International Journal of Wavelets, Multiresolution
and Information Processing
Vol. 5, No. 5 (2007) 709–724
© World Scientific Publishing Company

FORECASTING TIME SERIES USING WAVELETS

MINA AMINGHAFARI

Laboratoire de Mathématique–U.M.R. C 8628
"Probabilités, Statistique et Modélisation"
Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France
Mina.Aminghafari@math.u-psud.fr

and

Amirkabir University of Technology, Faculty of Mathematics
and Statistical Research Center, Tehran, Iran
Aminghafari@aut.ac.ir

JEAN-MICHEL POGGI

Laboratoire de Mathématique–U.M.R. C 8628
"Probabilités, Statistique et Modélisation"
Université Paris-Sud, Bât. 425, 91405 Orsay Cedex, France

and

Université Paris 5, Paris, France
Jean-Michel.Poggi@math.u-psud.fr

Received 8 July 2006
Revised 30 January 2007

This paper deals with wavelets in time series, focusing on statistical forecasting purposes. Recent approaches involve wavelet decompositions in order to handle non-stationary time series in such context. A method, proposed by Renaud *et al.*,[11] estimates directly the prediction equation by direct regression of the process on the Haar non-decimated wavelet coefficients depending on its past values. In this paper, this method is studied and extended in various directions. The new variants are used first for stationary data and after for stationary data contaminated by a deterministic trend.

Keywords: Forecasting; non-stationary; time series; wavelets.

Mathematics Subject Classification 2000: 62M20, 62M10, 65T60

1. Introduction

Wavelets have been used for various purposes in statistics including denoising, nonparametric function estimation, data compression as well as process synthesis, see for example Antoniadis.[2] An interesting example considering a time-varying autoregressive process for time series is given by Dahlhaus *et al.*[4] In a recent book, Percival and Walden[9] address a lot of problems and methods involving wavelets and time series, nevertheless, it should be noted that forecasting is not considered. To our knowledge, only a small number of papers consider this topic which is obviously of interest since wavelets are, for intrinsic reasons, well suited for dealing with non-stationary time series.

Classical assumptions including stationarity and linearity for instance, are approximately valid in numerous situations but put, of course, serious limitations in many cases. The idea is to replace the Fourier transform used for spectral representation of a stationary process, by the wavelet transform. It allows to analyze the local characteristics of a signal around every position in time and for all scales simultaneously. Nason *et al.*[8] define a representation of non-stationary stochastic process in terms of the discrete non-decimated wavelet coefficients called Locally Stationary Wavelet (LSW) processes. This leads to the first approach for forecasting. It proposes to relax the stationarity assumption by introducing the wavelet spectrum as a local version of the well known spectrum and leads to solve a generalization of the Yule–Walker equations involving time-dependent coefficients weighting the past observations and use it for forecasting (see Fryzlewicz *et al.*[6]). A second approach considered by Renaud *et al.*[11] starts from an entirely different viewpoint and estimates the prediction equation by direct regression of the process on the Haar non-decimated wavelet coefficients depending on its past values.

This paper extends on this second approach, containing a new proposal for prediction by regression on wavelet coefficients. This prediction method is used first for stationary data and after for stationary data contaminated by a deterministic trend.

2. Why Wavelets for Time Series Forecasting?

Let us give some key arguments for using wavelet decomposition for forecasting time series. Let us assume that the observed series is of the form: $Y(t) = f(t) + X(t)$, where X is a stochastic process and f is a deterministic smooth function. The deterministic part of the series (trend or smooth part) can be estimated, for example, by polynomial fitting of the approximation coefficients of a suitably chosen decomposition level and the detail coefficients are used for the prediction of the purely stochastic part.

Hence, from this veiwpoint and assuming the stationarity of X and a sufficiently regular wavelet, wavelet transform automatically filters the non-stationary component of the signal, instead of trying to detrend or suppress quasi-periodic smooth components as in the classical non-stationary $ARIMA$ approach. The complex multiscale structure of the observed signal can often be simplified using wavelets leading to signals of simpler structure. The coefficients at a given scale, of a long memory stochastic signal are of short memory (see for example Percival and Walden,[9] Chap. 9). The wavelet coefficients depend on the values of a signal only within an interval centered around the corresponding position and of length proportional to the associated scale. It follows that if the considered series is not too far from stationarity, procedures using wavelets are naturally localized. This key idea is used in the LSW model (see Nason *et al.*[8]), which proposes to define local stationarity using a wavelet spectrum.

3. Non-Decimated Wavelet Transform

For basics on wavelets for statistics, the reader can refer to Percival and Walden.[9] The so-called discrete wavelet transform (DWT) is often used for estimation purposes. But this transform is not translation invariant. This lack of translation invariance has both practical and theoretical consequences. On one hand, the computations need to be performed again when a new observation is available, and on the other hand, trying to fit a regression model between a future value of a series and some DWT coefficients depending on the past values, is ill-posed since the explanatory variables change with the instant of the last observed value. A classical way to circumvent this drawback is to use the non-decimated wavelet transform (NDWT).[9] Of course, NDWT does not perform a decomposition on an orthogonal basis but provides all the possible decompositions corresponding to the possible combinations of odd or even subsampling at each DWT decomposition step. It leads to a huge redundancy of information.

4. Prediction by Regression on Wavelet Coefficients

An approach to predict X_{N+1} based on wavelet coefficients, using N observations is presented by Renaud *et al.*[11] For the Haar wavelet, the reconstruction formula for $t = N + 1$ can be written as follows:

$$X_{N+1} = c_{J,N+1} + \sum_{j=1}^{J} w_{j,N+1}.$$

Then, to predict X_{N+1}, it suffices to predict the NDW approximation and detail coefficients $c_{J,N+1}$ and $w_{j,N+1}$. Hence, the idea is to predict for each scale, the unknown NDW coefficient by a linear combination of their past values dyadically lagged starting from N; that is to write:

$$\hat{w}_{j,N+1} = \sum_{k=1}^{r_j} a_{j,k} w_{j,N-2^j(k-1)},$$

and

$$\hat{c}_{J,N+1} = \sum_{k=1}^{r_{J+1}} a_{J+1,k} c_{J,N-2^J(k-1)}.$$

Let us note that the past values appearing on the right hand sides of the two previous equations must depend only on the past observations of the process itself. The explanatory variables are selected using dyadically lagged values of the NDW coefficients in order to extract the coefficients of the non redundant (i.e. the decimated) wavelet transform corresponding to the dyadic grid adapted to (i.e. ending at) the last observed value.

Then the complete prediction equation of X_{N+1}, when N observations $X_1, \ldots, X_N$ are given, is of the following form:

$$\hat{X}_{N+1} = \sum_{j=1}^{J} \sum_{k=1}^{r_j} a_{j,k} w_{j,N-2^j(k-1)} + \sum_{k=1}^{r_{J+1}} a_{J+1,k} c_{J,N-2^J(k-1)}. \tag{4.1}$$

Denoting the explanatory variables vector and parameter vector by

$$D_t = [w_{1,t}, \ldots, w_{1,t-2r_1}, \ldots, w_{J,t}, \ldots, w_{J,t-2^J r_J}, c_{J,t}, \ldots, c_{J,t-2^J r_{J+1}}]^T \tag{4.2}$$

$$\alpha = [a_{1,1}, \ldots, a_{1,r_1}, \ldots, a_{J,1}, \ldots, a_{J,r_J}, \ldots, a_{J+1,1}, \ldots, a_{J+1,r_{J+1}}]^T, \tag{4.3}$$

then, the prediction equation can be written as $\hat{X}_{N+1} = {D_N}^T \alpha$. The sequence α is estimated by minimizing the empirical mean square prediction error:

$$\hat{\alpha}_N = \arg\min_{\alpha} \sum_{k=M}^{N} \left(X_k - D_{k-1}^T \alpha \right)^2, \tag{4.4}$$

where M is a fixed integer. Of course, if X_t is stationary and Gaussian, the optimal prediction with respect to the mean square prediction error is given by a linear combination of the past values of the process. Since the extracted decimated coefficients are obtained by applying an invertible linear transformation to the original process, it follows that the two strategies differ only from the parametrization view-point. Since in Eq. (4.1), suitably chosen decimated coefficients of the NDW decomposition are involved, r_j could be selected as the order obtained from the fit of an AR process on the decimated coefficients of level j. Following this idea, the identification step performed can be followed by an estimation step and the corresponding parameters could be estimated by the previous AR fit. Let us note that the number of parameters (the length of α) can be large and is always greater than $J + 1$.

5. A New Proposal for Prediction by Regression on Wavelet Coefficients

The first change we propose starting from the procedure of Renaud *et al.*[11] is the extension to an arbitrary compactly supported orthogonal wavelet.

Indeed, the previous strategy involving the Haar wavelet seems to be designed in such a way that the reconstruction step is obvious which could be useful if the prediction is obtained from the reconstruction of scale-by-scale forecasted coefficients (as in the paper by Soltani *et al.*[12]). But, since the prediction equation (4.1) is directly fitted from the data, the reconstruction equation which is particularly simple for the Haar wavelet is never used.

In fact, it is possible to relax the constraint and allowed to use more regular wavelets. It suffices to modify the definition of the NDWT for coefficients involving the boundaries of the signal, by using zero-padding. The NDW coefficients of $(X_1, \ldots, X_N)$ are defined by

$$w_{j,s} = \sum_{k=0}^{L_j-1} g_{j,k} X_{s-k} \mathbb{1}_{s-k>0}(s)$$

and

$$c_{j,s} = \sum_{k=0}^{L_j-1} h_{j,k} X_{s-k} \mathbb{1}_{s-k>0}(s) \tag{5.1}$$

for $s = 1, 2, \ldots, N$, where $\mathbb{1}_H(x)$ is the indicator function of set H. The filters $\{h_{j,k}\}_{j,k}$ and $\{g_{j,k}\}_{j,k}$ are the filters corresponding to the NDWT wavelet and obtained by convolving upsampled versions of h, low-pass and g, high-pass filters associated with the orthogonal wavelet, normalized in l^1 and supposed to be of length L (see Percival and Walden,[9] p. 169). In addition, $L_j = (2^j - 1)(L - 1) + 1$.

This preserves the two following desired properties: non-anticipativeness and the fact that for any $t \leq N$ the transform of the t first samples of the signal is the same as the truncated at t of the transform of the complete signal.

The main advantages of the new degree of freedom are: to allow more regular wavelets; to adapt its choice to the particularities of the data; and to make profit from the automatic filtering capabilities of the wavelet transform to decompose series with a sufficiently regular wavelet in order to concentrate the deterministic part on the approximation coefficients. In the sequel of this section, we study the prediction procedure from an experimental point of view, having in mind the underlying additive decomposition of the observation $Y = X + f$, where, roughly speaking, X is a random stationary process and f captures the trend and low frequency components and can be considered as a deterministic component for short term prediction. Then, we first define a procedure suitable for stationary signals and then extend it to deal with additional deterministic trend.

5.1. *Stationary case: The procedure*

The proposed method to predict a stationary signal is to estimate, thanks to the past observations, the parameters of the prediction equation (4.1) where the NDW coefficients are calculated using an arbitrary compactly supported orthogonal wavelet. This step involves a stepwise ascending method to select the convenient explanatory variables. Then using Eq. (4.1) with the estimated parameters, the next future value is predicted. For a given wavelet (whose associated l^1-normalized filters are h and g), the method can be described as follows:

- Step 1: Compute wavelet decomposition.
  Perform NDW decomposition at level J of observed time series, $(X_1, \ldots, X_N)$ using h and g. This step produces $J + 1$ vectors of size N.

- Step 2: Set prediction equation.
  Select $(r_1^{(1)}, \ldots, r_{J+1}^{(1)})$ in the prediction equation, the maximum numbers of lagged explanatory variables at each level of decomposition. For $1 \leq j \leq J$, $r_j^{(1)}$ is taken as the order of AR process fitted on the dyadically downsampled version of W_j (or C_J) starting from the last coefficient. The prediction equation (4.1) involves D_t the vector of dyadically lagged NDW coefficients and can be written as $\hat{X}_{t+1} = X_{t+1} = D_t^T \alpha$ for past observations.
- Step 3: Estimate prediction equation.
  Perform estimation using *ascending stepwise regression* between X_t and D_{t-1} thanks to the observations of index up to N and estimate α by $\hat{\alpha}_N$.
- Step 4: Compute prediction $\hat{X}_{N+1} = D_N^T \hat{\alpha}_N$.

Let us first give some comments on Step 1. The scaling filter h and its quadrature mirror filter g are used for decomposition, i.e. to calculate the NDW coefficients. Of course, we can also use the reversed filters denoted by $\bar{g}$ and $\bar{h}$ to calculate the NDW coefficients following Percival and Walden.[9] Let us note that this first choice is the usual one. The choice between the two alternative solutions could be data dependent. For example, let us consider the Daubechies wavelet *db*2 (two vanishing moments), the filter g gives larger weights to the more recent values of the signal in the detail coefficients calculations and the opposite happens for approximation coefficients. Using reversed filters $\bar{h}$ the opposite occurs. We will see later that the performance of prediction can be improved by using $\bar{h}$ and $\bar{g}$ for signals with the large low-frequency component, as excepted since less information is lost in the lag. Hence a useful alternative to Step 1 is:

- Step 1′: Perform NDW decomposition at level J of observed time series using the reversed filters $\bar{h}$ and $\bar{g}$.

Then, a remark about Step 2 is that only dyadically lagged variables of the NDW coefficients are considered. A variant is, in addition, to consider as candidates all the lagged variables filled the gaps and to leave the variable selection estimation algorithm to select the influential ones. Then, the alternative prediction equation can written as:

$$\hat{X}_{N+1} = \sum_{j=1}^{J} \sum_{k=1}^{r_j} a_{j,k} w_{j,N-k+1} + \sum_{k=1}^{r_{J+1}} a_{J+1,k} c_{J,N-k+1}, \tag{5.2}$$

where the number of explanatory variables $(r_j)_{1 \leq j \leq J+1}$ are chosen as described below. Similarly to the previous definitions of D_t, if we set

$$D_t = [w_{1,t}, \ldots, w_{1,t-r_1+1}, \ldots, w_{J,t-r_J+1}, c_{J,t}, \ldots, c_{J,t-r_{J+1}+1}]^T \tag{5.3}$$

$$\alpha = [a_{1,1}, \ldots, a_{1,r_1}, \ldots, a_{J,1}, \ldots, a_{J,r_J}, \ldots, a_{J+1,1}, \ldots, a_{J+1,r_{J+1}}]^T, \tag{5.4}$$

stands for the alternative parameter vector , therefore, the prediction equation can be written as: $\hat{X}_{N+1} = D_N^T \alpha$. Then an alternative to Step 2 is:

- Step $2'$: Select $(r_1^{(2)}, \ldots, r_{J+1}^{(2)})$ in the prediction equation, the maximum numbers of explanatory variables at each level of decomposition. For $1 \leq j \leq J$, take $r_j^{(2)} = 2^j \log_2(p_j)$ where p_j is the order of AR process fitted on W_j and $r_{J+1}^{(2)} = 2^J \log_2(p_{J+1})$ where p_{J+1} is the order of AR process fitted on C_J.

5.2. *Stationary case: Simulated examples*

5.2.1. *The considered models*

In this section, we consider some selected examples of stationary (or at least asymptotically stationary) models, given in Table 1:

- one high order $AR(14)$ to experiment longer short-dependence;
- one *FARIMA* model to cope with the long-memory case. Let us recall that a *FARIMA* process exhibits long memory characteristics: its autocovariance decays polynomially to zero instead of exponentially for the AR models; and
- a highly nonlinear model i.e. generalized thresholded autoregressive ($GTAR$) used in Poggi and Portier.[10]

5.2.2. *The prediction methods*

Two different wavelets are used and compared: *haar* and *db*2 (and the associated reversed filters).

For each wavelet, three prediction methods are compared:

- the first one, the reference scheme, involves lagged NDW coefficients as explanatory variables, and corresponds to Steps $1, 2, 3, 4$;
- the second one consider in addition the NDW coefficients filled in the gaps of the previous lagged variables and corresponds to Steps $1, 2', 3, 4$;
- the last one uses $\bar{h}$ and $\bar{g}$ the reversed versions of the filters and corresponds to Steps $1', 2', 3, 4$.

Table 1. The considered *AR FARIMA* and *GTAR* models where ε_t is a Gaussian white noise with zero mean and unit variance.

Model	Equation	Parameters
$\mathcal{M}_{AR}$	$(1 - \phi_1 B - \cdots - \phi_p B^p) X_t = \varepsilon_t$ $\phi = (\phi_1, \ldots, \phi_p), p = 14$	$\phi = (0.1, 0.23, 0.1, -0.2, -0.1, 0.1, 0,$ $0.2, -0.03, 0.1, 0.1, 0.02, 0.1, 0.01)$
$\mathcal{M}_{FARIMA}$	$(1 - \phi_1 B)(1 - B)^d X_t = \varepsilon_t$	$\phi_1 = 0.5, d = 0.4$
$\mathcal{M}_{GTAR}$	$X_{t+1} = \eta_1 X_t \mathrm{ll}_{\lvert X_t \rvert > 1}$ $-(\eta_2 + \eta_3 \lvert X_t \rvert) X_t \mathrm{ll}_{\lvert X_t \rvert \leq 1} + \varepsilon_{t+1}$	$\eta_1 = -0.5, \eta_2 = 0.1, \eta_3 = 0.4$

Since the prediction scheme defined by Steps $1', 2, 3, 4$ gives similar results to the reference method, the associated results are not reported here.

Remark 5.1. Let us mention that since in Renaud *et al.*,[11] the results are only given graphically, a careful comparison of simulation results is difficult. In addition, they restrict wavelet prediction to the use of the reference method but only for the Haar wavelet (let us quote that in that case the Steps 1 and $1'$ give the same prediction results).

The simulated time series considered below are of length $N = 2000$, so the decomposition level J can be at most $\lfloor \log_2(N) \rfloor \simeq 10$. In this section, the maximum decomposition level is set to $J = 4$. This moderately large value is a compromise between the smoothing effect provided by large values of J and the number of parameters to estimate as well as the number of boundary coefficients, both increasing with J (let us mention that an additional remark about the choice of J can be found in Aminghafari[1]).

5.2.3. *The experimental framework*

For each model, we simulate 20 realizations of time series of size $N = 2000$ denoted by $(x_1^k, \ldots, x_N^k)_{k=1,\ldots,20}$.

Each realization of size $N = 2000$ observations is divided in two groups (standing for the past and the future respectively) of size $n = 1500$ (for the past) and $N - n = 500$ (for the future). The first group (the training set of observations) $x_1, \ldots, x_n$ is used to select the explanatory variables and to estimate α by $\hat{\alpha}$ in the prediction equation, which is kept fixed for all the forecasts thanks to stationarity. The second group (the test set) is used to evaluate performance of the one-step forecasting procedure.

The performance of the three considered prediction schemes is assessed by computing the standard deviation of the prediction errors on the test sample, which is an estimate of the square root of the Mean Square Prediction Error (MSPE):

$$R_{pred}(x^k, \hat{x}^k) = \sqrt{\frac{1}{N-n} \sum_{t=n+1}^{N} (\hat{x}_t^k - x_t^k)^2}, \quad k = 1, \ldots, 20, \tag{5.5}$$

where $\hat{x}_t^k$ denotes the prediction of x_t^k.

For each model, the average over the 20 realizations of the defined criteria: $\bar{R}_{pred}$ is computed. The closer $\bar{R}_{pred}$ to 1 (the noise standard deviation), the better the prediction. The performance results are given in Table 2.

Let us first examine the autoregressive data generated from the $\mathcal{M}_{AR}$ an $AR(14)$ model, to study a situation exhibiting a "long" short dependency. Except for the third method, *haar* outperforms *db*2. But, applying the third method, which reverses the filters, the use of *db*2 leads to similar performance, as good as the *haar* one. The distribution of energy, similar for the two wavelets, is given in Table 3.

Table 2. $\mathcal{M}_{AR}$, $\mathcal{M}_{FARIMA}$ and $\mathcal{M}_{GTAR}$: prediction performance.

Wavelet	Method		$\bar{R}_{pred}$	
		$\mathcal{M}_{AR}$	$\mathcal{M}_{FARIMA}$	$\mathcal{M}_{GTAR}$
haar	steps $1, 2, 3, 4$	1.03	1.02	1.01
	steps $1, 2', 3, 4$	1.02	1.02	1.02
*db*2	steps $1, 2, 3, 4$	1.10	2.08	1.07
	steps $1, 2', 3, 4$	1.06	1.03	1.04
	steps $1', 2', 3, 4$	1.02	1.04	1.14

Table 3. $\mathcal{M}_{AR}$, $\mathcal{M}_{FARIMA}$ and $\mathcal{M}_{GTAR}$: Distribution of energy across scales, of the NDW coefficients using *db*2.

Signal	W_1	W_2	W_3	W_4	C_4
Energy percentage of $\mathcal{M}_{AR}$	43	18	19	4	16
Energy percentage of $\mathcal{M}_{FARIMA}$	6	9	13	13	59
Energy percentage of $\mathcal{M}_{GTAR}$	77	14	5	2	2

For $\mathcal{M}_{FARIMA}$, the approximation coefficients captures, as expected, the major part (about 60%) of the total energy of the signal. Except for the first method, *db*2 performs as well as *haar*. It is clear in this example that the first method does not consider a sufficiently large number of candidate variables and so a very small model of poor performance is delivered. Increasing the number of candidate variables suffices to obtain satisfactory behavior but leads to a complex prediction equation.

For $\mathcal{M}_{GTAR}$, the results for *haar* are of good quality and only the second method reaches comparable performance. The effect of reversing the initial filters leads to a strong degradation (around 10%), since the distribution of the NDW *db*2-coefficients is concentrated at the first level (around 80%, see Table 3).

Remark 5.2. Let us give a short conclusion on the proposed procedure for the stationary case, taking Renaud *et al.*[11] as the starting reference:

- the extension of the prediction procedure to *db*2 leads to similar or slightly degraded performance, at least when the convenient variants are used;
- the various parameters of the prediction methods ($(r_j)_j$ and the variables involved in the prediction equation) are chosen automatically; and
- different variants have been considered and the most promising seems to be the following: to apply the Step $2'$ to generate the candidate variables and to use the filters $(g, \bar{h})$ for coefficients calculation.

The main conclusion is that the use of a wavelet more regular than *haar* can be considered to handle the stationary part of the signal while the non-stationary component captured by the approximation coefficients can be forecasted following different ideas explained in the next section.

5.3. *Non-stationary case: The procedure*

The prediction procedure sketched previously, is not designed to handle time series involving a deterministic trend. Suppose that the observed time series are of the form $Y_t = X_t + f(t)$, where X_t is a purely stochastic time series and $f(t)$ is a deterministic component. From the reconstruction equation (see Percival and Walden,[9] p. 173) we can write

$$Y_t = (\mathcal{A}_J(X))_t + \left(\sum_{j=1}^{J} \mathcal{D}_j(X)\right)_t + f(t), \quad 1 \le t \le N, \tag{5.6}$$

where $\mathcal{A}_j(X)$ and $\mathcal{D}_j(X)$ are suitably chosen reconstructed versions of approximation and detail of level j respectively, obtained from the NDW coefficients of X.

So, a procedure for the prediction of a signal contaminated by a trend can be proposed by extending the previous one:

- Step 1: The term $\mathcal{D}_t = (\sum_{j=1}^{J} \mathcal{D}_j(X))_t$ can be predicted using the previous procedure since the details are supposed to be free of f for a convenient wavelet choice. Of course, this step must reject the approximation coefficients in the prediction equation (5.2) by taking $r_{J+1} = 0$. Let us denote by $\hat{\mathcal{D}}_{N+1}$, the prediction of the high frequency stationary components delivered by this step.
- Step 2: Then,

$$Z_t = Y_t - \widehat{\left(\sum_{j=1}^{J} \mathcal{D}_j(X)\right)_t} = Y_t - \hat{\mathcal{D}}_t$$

  estimates $(\mathcal{A}_J(X))_t + f(t)$. This signal can be extrapolated following various deterministic or stochastic ways. For example, by polynomial fitting for the first case or by exponential smoothing for the second one. In the sequel, we use polynomial fitting (globally or only locally) of order selected by polynomial stepwise regression. Let us denote by $\hat{Z}_{N+1}$, the extrapolation of the low frequency components given by this second step.
- Step 3: Finally, the prediction of the top level time series Y is given by:

$$\hat{Y}_{N+1} = \hat{\mathcal{D}}_{N+1} + \hat{Z}_{N+1}.$$

5.4. *Non-stationary case: Real data*

This method has been studied (see Aminghafari[1]) using the Monte–Carlo approach applied to the synthetic stationary processes, previously examined, contaminated by various deterministic functions (linear, quadratic, piecewise monotonic and

Table 4. Nile data: Prediction performance.

Method	J	Bandwidth	Wavelet	R_{pred}
first	4		*haar*	48.75
first	2		*haar*	48.77
first	4		*db*2	50.09
first	2		*db*2	49.24
local	4	20	*haar*	54.56
local	4	15	*haar*	54.06
local	4	10	*haar*	56.76
local	4	20	*db*2	51.31
local	4	15	*db*2	**48.50**
local	4	10	*db*2	**48.83**

periodic). These experiments are not detailed here. Let us compare on a real data set, the method sketched in the last paragraph with a first method taken as a reference:

- *First Method*: it is simply the prediction procedure designed for stationary process without any specific adaptation.
- *Local Method*: the method of the previous section and for Step 2, a polynomial fitting on all the available past values of $Z(t)$ is performed only locally using a window (a bandwidth from 5 to 40).

The indication about choosing the bandwidth using a kind of hold-out procedure can be found in Aminghafari.[1]

Let us consider Nile data provides a standard benchmark data set. It exhibits long-range dependence and contains significant low frequency components even if no obvious trend can be identified. Let us divide the data of length 647 in two parts, $n = 600$ for the "so-called" past data and $N - n = 47$ for the data to predict. The window bandwidth for the local method can be fixed to 10, 15 or 20. The performance is computed by the standard deviation of the prediction errors on the test sample and is denoted by R_{pred}. The performance results are given in Table 4. The best performance obtained for the local method using $db2$ and a window bandwidth 15, is equal to 48.50. This performance is close to the best performance obtained for the first method i.e. 48.21 obtained for $J = 2$ and using $db2$ with filters $(\bar{g}, \bar{h})$ (not shown in the table). The performance obtained by taking 10 as window bandwidth (the automatically selected bandwidth in Aminghafari[1]) is equal to 48.83. The performance of the local method using $db2$ is better than the same one using $haar$.

Figure 1 shows the use of Nile data to predict on the left and the corresponding predictions based on the local method using $db2$ with 15 as bandwidth and predictions based on the first method using $db2$ (the couple of filters $(g, \bar{h})$) and $J = 2$ on the right. The fluctuations of the data are correctly predicted by both methods but with an amplitude slightly minored especially around the picks. We can see the

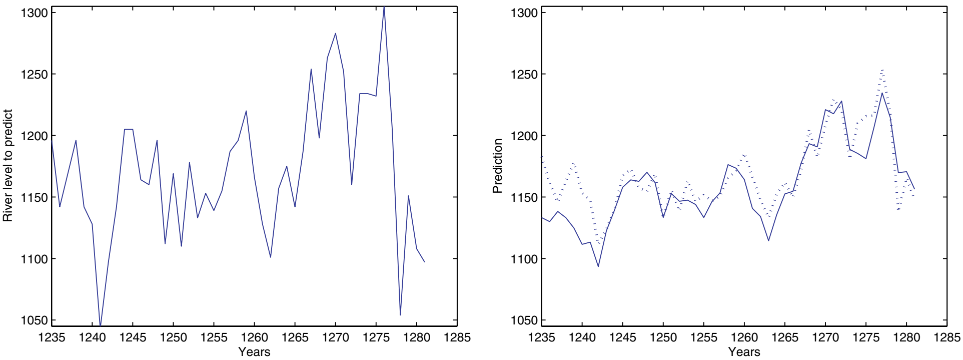

Fig. 1. On the left: Using Nile data to predict. On the right: The corresponding predictions based on the local method using $db2$ with 15 as bandwidth (solid line) and based on the first method using $db2$ and $J = 2$ (dotted line).

prediction based on the local method is smoother than the prediction based on the first method, but the local prediction exhibits less delay than the second one.

6. Theoretical Results

In this section devoted to theoretical results, we consider the prediction of a purely stochastic signal using an arbitrary orthogonal compactly supported wavelet.

Renaud *et al.*[11] proved that if X_t is a linear AR process and if the Haar wavelet is used to calculate NDW coefficients, then $N^{1/2}(\hat{\alpha}_N - \alpha)$ converges in distribution to a centered Gaussian distribution where α is a linear transform of autoregressive parameters. The AR assumption is not a necessary condition and it is probably used to simplify technical problems. We will try to relax this linearity condition and consider a class of functional AR processes. In addition, we allow to use any orthogonal compactly supported wavelet and not necessarily the Haar one.

In Sec. 6.1, we first establish an almost sure convergence result for $\hat{\alpha}_N$ for an ergodic process then we introduce a more constrained framework of functional AR in Sec. 6.2 and give an almost sure convergence and CLT. Let us note that, for the asymptotic results, $(r_j)_{1 \leq j \leq J+1}$ are supposed to be known and an ordinary regression is performed on dyadically lagged NDW-coefficients of an orthogonal compactly supported wavelet.

6.1. *Asymptotic results for ergodic process*

Let $X_k^{(s)}$ be a vector in $\mathbb{R}^s$ equal to $(X_k, X_{k-1}, \ldots, X_{k-s+1})^T$. For a given orthogonal compactly supported wavelet, the discrete wavelet transform of vector $X_k^{(s)}$ can be written as $\mathcal{W}X_k^{(s)}$ where $\mathcal{W}$ is a $s \times s$ orthogonal matrix defined using the wavelet and scaling filters impulse responses (see Percival and Walden,[9] p. 57). We look for the relationship between the vector D_t defined in (4.2) and the realization of series, and use a sub-matrix of $\mathcal{W}$ to construct D_t. Since the explanatory variables D_t

involve the dyadically lagged NDW coefficients, we can find the orthogonal matrix $\mathcal{W}$ of some dimension B such that $\mathcal{W}X_t^{(B)}$ contains all coefficients in D_t. Hence, we only need to find the dimension B to determine the relationship between D_t and the observations of series.

Lemma 6.1. *Let B be an integer multiple of a power of two (i.e. $B = k2^{J_0}$ for some J_0) greater than or equal to the maximum of $\{2^j(r_j - 1) + L_j, j = 1, \ldots, J\}$ and $2^J(r_{J+1} - 1) + L_J)$ where $L_j = (2^j - 1)(L - 1) + 1$, J is the decomposition level and L is the length of considered wavelet. Then for any given t, $\mathcal{W}X_t^{(B)}$ contains all the coefficients involved in D_t.*

We can find a sub-matrix of $\mathcal{W}$, $\mathcal{RW}$ by omitting some rows of $\mathcal{W}$ such that

$$D_t = \mathcal{RW}[X_t, \ldots, X_{t-B+1}]^T = \mathcal{RW}X_t^{(B)}, \tag{6.1}$$

where $\mathcal{R}$ is the $Q \times B$ matrix containing 0 and 1, i.e. $\mathcal{R}$ is a sub matrix of $Q \times Q$ identity matrix where $Q = \sum_{j=1}^{J+1} r_j$.

The sequence $\hat{\alpha}_N$ in Eq. (4.4) satisfies the following equation

$$\hat{\alpha}_N = S_N^{-1} \sum_{k=M}^{N} D_{k-1}X_k, \tag{6.2}$$

where $S_N = \mathcal{Q} + \sum_{k=M}^{N} D_{k-1}D_{k-1}^T$ and $\mathcal{Q}$ is a hermitian positive definite matrix. Under ergodicity assumption, we show *almost sure* convergence of $\hat{\alpha}_N$.

Theorem 6.1. *Suppose $\{X_t\}_t$ be an ergodic strictly stationary process with $E|X_t| < \infty$ and $EX_t^2 < \infty$. Let B be an integer multiple of a power of two larger than or equal to the maximum of $\{2^j(r_j - 1) + L_j, j = 1, \ldots, J\}$, $2^J(r_{J+1} - 1) + L_J$ and p. Let M be greater than $B - 1$. Suppose that the covariance matrix of $X_t^{(s)}$ be invertible for any s. Then we have $\hat{\alpha}_N \xrightarrow[N \to \infty]{a.s.} \alpha$ where $\alpha = \operatorname{argmin}_{\theta \in \mathbb{R}^Q} \mathbb{E}[X_{N+1} - D_N^T\theta]^2$.*

Sketch of Proof: The result is a direct application of Theorem 2 of Hannan,[7] p. 203. The detailed proof is given in Aminghafari,[1] pp. 112–113.
This theorem expresses under very general condition (i.e. ergodicity) that the sequence $\hat{\alpha}_N$ converges to the parameter of the best linear prediction based on the NDW coefficients of past values. For example, this theorem holds if $\{X_t\}$ is mixing or for any Gaussian stationary process with continuous spectral distribution.

6.2. *Asymptotic results for functional AR*

In this section, we introduce a more constrained framework of the functional *AR* and we give an almost sure convergence and CLT.

We consider a functional *AR* model of order p of the following form:

$$X_{n+1} = f(X_n, X_{n-1}, \ldots, X_{n-p+1}) + \xi_{n+1}, \quad n \in \mathbb{N},$$

where the function f from $\mathbb{R}^p$ to $\mathbb{R}$ is unknown, the model order p is known and ξ_n is a Zero mean white noise with variance σ^2. Let given initial value of $X_0^{(p)}$ be independent of $(\xi_n, n \in \mathbb{N})$. We make the following assumptions on the functional AR process.

Assumptions A.

- Assumptions $A1$: f is a continuous function and there exist non-negative constants $\lambda_1, \lambda_2, \dots, \lambda_p$ with $\sum_{i=1}^p \lambda_i < 1$ such that

$$|f(x) - f(y)| \leq \sum_{i=1}^{p} \lambda_i |x_i - y_i|, \quad \text{for any}\, x, y \in \mathbb{R}^p,$$

  where $x = (x_1, x_2, \dots, x_p)$ and $y = (y_1, y_2, \dots, y_p)$.
- Assumption $A2$: The noise $(\xi_n)_{n\geq 0}$ has a moment of order $m > 2$ and it has a probability density function.

Under these assumptions X_t is an asymptotically stationary process with a stationary measure μ (for more details see Duflo,[5] Chap. 6). Nevertheless the conditions A exclude some linear models (see Poggi and Portier,[10] p. 619).

We establish two convergence results of the least squares estimator given by (6.4).

First, we establish the almost sure convergence of $\hat{\alpha}_N$. Second, we give the asymptotic distribution of $\hat{\alpha}_N$. The sequence $\hat{\alpha}_N$ in Eq. (4.4), satisfies the following equation

$$\hat{\alpha}_N = S_N^{-1} \sum_{k=M}^{N} D_{k-1} X_k, \tag{6.3}$$

where $S_N = S_{N-1} + D_{N-1} D_{N-1}^T$ and $S_{M-1} \geq \lambda I$ for some $\lambda > 0$.

To reduce computational cost, the sequence $\hat{\alpha}_N$ can be written as a recursive estimator (see Duflo,[5] p. 136) as following

$$\hat{\alpha}_{N+1} = \hat{\alpha}_N + S_{N+1}^{-1} D_N (X_{N+1} - D_N^T \hat{\alpha}_N). \tag{6.4}$$

Theorem 6.2. *Consider a functional AR of order p satisfying assumptions A. Let B be an integer multiple of a power of two greater than or equal to the maximum of* $\{2^j(r_j - 1) + L_j, j = 1, \dots, J\}$, $2^J(r_{J+1} - 1) + L_J$ *and p. Let M be an integer greater than* $B - 1$. *Suppose that the covariance matrix* $\Gamma_s = \mathbb{E} X_t^{(s)} (X_t^{(s)})^T$ *is positive definite for each s. Then* $\hat{\alpha}_N$ *defined by* (6.4) *satisfies*:

(i) $\hat{\alpha}_N \xrightarrow[N \to \infty]{a.s} \alpha$ *where* $\alpha = \arg\min_{\theta \in \mathbb{R}^Q} \mathbb{E}[X_{N+1} - D_N^T \theta]^2$ *exists.*

(ii) *If moreover* $(\xi_n)_{n \geq 0}$ *has a moment of order* $m > 4$, *we have a CLT for* $\hat{\alpha}_N$,

$$\sqrt{N}(\hat{\alpha}_N - \alpha) \xrightarrow[N \to \infty]{\mathcal{L}} \mathcal{N}(0, \sigma^2 S^{-1}), \tag{6.5}$$

where $S = \mathcal{R} \mathcal{W} \Gamma_B \mathcal{W}^T \mathcal{R}^T$.

Sketch of Proof: The detailed proof is given in Aminghafari,[1] pp. 113–118. The proof of the almost sure convergence is directly adapted from the proof of Theorem 1 in Poggi and Portier[10] and Theorem 6 of Duflo.[5] The proof of CLT is based on Lemma C1 in Bercu[2] and Theorem 6 of Duflo.[5]

Even if the function f is nonlinear, $\hat{\alpha}_N$ converges towards the parameters of the best prediction by a linear predictor obtained by NDW coefficients of the past observations in the sense of the mean-square error under the stationary distribution. Otherwise, we find the well-known convergence results of the least squares estimator.

Remark 6.1. Let us remark that the two theorems hold true for any orthogonal matrix $\mathcal{W}$.

7. Conclusion

In this paper, we have studied a generalization of a prediction procedure proposed by Renaud *et al.*[11] both from theoretical and practical viewpoints. Focusing on this last issue, the extension to orthogonal wavelets which are more regular than the Haar one and various data-driven ways to design the prediction equation are considered.

The main conclusion is two-fold. First, from the performance viewpoint, it is clear that the new scheme, even if it performs well, is not often better than the simple original method involving the Haar wavelet (the two methods achieve similar results). The reason is certainly that the original method includes directly the approximation components in the prediction equation instead of predicting it together with the deterministic trend by a polynomial extrapolation. This last strategy is not efficient to predict stochastic signals and then could be improved. Second, it should be noted that, if the obtained performance is close to each other, the model can nevertheless be considerably simplified by splitting the signal in two components and by using *db2* or *db3* instead of *haar*. The complexity of the prediction equation generated by the application of the first method to non-stationary signals grows rapidly and the associated parameters can be badly identified and poorly estimated.

Acknowledgments

The first author is grateful to Amirkabir University of Technology, Statistical Research Center and Paris XI University for their supports.

References

1. M. Aminghafari, Méthodes d'ondelettes en statistique des signaux temporels uni et multivariés, PhD thesis, Orsay, France (2006).
2. A. Antoniadis, Wavelet in statistics: A review, *J. Ital. Statist. Soc.* **6** (1997) 97–144.

3. B. Bercu, Central limit theorem and law of iterated logarithm for least squares algorithms in adaptive tracking, *SIAM J. Control Optim.* **36**(3) (1998) 910–928 (electronic).
4. R. Dahlhaus, M. H. Neumann and R. Von Sachs, Nonlinear wavelet estimation of time-varying autoregressive processes, *Bernoulli* **5**(5) (1999) 873–906.
5. M. Duflo, *Random Iterative Models* (Springer–Verlag, 1997).
6. P. Fryzlewicz, S. Van Bellegem and R. Von Sachs, Forecasting non-stationary time series by wavelet process modelling, *Ann. Inst. Statist. Math.* **55**(4) (2003) 737–764.
7. E. J. Hannan, *Multiple Time Series* (John Wiley and Sons, 1970).
8. G. P. Nason, R. Von Sachs and G. Kroisandt, Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum, *J. R. Stat. Soc. Ser. B* **62**(2) (2000) 271–292.
9. D. B. Percival and A. T. Walden, *Wavelet Methods for Time Series Analysis* (Cambridge University Press, 2000).
10. J.-M. Poggi and B. Portier, A test of linearity for functional autoregressive models, *J. Time Ser. Anal.* **18**(6) (1997) 615–639.
11. O. Renaud, J.-L. Starck and F. Murtagh, Prediction based on a multiscale decomposition, *Int. J. Wavelets Multires. Inform. Process.* **1** (2003) 217–232.
12. S. Soltani, D. Boichu, P. Simard and S. Canu, The long-term memory prediction by multiscale decomposition, *Signal Process.* **80** (2000) 2195–2205.