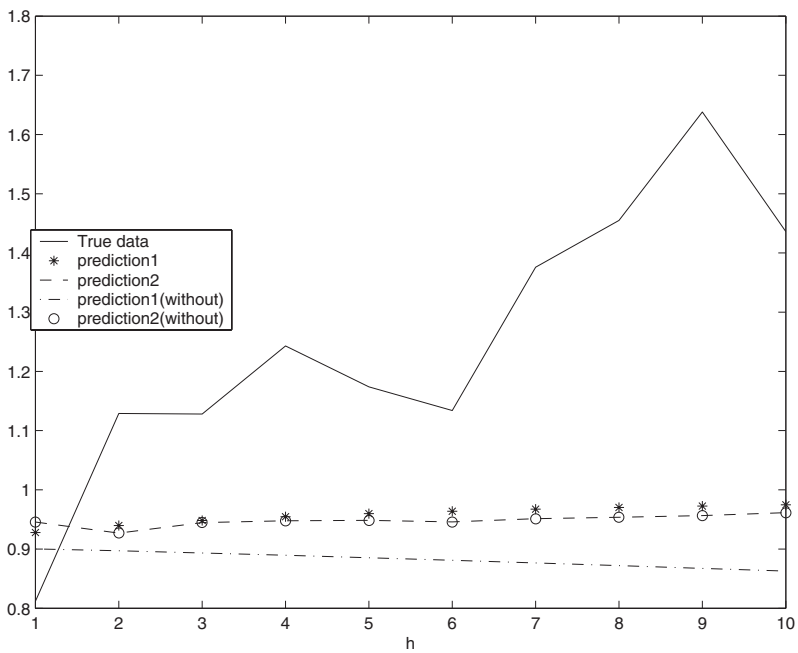

Figure 3. Forecasts of tree-ring widths in Arizona with and without subtracting the mean, by Method 1 (prediction 1) and Method 2 (prediction 2)

DOI: 10.1002/for

Table IV. Forecasts of the chemical process temperature readings without subtracting the mean

True data (chemical)	22.2	21.8	21.3	20.8	20.2	19.7	19.3	19.1	19.0	18.8	MSEP
Prediction 1 (without)	22.174	21.928	21.701	21.416	21.140	20.849	20.515	20.187	19.800	19.418	0.6444
Prediction 2 (without)	22.221	22.049	21.882	21.728	21.578	21.429	21.280	21.127	20.977	20.823	2.2125

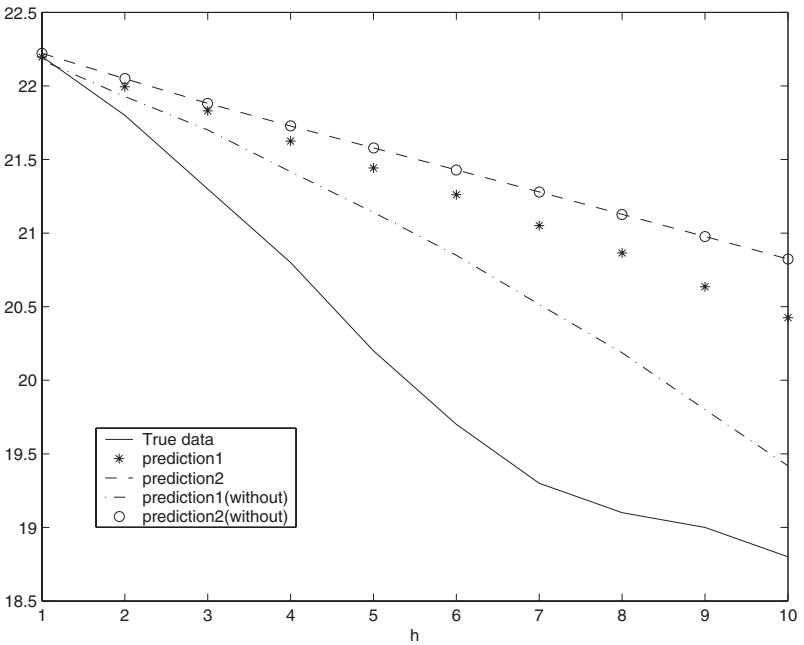

Figure 4. Forecasts of the chemical process temperature readings with and without subtracting the mean, by Method 1 (prediction 1) and Method 2 (prediction 2)

1 and hence is non-mean reverting. It seems that Method 1 can give superior forecasts in the case of non-mean reverting time series than Method 2, and is more robust to a possibly structural change in mean (see Figure 2).

CONCLUSION

It is not difficult to show that $\mathbf{\Psi}^{-1} = \mathbf{\Gamma}^{-1}(d_1)\mathbf{T}\mathbf{\Phi}^{-1}$; this equality implies that the two methods of forecast (Method 1 and Method 2) presented in this paper lead to the same forecasts $\hat{y}_{t+h}$ and the same mean squared errors $\sigma^2(h)$, $h = 1, \ldots$ This, of course, holds if we assume that (d, s, d_1) are known, $d = s + d_1$. However, in practice the parameters (d, s, d_1) will be estimated from the data. The existing estimators of the long-range dependent parameter are not, in general, invariant to differences; i.e., $\hat{d} \neq \hat{s} + \hat{d}_1$, where $\hat{s}$ is an estimator of s, which is the number of times we need to take differ-

DOI: 10.1002/for

ences to obtain a stationary and invertible series. We can remark that the Geweke and Porter-Hudak (1983) estimator is not invariant to first (second) differences; i.e., the estimator $\hat{d}$ based on the original data is not in general equal to one (two) plus the estimator $\hat{d}_1$ based on the differentiated data. For Arizona data we obtain $\hat{d}_1 + 1 = -0.4891 + 1 = 0.5109 \neq \hat{d} = 0.5631$. For the chemical process temperature readings we obtain $\hat{d}_1 + 2 = -0.2567 + 2 = 1.7433 \neq \hat{d} = 2.3434$. This non-invariability was pointed out by many authors; for instance, Agiakloglou *et al.* (1993) obtained $\hat{d}_1 = -0.26$ for the first difference of a series of US unemployment figures, whereas their estimate of d for the undifferentiated series was $\hat{d} = 0.9997$. A direct consequence of this non-invariability is that the two fitted ARMA(p, q) models given by equations (4) and (28) are not identical. Therefore the two proposed methods will give different forecasts and also different mean squared errors. By comparing the MSEP obtained for the two methods (by using simulated and real data), we can conclude that Method 1 is slightly superior to Method 2. This conclusion is in accordance with that of Parzen (1982), who suggested that estimating a nonstationary filter is superior to the approach of Box and Jenkins, who recommended taking successive differences until a stationary series was obtained.

REFERENCES

Agiakloglou C, Newbold P, Wohar M. 1993. Bias in an estimator of the fractional difference parameter. *Journal of Time Series Analysis* **14**: 235–246.

Beran J. 1995. Maximum likelihood estimation of the differencing parameter for invertible short and long-memory ARIMA models. *Journal of Royal Statistical Society* **B57**: 659–672.

Beran J, Bhansali RJ, Ocker D. 1998. On unified model selection for stationary and nonstationary short and long-memory autoregressive processes. *Biometrika* **85**: 921–934.

Bhansali RJ, Kokoszka PS. 2001. Prediction of long-memory time series: an overview. *Estadistica* **160–161**: 41–96.

Bhansali RJ, Kokoszka PS. 2003. Prediction of long-memory time series. In *Theory and Applications of Long-Range Dependence*, Doukhan P, Oppenheim G, Taqqu MS (eds). Birkhaüser: Baston, MA.

Box GEP, Jenkins GM. 1976. *Time Series Analysis, Forecasting and Control.* Holden Day: San Francisco.

Brockwell PJ, Davis RA. 1991. *Time Series: Theory and Methods.* Springer: Berlin.

Crato N, Ray BK. 1996. Model selection of long-range dependent processes: results of simulation study. *Journal of forecasting* **15**: 107–125.

Fox R, Taqqu MS. 1986. Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Annals of Statistics* **14**: 517–532.

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

Haslett J, Raftery AE. 1989. Space time modelling with long-memory dependence: assessing Ireland's wind power resource. *Applied Statistics* **38**: 1–50.

Hosking JRM. 1981. Fractional differencing. *Biometrika* **68**: 165–176.

Parzen E. 1982. ARARMA models for time series analysis and forecasting. *Journal of forecasting* **1**: 67–82.

Peiris MS, Perera BJC. 1988. On prediction with fractionally differenced ARIMA models. *Journal of Time Series Analysis* **9**(3): 215–220.

Robinson PM. 1995b. Gaussian semiparametric estimation of long range dependence. *Annals of Statistics* **23**: 1630–1661.

Velasco C. 1999a. Non-stationary log-periodogram regression. *Journal of Econometrics* **91**: 325–371.

Velasco C. 1999b. Gaussian semiparametric estimation of non-stationary time series. *Journal of Time Series Analysis* **20**(1): 87–127.

Velasco C, Robinson PM. 2000. Whittle Pseudo–Maximum likelihood Estimation for Nonstationary Time Series. *Journal of the American Statistical Association* **95**(452): 1229–1243.

DOI: 10.1002/for

Author's biography:
Mohamed Boutahar obtained his PhD from the Provence University, Marseille 1992. His main research interest is time series: asymptotic theory, nonstationary models, long-memory models and modelling in particular. He is researcher in the Department of Statistics and Econometrics of GREQAM.

Author's address:
Mohamed Boutahar, GREQAM, 2 rue de la vieille charité, and Department of Mathematics, Luminy Faculty of Sciences, 163 Av. de Luminy, 13288 Marseille Cedex 9, France.

DOI: 10.1002/for