Forecasting non-stationary time series by wavelet process modelling

P. Fryźlewicz[1] S. Van Bellegem[2, 4, *] R. von Sachs[3, 4]

December 16, 2002

Abstract

Many time series in the applied sciences display a time-varying second order structure. In this article, we address the problem of how to forecast these non-stationary time series by means of non-decimated wavelets. Using the class of Locally Stationary Wavelet processes, we introduce a new predictor based on wavelets and derive the prediction equations as a generalisation of the Yule-Walker equations. We propose an automatic computational procedure for choosing the parameters of the forecasting algorithm. Finally, we apply the prediction algorithm to a meteorological time series.

Keywords: Local stationarity, non-decimated wavelets, prediction, time-modulated processes, Yule-Walker equations.
Running head: Forecasting non-stationary processes by wavelets

[1]University of Bristol, Department of Mathematics, Bristol, UK. E-mail: P.Z.Fryzlewicz@bristol.ac.uk

[2]Research Fellow of the National Fund for Scientific Research (F.N.R.S.). Université catholique de Louvain, Institut de statistique, Louvain-la-Neuve, Belgium. E-mail: vanbellegem@stat.ucl.ac.be

[3]Université catholique de Louvain, Institut de statistique, Louvain-la-Neuve, Belgium. E-mail: vonsachs@stat.ucl.ac.be

[4]Financial support from the contract 'Projet d'Actions de Recherche Concertées' nr. 98/03-217 of the Belgian Government and from the IAP research network No. P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs) are gratefully acknowledged.

* *Corresponding author. Address for correspondence: Université catholique de Louvain, Institut de statistique, Voie du Roman Pays, 20, B-1348 Louvain-la-Neuve, Belgium. Fax: +32 10 47.30.32*

1 Introduction

In a growing number of fields, such as biomedical time series analysis, geophysics, telecommunications, or financial data analysis, to name but a few, explaining and inferring from observed serially correlated data calls for non-stationary models of their second order structure. That is, variance and covariance, or equivalently the spectral structure, are likely to change over time.

In this article, we address the problem of whether and how wavelet methods can help in forecasting non-stationary time series. Recently, Antoniadis and Sapatinas (2002) used wavelets for forecasting time-continuous stationary processes. The use of wavelets has proved successful in capturing *local* features of observed data. There arises a natural question of whether they can also be useful for prediction in situations where too little homogeneous structure at the end of the observed data set prevents the use of classical prediction methods based on stationarity. Obviously, in order to develop a meaningful approach, one needs to control this deviation from stationarity, and hence one first needs to think about what kind of non-stationary *models* to fit to the observed data. Let us give a brief overview of the existing possibilities.

Certainly the simplest approach consists in assuming piecewise stationarity, or approximate piecewise stationarity, where the challenge is to find the stretches of homogeneity optimally in a data-driven way (Ombao *et al.*, 2001). The resulting estimate of the time-varying second order structure is, necessarily, rather blocky over time, so some further thoughts on how to cope with these potentially artificially introduced discontinuities are needed. To name a few out of the many models which allow a smoother change over time, we cite the following approaches to the idea of "local stationarity": the work of Mallat *et al.* (1998), who impose bounds on the derivative of the Fourier spectrum as a function of time, and the approaches which allow the coefficients of a parametric model (such as AR) to vary slowly with time (e.g. Mélard and Herteleer-De Schutter (1989), Dahlhaus *et al.* (1999) or Grillenzoni (2000)). The following fact is a starting point for several other more general and more non-parametric approaches: every covariance-stationary process X_t has a Cramér representation

$$X_t = \int_{(-\pi,\pi]} A(\omega) \exp(i\omega t) dZ(\omega), \quad t \in \mathbb{Z}, \tag{1.1}$$

where $Z(\omega)$ is a stochastic process with orthonormal increments. Non-stationary processes are defined by assuming a *slow* change over time of the amplitude $A(\omega)$ (Priestley (1965), Dahlhaus (1997), Ombao *et al.* (2002)). All the above models are of the "time-frequency" type as they use, directly or indirectly, the concept of a time-varying spectrum, being the Fourier transform of a time-varying autocovariance.

The work of Nason, von Sachs and Kroisandt (2000) adopts the concept of local stationarity but replaces the aforementioned spectral representation with respect to the Fourier basis by a representation with respect to non-decimated (or translation-invariant) wavelets. With their model of "Locally Stationary Wavelet" (LSW) processes, the authors introduce a time-scale representation of a stochastic process. The representation allows for a rigorous theory of how to estimate the *wavelet spectrum*, i.e. the coefficients of the resulting representation of the local autocovariance function with respect to *autocorrelation wavelets*. This theory parallels the one developed by Dahlhaus (1997), where rescaling the time argument of the autocovariance and the Fourier spectrum makes it possible to embed the estimation in the non-parametric regression setting, including asymptotic considerations of consistency

and inference. Nason *et al.* (2000) also propose a fast and easily implementable estimation algorithm which accompanies their theory.

As LSW processes are defined with respect to a wavelet system, they have a mean-square representation in the time-scale plane. It is worth recalling that many time series in the applied sciences are believed to have an inherent "multiscale" structure (e.g. financial log-return data, see Calvet and Fisher (2001)). In contrast to Fourier-based models of nonstationarity, the LSW model offers a multiscale representation of the (local) covariance (see Section 2). This representation is often sparse, and thus the covariance may be estimated more easily in practice. The estimator itself is constructed by means of the *wavelet periodogram*, which mimicks the structure of the LSW model and is naturally localised.

Given all these benefits, it seems appropriate to us to use the (linear) LSW model to generalise the stationary approach of forecasting X_t by means of a predictor based on the previous observations up to time $t-1$. While the classical linear predictor can be viewed as based on a non-local Fourier-type representation, our generalisation uses a local wavelet-based approach.

The paper is organised as follows: Section 2 familiarises the reader with the general LSW model, as well as with the particular subclass of time-modulated processes. These are stationary processes modulated by a time-varying variance function, and have proved useful, for instance, in modelling financial log-return series (Van Bellegem and von Sachs (2002)). In the central Section 3, we deal with the theory of prediction for LSW processes, where the construction of our linear predictor is motivated by the approach in the stationary case, i.e. the objective is to minimise the mean-square prediction error (MSPE). This leads to a generalisation of the Yule-Walker equations, which can be solved numerically by matrix inversion or standard iterative algorithms such as the innovations algorithm (Brockwell and Davis, 1991), provided that the non-stationary covariance structure is known. However, the estimation of a non-stationary covariance structure is the main challenge in this context, and this issue is addressed in Section 4. In the remainder of Section 3, we derive an analogue of the classical Kolmogorov formula for the theoretical prediction error, and we generalise the one-step-ahead to h-step-ahead prediction.

Section 4 deals with estimation of the time-varying covariance structure. We discuss some asymptotic properties of our estimators based on the properties of the corrected wavelet periodogram, which is an asymptotically unbiased, but not consistent, estimator of the wavelet spectrum. To achieve consistency, we propose an automatic smoothing procedure, which forms an integral part of our new algorithm for forecasting non-stationary time series. The algorithm implements the idea of adaptive forecasting (see Ledolter (1980)) in the LSW model. In Section 5 we apply our algorithm to a meteorological time series.

We close with a conclusions section and we present our proofs in two appendices. Appendix A contains all the results related to approximating the finite-sample covariance structure of the non-stationary time series by the locally stationary limit. In Appendix B, we show some relevant basic properties of the system of autocorrelation wavelets, and provide the remaining proofs of the statements made in Section 3 and 4.

2 Locally Stationary Wavelet processes

LSW processes are constructed by replacing the amplitude $A(\omega)$ in the Cramér representation (1.1) with a quantity which depends on time (this ensures that the second-order structure of the process changes over time), as well as by replacing the Fourier harmonics $\exp(i\omega t)$

with non-decimated discrete wavelets $\psi_{jk}(t)$, $j = -1, -2, \ldots$, $k \in \mathbb{Z}$. Here, j is the scale parameter (with $j = -1$ denoting the finest scale) and k is the location parameter. Note that unlike decimated wavelets, for which the permitted values of k at scale j are restricted to the set $\{c2^{-j}, c \in \mathbb{Z}\}$, non-decimated wavelets can be shifted to any location defined by the finest resolution scale, determined by the observed data ($k \in \mathbb{Z}$). As a consequence, non-decimated wavelets do not constitute bases for ℓ_2 but overcomplete sets of vectors. The reader is referred to Coifman and Donoho (1995) for an introduction to non-decimated wavelets.

By way of example, we recall the simplest discrete non-decimated wavelet system: the Haar wavelets. They are defined by

$$\psi_{j0}(t) = 2^{j/2}\mathbb{I}_{\{0,1,\ldots,2^{-j-1}-1\}}(t) - 2^{j/2}\mathbb{I}_{\{2^{-j-1},\ldots,2^{-j}-1\}}(t) \quad \text{for } j = -1, -2, \ldots \text{ and } t \in \mathbb{Z}\ ,$$

and $\psi_{jk}(t) = \psi_{j0}(t-k)$ for all $k \in \mathbb{Z}$, where $\mathbb{I}_{\mathcal{A}}(t)$ is 1 if $t \in \mathcal{A}$ and 0 otherwise.

We are now in a position to quote the formal definition of an LSW process from Nason, von Sachs and Kroisandt (2000).

Definition 1. *A sequence of doubly-indexed stochastic processes $X_{t,T}$ $(t = 0, \ldots, T-1)$ with mean zero is in the class of* LSW processes *if there exists a mean-square representation*

$$X_{t,T} = \sum_{j=-J}^{-1} \sum_{k=-\infty}^{\infty} w_{j,k;T}\ \psi_{jk}(t)\ \xi_{jk}, \tag{2.1}$$

where $\{\psi_{jk}(t)\}_{jk}$ is a discrete non-decimated family of wavelets for $j = -1, -2, \ldots, -J$, based on a mother wavelet $\psi(t)$ of compact support and $J = -\min\{j : \mathcal{L}_j \leqslant T\} = O(\log(T))$, where $\mathcal{L}_j$ is the length of support of $\psi_{j0}(t)$. Also,

1. *ξ_{jk} is a random orthonormal increment sequence with $\mathrm{E}\xi_{jk} = 0$ and $\mathrm{Cov}\,(\xi_{jk}, \xi_{\ell m}) = \delta_{j\ell}\ \delta_{km}$ for all j, ℓ, k, m; where $\delta_{j\ell} = 1$ if $j = \ell$ and 0 otherwise;*

2. *For each $j \leqslant -1$, there exists a Lipschitz-continuous function $W_j(z)$ on $(0,1)$ possessing the following properties:*

   - *$\sum_{j=-\infty}^{-1} |W_j(z)|^2 < \infty$ uniformly in $z \in (0,1)$;*
   - *there exists a sequence of constants C_j such that for each T*

     $$\sup_{k=0,\ldots,T-1} \left| w_{j,k;T} - W_j\left(\frac{k}{T}\right)\right| \leqslant \frac{C_j}{T}\ ; \tag{2.2}$$

   - *the constants C_j and the Lipschitz constants L_j are such that $\sum_{j=-\infty}^{-1} \mathcal{L}_j(C_j + \mathcal{L}_j L_j) < \infty$.*

LSW processes are not uniquely determined by the sequence $\{w_{jk;T}\}$. However, Nason *et al.* (2000) develop a theory which defines a unique spectrum. This spectrum measures the power of the process at a particular scale and location. Formally, the *evolutionary wavelet spectrum* of an LSW process $\{X_{t,T}\}_{t=0,\ldots,T-1}$, with respect to ψ, is defined by

$$S_j(z) = |W_j(z)|^2\ ,\ z \in (0,1) \tag{2.3}$$

and is such that, by definition of the process, $S_j(z) = \lim_{T\to\infty} |w_{j,[zT];T}|^2$ for all z in $(0,1)$.

Remark 1 (Rescaled time). In Definition 1, the functions $\{W_j(z)\}_j$ and $\{S_j(z)\}_j$ are defined on the interval $(0,1)$ and not on $\{0,\ldots,T-1\}$. Throughout the paper, we refer to z as the *rescaled time*. This idea goes back to Dahlhaus (1997), who shows that the time-rescaling permits an asymptotic theory of statistical inference for a time-varying Fourier spectrum. The rescaled time is related to the *observed time* $t \in \{0,\ldots,T-1\}$ by the natural mapping $t = [zT]$, which implies that as $T \to \infty$, functions $\{W_j(z)\}_j$ and $\{S_j(z)\}_j$ are sampled on a finer and finer grid. Due to the rescaled time concept, the estimation of the wavelet spectrum $\{S_j(z)\}_j$ is a statistical problem analogous to the estimation of a regression function (see also Dahlhaus (1996a)).

In the classical theory of stationary processes, the spectrum and the autocovariance function are Fourier transforms of each other. To establish an analogous relationship for the wavelet spectrum, observe that the autocovariance function of an LSW process can be written as

$$c_T(z,\tau) = \mathrm{Cov}\left(X_{[zT],T}, X_{[zT]+\tau,T}\right)$$

for $z \in (0,1)$ and τ in $\mathbb{Z}$, and where $[\,\cdot\,]$ denotes the integer part of a real number. The next result shows that this covariance tends to a local covariance as T tends to infinity. Let us introduce the *autocorrelation wavelets* as

$$\Psi_j(\tau) = \sum_{k=-\infty}^{\infty} \psi_{jk}(0)\ \psi_{jk}(\tau)\ , \qquad j<0,\ \tau \in \mathbb{Z}.$$

Some useful properties of the system $\{\Psi_j\}_{j<0}$ can be found in Appendix B. By definition, the *local autocovariance function* of an LSW process with evolutionary spectrum (2.3) is given by

$$c(z,\tau) = \sum_{j=-\infty}^{-1} S_j(z)\Psi_j(\tau) \tag{2.4}$$

for all $\tau \in \mathbb{Z}$ and z in $(0,1)$. In particular, the local variance is given by the multiscale decomposition

$$\sigma^2(z) = c(z,0) = \sum_{j=-\infty}^{-1} S_j(z) \tag{2.5}$$

as $\Psi_j(0) = 1$ for all scales j.

Proposition 1 (Nason *et al.* (2000)). *Under the assumptions of Definition 1, if $T \to \infty$, then $|c_T(z,\tau) - c(z,\tau)| = O(T^{-1})$ uniformly in $\tau \in \mathbb{Z}$ and $z \in (0,1)$.*

Note that formula (2.4) provides a decomposition of the autocovariance structure of the process over scales and rescaled-time locations. In practice, it often turns out that spectrum $S_j(z)$ is only significantly different from zero at a limited number of scales (Fryźlewicz, 2002). If this is the case, then the local autocovariance function $c(z,\tau)$ has a sparse representation and can thus be estimated more easily.

Remark 2 (Stationary processes). A stationary process with an absolutely summable autocovariance function is an LSW process (Nason *et al.*, 2000, Proposition 3). Stationarity

is characterised by a wavelet spectrum which is constant over rescaled time: $S_j(z) = S_j$ for all $z \in (0,1)$.

Remark 3 (Time-modulated processes). Time-modulated (TM) processes constitute a particularly simple class of non-stationary processes. A TM process $X_{t,T}$ is defined as

$$X_{t,T} = \sigma\left(\frac{t}{T}\right) Y_t, \tag{2.6}$$

where Y_t is a zero-mean stationary process with variance one, and the local standard deviation function $\sigma(z)$ is Lipschitz continuous on $(0,1)$ with the Lipschitz constant D. Process $X_{t,T}$ is LSW if

- the autocovariance function of Y_t is absolutely summable (so that Y_t is LSW with a time-invariant spectrum $\{S_j^Y\}_j$);
- and if the Lipschitz constants $L_j^X = D(S_j^Y)^{1/2}$ satisfy the requirements of Definition 1.

If these two conditions hold, then the spectrum $S_j(z)$ of $X_{t,T}$ is given by the formula $S_j(z) = \sigma^2(z) S_j^Y$. The local autocorrelation function $\rho(\tau) = c(z,\tau)/c(z,0)$ of a TM process is independent of z.

However, the real advantage of introducing general LSW processes lies in their ability to model processes whose both variance and autocorrelation function vary over time. Figure 1 shows simulated examples of LSW processes in which the spectrum is only non-zero at a limited number of scales. A sample realisation of a TM process is plotted in Figure 1(c), and Figure 1(d) shows a sample realisation of an LSW process which cannot be modelled as a TM series.

Figure 1 here

3 The predictor and its theoretical properties

In this section, we define and analyse the general linear predictor for non-stationary data that are modelled to follow the LSW process representation given in Definition 1.

3.1 Definition of the linear predictor

Given t observations $X_{0,T}, X_{1,T}, \ldots, X_{t-1,T}$ of an LSW process, we define the *h-step-ahead predictor* of $X_{t-1+h,T}$ by

$$\hat{X}_{t-1+h,T} = \sum_{s=0}^{t-1} b_{t-1-s;T}^{(h)} \, X_{s,T} \ , \tag{3.1}$$

where the coefficients $b_{t-1-s;T}^{(h)}$ are such that they minimise the Mean Square Prediction Error (MSPE). The MSPE is defined by

$$\mathrm{MSPE}(\hat{X}_{t-1+h,T}, X_{t-1+h,T}) = \mathrm{E}\left(\hat{X}_{t-1+h,T} - X_{t-1+h,T}\right)^2 \ .$$

The predictor (3.1) is a linear combination of doubly-indexed observations where the weights need to follow the same doubly-indexed framework. This means that as $T \to \infty$, we augment our knowledge about the local structure of the process, which allows us to fit coefficients $b^{(h)}_{t-1-s;T}$ more and more accurately. The double indexing of the weights is necessary due to the non-stationary nature of the data. This scheme is different to the traditional filtering of the data $X_{s,T}$ by a linear filter $\{\boldsymbol{b}_t\}$. In particular, we do not assume the (square) summability of the sequence $\boldsymbol{b}_t$ because (3.1) is a relation which is written in rescaled time.

The following assumption holds in the sequel of the paper.

Assumption 1. If h is the prediction horizon and t is the number of observed data, then we set $T = t + h$ and we assume $h = o(T)$.

Remark 4 (Prediction domain in the rescaled time). With this assumption, the last observation of the LSW process is denoted by $X_{t-1,T} = X_{T-h-1,T}$, while $\hat{X}_{T-1,T}$ is the last possible forecast (h steps ahead). Consequently, in the rescaled time (see Remark 1), the evolutionary wavelet spectrum $S_j(z)$ can only be estimated on the interval

$$\left[0, 1 - \frac{h+1}{T}\right]. \tag{3.2}$$

The rescaled-time segment

$$\left(1 - \frac{h+1}{T}, 1\right) \tag{3.3}$$

accommodates the predicted values of $S_j(z)$. With Assumption 1, the estimation domain (3.2) asymptotically tends to $[0, 1)$ while the prediction domain (3.3) shrinks to an empty set in the rescaled time. Thus, Assumption 1 ensures that asymptotically, we acquire knowledge of the wavelet spectrum over the full interval $[0, 1)$.

3.2 Prediction in the wavelet domain

There is an interesting link between the above definition of the linear predictor (3.1) and another, "intuitive" definition of a predictor in the LSW model. For ease of presentation, let us suppose the forecasting horizon is $h = 1$, so that $T = t + 1$. Given observations up to time $t - 1$, a natural way of defining a predictor of $X_{t,T}$ is to mimic the structure of the LSW model itself by moving to the wavelet domain. The empirical wavelet coefficients are defined by

$$d_{jk;T} = \sum_{s=0}^{t-1} X_{s,T}\ \psi_{jk}(s)$$

for all $j = -1, \ldots, -J$ and $k \in \mathbb{Z}$. Then, the one-step-ahead predictor is constructed as

$$\hat{X}_{t,T} = \sum_{j=-J}^{-1} \sum_{k \in \mathbb{Z}} d_{jk;T}\ a^{(1)}_{jk;T}\ \psi_{jk}(t)\ , \tag{3.4}$$

where the coefficients $a^{(1)}_{jk}$ have to be estimated and are such that they minimise the MSPE. This predictor (3.4) may be viewed as a projection of $X_{t,T}$ on the space of random variables

spanned by $\{d_{j,k;T} | j = -1, \ldots, -J \text{ and } k = 0, \ldots, T-1\}$.

It turns out that due to the redundancy of the non-orthogonal wavelet system $\{\psi_{jk}(t)\}$, the predictor (3.4) does not have a unique representation: there exists more than one solution $\{a^{(1)}_{jk}\}$ minimising the MSPE, but each solution gives the same predictor (expressed as a different linear combination of the redundant functions $\{\psi_{jk}(t)\}$). One can easily verify this observation by considering, for example, the stationary process $X_s = \sum_{k=-\infty}^{\infty} \psi_{-1k}(s)\zeta_k$, where ψ_{-1} is the non-decimated discrete Haar wavelet at scale -1 and ζ_k is an orthonormal increment sequence.

It is not surprising that the wavelet predictor (3.4) is related to the linear predictor (3.1) by

$$b^{(1)}_{t-s;T} = \sum_{j=-J}^{-1} \sum_{k \in \mathbb{Z}} a^{(1)}_{jk;T} \; \psi_{jk}(t) \; \psi_{jk}(s).$$

Because of the redundancy of the non-decimated wavelet system, for a fixed sequence $b^{(1)}_{t-s;T}$, there exists more than one sequence $a^{(1)}_{jk;T}$ such that this relation holds. For this reason, we prefer to work directly with the general linear predictor (3.1), bearing in mind that it can also be expressed as a (non-unique) projection onto the wavelet domain.

3.3 One-step ahead prediction equations

In this subsection, we consider a forecasting horizon $h = 1$ (so that $T = t + 1$) and want to minimise the mean square prediction error $\text{MSPE}(\hat{X}_{t;T}, X_{t;T})$ with respect to $b^{(1)}_{t-s;T}$. This quadratic function may be written as

$$\text{MSPE}(\hat{X}_{t;T}, X_{t;T}) = \boldsymbol{b}_t' \boldsymbol{\Sigma}_{t;T} \boldsymbol{b}_t \; ,$$

where $\boldsymbol{b}_t$ is the vector $(b^{(1)}_{t-1;T}, \ldots, b^{(1)}_{0;T}, -1)$ and $\boldsymbol{\Sigma}_{t;T}$ is the covariance matrix of $X_{0;T}, \ldots, X_{t;T}$. However, the matrix $\boldsymbol{\Sigma}_{t;T}$ depends on $w^2_{jk;T}$ which cannot be estimated, as they are not identifiable (recall that the representation (2.1) is not unique due to the redundancy of the system $\{\psi_{jk}\}$). The next proposition shows that the MSPE may be approximated by $\boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t$, where $\boldsymbol{B}_{t;T}$ is a $(t+1) \times (t+1)$ matrix whose $(m,n)-$th element is given by

$$\sum_{j=-J}^{-1} S_j \left(\frac{n+m}{2T} \right) \Psi_j(n-m) \; ,$$

and can be estimated by estimating the (uniquely defined) wavelet spectrum S_j. We first consider the following assumptions on the evolutionary wavelet spectrum.

Assumption 2. The evolutionary wavelet spectrum is such that

$$\sum_{\tau=0}^{\infty} \sup_z |c(z,\tau)| < \infty, \tag{3.5}$$

$$C_1 := \operatorname{ess\,inf}_{z,\omega} \sum_{j<0} S_j(z) |\hat{\psi}_j(\omega)|^2 > 0, \tag{3.6}$$

where $\hat{\psi}_j(\omega) = \sum_{s=-\infty}^{\infty} \psi_{j0}(s) \exp(i\omega s)$.

Note that if (3.5) holds, then

$$C_2 := \operatorname*{ess\,sup}_{z,\omega} \sum_{j<0} S_j(z)|\hat{\psi}_j(\omega)|^2 < \infty. \tag{3.7}$$

Assumption (3.5) ensures that for each z, the local covariance $c(z,\tau)$ is absolutely summable, so the process is short-memory (in fact, Assumption (3.5) is slightly stronger than that, for technical reasons). Assumption (3.6) and formula (3.7) become more transparent when we recall that for a stationary process X_t with spectral density $f(\omega)$ and wavelet spectrum S_j, we have $f(\omega) = \sum_j S_j|\hat{\psi}_j(\omega)|^2$ (the Fourier transform of equation (2.4) for stationary processes). In this sense, (3.6) and (3.7) are "time-varying" counterparts of the classical assumptions of the (stationary) spectral density being bounded away from zero, as well as bounded from above.

Proposition 2. *Under Assumptions (3.5) and (3.6), the mean square one-step-ahead prediction error may be written as*

$$\mathrm{MSPE}(\hat{X}_{t;T}, X_{t;T}) = \boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t \ (1 + o_T(1)) \ . \tag{3.8}$$

Moreover, if $\{b^{(1)}_{s;T}\}$ are the coefficients which minimise $\boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t$, then $\{b^{(1)}_{s;T}\}$ solve the following linear system

$$\sum_{m=0}^{t-1} b^{(1)}_{t-1-m;T} \left\{ \sum_{j=-J}^{-1} S_j\left(\frac{n+m}{2T}\right) \Psi_j(m-n) \right\} = \sum_{j=-J}^{-1} S_j\left(\frac{t+n}{2T}\right) \Psi_j(t-n) \tag{3.9}$$

for all $n = 0, \ldots, t-1$.

The proof of the first result can be found in Appendix A (see Lemma 5) and uses standard approximations of covariance matrices of locally stationary processes. The second result is simply the minimisation of the quadratic form (3.8) and the system of equations (3.9) is called the *prediction equations*. The key observation here is that minimising $\boldsymbol{b}_t' \boldsymbol{\Sigma}_{t;T} \boldsymbol{b}_t$ is asymptotically equivalent to minimising $\boldsymbol{b}_t' \boldsymbol{B}_{t;T} \boldsymbol{b}_t$. Bearing in mind the relation of formula (2.4) between the wavelet spectrum and the local autocovariance function, the prediction equations can also be written as

$$\sum_{m=0}^{t-1} b^{(1)}_{t-1-m;T} \, c\left(\frac{n+m}{2T}, m-n\right) = c\left(\frac{n+t}{2T}, t-n\right). \tag{3.10}$$

The following two remarks demonstrate how the prediction equations simplify in the case of two important subclasses of locally stationary wavelet processes.

Remark 5 (Stationary processes). If the underlying process is stationary, then the local autocovariance function $c(z,\tau)$ is no longer a function of two variables, but only a function of τ. In this context, the prediction equations (3.10) become

$$\sum_{m=0}^{t-1} b^{(1)}_{t-1-m} \, c(m-n) = c(t-n)$$

for all $n = 0, \ldots, t-1$, which are the standard Yule-Walker equations used to forecast

stationary processes.

Remark 6 (Time-modulated processes). For the processes considered in Remark 3 (equation (2.6)), the local autocovariance function has a multiplicative structure: $c(z,\tau) = \sigma^2(z)\rho(\tau)$. Therefore, for these processes, prediction equations (3.10) become

$$\sum_{m=0}^{t-1} b^{(1)}_{t-1-m;T}\sigma^2\left(\frac{n+m}{2T}\right)\rho(m-n) = \sigma^2\left(\frac{n+t}{2T}\right)\rho(t-n).$$

We will now study the inversion of the system (3.9) in the general case, and the stability of the inversion. Denote by $\boldsymbol{P}_t$ the matrix of this linear system, i.e.

$$(\boldsymbol{P}_t)_{nm} = \sum_{j=-J}^{-1} S_j\left(\frac{n+m}{2T}\right)\Psi_j(m-n)$$

for $n, m = 0, \ldots, t-1$. Using classical results of numerical analysis (see for instance Kress (1991, Theorem 5.3)) the measure of this stability is given by the so-called *condition number*, which is defined by $\mathrm{cond}\,(\boldsymbol{P}_t) = \|\boldsymbol{P}_t\|\,\|\boldsymbol{P}_t^{-1}\|$. It can be proved along the lines of Lemma 3 (Appendix A) that, under Assumptions (3.5) and (3.6), $\mathrm{cond}\,(\boldsymbol{P}_t) \leqslant C_1\,C_2$.

3.4 The prediction error

The next result generalises the classical Kolmogorov formula for the theoretical one-step-ahead prediction error (Brockwell and Davis, 1991, Theorem 5.8.1). It is a direct modification of a similar result stated by Dahlhaus (1996b, Theorem 3.2(i)) for locally stationary Fourier processes.

Proposition 3. *Suppose that Assumptions (3.5) and (3.6) hold. Given t observations $X_{0,T}, \ldots, X_{t-1,T}$ of the LSW process $\{X_{t,T}\}$ (with $T = t+1$), the one-step ahead mean square prediction error σ^2_{OSPE} in forecasting $\hat{X}_{t,T}$ is given by*

$$\sigma^2_{\mathrm{OSPE}} = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} d\omega\ \ln\left[\sum_{j=-\infty}^{-1} S_j\left(\frac{t}{T}\right)|\hat{\psi}_j(\omega)|^2\right]\right\}(1+o_T(1))\ .$$

Note that due to Assumption (3.6), the sum $\sum_j S_j(t/T)|\hat{\psi}_j(\omega)|^2$ is strictly positive, except possibly on a set of measure zero.

3.5 h-step-ahead prediction

The one-step-ahead prediction equations have a natural generalisation to the h-step-ahead prediction problem with $h > 1$. The mean square prediction error can be written

$$\mathrm{MSPE}(\hat{X}_{t+h-1,T}, X_{t+h-1,T}) = \mathrm{E}\left(\hat{X}_{t+h-1,T} - X_{t+h-1,T}\right)^2 = \boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1},$$

where $\boldsymbol{\Sigma}_{t+h-1;T}$ is the covariance matrix of $X_{0,T}, \ldots, X_{t+h-1,T}$ and $\boldsymbol{b}_{t+h-1}$ is the vector $(b^{(h)}_{t-1}, \ldots, b^{(h)}_0, b^{(h)}_{-1}, \ldots, b^{(h)}_{-h})$, with $b^{(h)}_{-1}, \ldots, b^{(h)}_{-h+1} = 0$ and $b^{(h)}_{-h} = -1$. Like before, we approximate the mean square error by $\boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1}$, where $\boldsymbol{B}_{t+h-1;T}$ is a $(t+h)\times(t+h)$

matrix whose (m,n)-th element is given by

$$\sum_{j=-J}^{-1} S_j \left(\frac{n+m}{2T}\right) \Psi_j(n-m) \ .$$

Proposition 4. *Under Assumptions (3.5) and (3.6), the mean square prediction error may be written as*

$$\mathrm{MSPE}(\hat{X}_{t+h-1;T}, X_{t+h-1;T}) = \boldsymbol{b}'_{t+h-1} \boldsymbol{B}_{t+h-1;T} \boldsymbol{b}_{t+h-1} \ (1 + o_T(1)) \ .$$

4 Prediction based on data

Having treated the prediction problem from a theoretical point of view, we now address the question of how to estimate the unknown time-varying second order structure in the system of equations (3.9). In Subsection 4.3, we propose a complete algorithm for forecasting non-stationary time series using the LSW framework.

4.1 Estimation of the time-varying second-order structure

Our estimator of the local autocovariance function $c(z,\tau)$, with $0 < z < t/T$, is constructed by estimating the unknown wavelet spectrum $S_j(z)$ in the multiscale representation (2.4). Let us first define the function $J(t) = -\min\{j : \mathcal{L}_j \leqslant t\}$. Following Nason *et al.* (2000) we define the *wavelet periodogram* as the sequence of squared wavelet coefficients $d_{jk;T}$, where j and k are scale and location parameters, respectively:

$$I_j(k/T) = d^2_{jk;T} = \left(\sum_{s=0}^{t-1} X_{s,T} \ \psi_{jk}(s) \right)^2 \ , \quad -J(t) \leqslant j \leqslant -1, \ k = \mathcal{L}_j - 1, \ldots, t-1 \ .$$

Note that as ψ_{jk} is only nonzero for $s = 0, \ldots, \mathcal{L}_j - 1$, the estimator $I_j(k/T)$ is a function of $X_{t,T}$ for $t \leqslant k$. At the left edge, we set $I_j(k/T) = I_j((\mathcal{L}_j - 1)/T)$ for $k = 0, \ldots, \mathcal{L}_j - 2$.

From this definition, we define our multiscale estimator of the local variance function (2.5) as

$$\tilde{c}\left(\frac{k}{T}, 0\right) = \sum_{j=-J}^{-1} 2^j \ I_j\left(\frac{k}{T}\right) . \tag{4.1}$$

The next proposition concerns the asymptotic behaviour of the first two moments of this estimator.

Proposition 5. *The estimator (4.1) satisfies*

$$\mathrm{E}\ \tilde{c}\left(\frac{k}{T}, 0\right) = c\left(\frac{k}{T}, 0\right) + O\left(T^{-1}\log(T)\right) .$$

If, in addition, the increment process $\{\xi_{jk}\}$ in Definition 1 is Gaussian and (3.5) holds, then

$$\text{Var } \tilde{c}\left(\frac{k}{T},0\right) = 2\sum_{i,j=-J}^{-1} 2^{i+j}\left(\sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{in}(\tau)\psi_{jn}(0)\right)^2 + O(T^{-1}).$$

Remark 7 (Time-modulated processes). For Gaussian time-modulated processes considered in Remark 3 (formula (2.6)), the variance of estimator (4.1) reduces to

$$\text{Var } \tilde{c}\left(\frac{k}{T},0\right) = 2\sigma^4(k/T)\sum_{i,j=-J}^{-1} 2^{i+j}\left(\sum_{\tau}\rho(\tau)\sum_{n}\psi_{in}(\tau)\psi_{jn}(0)\right)^2 + O(T^{-1}), \tag{4.2}$$

where $\rho(\tau)$ is the autocorrelation function of Y_t (see equation (2.6)). If $X_{t,T} = \sigma(t/T)Z_t$, where Z_t are i.i.d. $N(0,1)$, then the leading term in (4.2) reduces to $(2/3)\sigma^4(k/T)$ for all compactly supported wavelets ψ. Other possible estimators of the local variance for time-modulated processes, as well as an empirical study of the explanatory power of these models as applied to financial time series, may be found in Van Bellegem and von Sachs (2002).

Remark 8. Proposition 5 can be generalised for the estimation of $c(z,\tau)$ for $\tau \neq 0$. Define the estimator

$$\tilde{c}\left(\frac{k}{T},\tau\right) = \sum_{j=-J}^{-1}\left(\sum_{\ell=-J}^{-1} A^{-1}_{j\ell}\Psi_\ell(\tau)\right) I_j\left(\frac{k}{T}\right), \quad k=0,\ldots,t-1, \tau\neq 0, \tag{4.3}$$

where the matrix $\boldsymbol{A} = (A_{j\ell})_{j,\ell<0}$ is defined by

$$A_{j\ell} := \langle \Psi_j, \Psi_\ell\rangle = \sum_{\tau}\Psi_j(\tau)\,\Psi_\ell(\tau)\,. \tag{4.4}$$

Note that the matrix $A_{j\ell}$ is not simply diagonal due to the redundancy in the system of autocorrelation wavelets $\{\Psi_j\}$. Nason *et al.* (2000) proved the invertibility of $\boldsymbol{A}$ if $\{\Psi_j\}$ is constructed using Haar wavelets. If other compactly supported wavelets are used, numerical results suggest that the invertibility of $\boldsymbol{A}$ still holds, but a complete proof of this result has not been established yet. Using Lemma 8, it is possible to generalise the proof of Proposition 5 for Haar wavelets to show that

$$\text{E } \tilde{c}\left(\frac{k}{T},\tau\right) = c\left(\frac{k}{T},\tau\right) + O\left(T^{-1/2}\right)$$

for $\tau \neq 0$ and, if Assumption (3.5) hold and if the increment process $\{\xi_{jk}\}$ in Definition 1 is Gaussian, then

$$\text{Var } \tilde{c}\left(\frac{k}{T},\tau\right) = 2\sum_{i,j=-J}^{-1} h_i(\tau)h_j(\tau)\left\{\sum_{\tau} c\left(\frac{k}{T},\tau\right)\sum_{n}\psi_{in}(\tau)\psi_{jn}(0)\right\}^2 + O\left(T^{-1}\log^2(T)\right)$$

for $\tau \neq 0$, where $h_j(\tau) = \sum_{\ell=-J}^{-1} A^{-1}_{j\ell}\Psi_\ell(\tau)$.

These results show the inconsistency of the estimator of the local (co)variance, which needs to be smoothed w.r.t. the rescaled time z (i.e. $\tilde{c}(\cdot,\tau)$ needs to be smoothed for all

τ). We use standard kernel smoothing where the problem of the choice of the bandwidth parameter g arises. The goal of Subsection 4.3 is to provide a fully automatic procedure for choosing g.

To compute the linear predictor in practice, we invert the generalised Yule-Walker equations (3.10) in which the theoretical local autocovariance function is replaced by the smoothed version of $\tilde{c}(k/T, \tau)$. However, in equations (4.1) and (4.3), our estimator is only defined for $k = 0, \ldots, t-1$ while the prediction equations (3.10) require the local autocovariance up to $k = t$ (for $h = 1$). This problem is inherent to our non-stationary framework. We denote the predictor of $c(t/T, \tau)$ by $\hat{c}(t/T, \tau)$ and, motivated by the slow evolution of the local autocovariance function, propose to compute $\hat{c}(t/T, \tau)$ by the local smoothing of the (unsmoothed) estimators $\{\tilde{c}(k/T, \tau), k = t-1, \ldots, t-\mu\}$. In practice, the smoothing parameter μ for prediction is set to be equal to gT, where g is the smoothing parameter (bandwidth) for estimation. They can be obtained by the data-driven procedure described in Subsection 4.3.

4.2 Future observations in rescaled time

For clarity of presentation, we restrict ourselves (in this and the following subsection) to the case $h = 1$.

In remarks 1 and 4, we recalled the mechanics of rescaled time for non-stationary processes. An important ingredient of this concept is that the data come in the form of a triangular array whose rows correspond to *different* stochastic processes, only linked through the asymptotic wavelet spectrum sampled on a finer and finer grid. This mechanism is inherently different to what we observe in practice, where, typically, observations arrive one by one and neither the values of the "old" observations, nor their corresponding second-order structure, change when a new observation arrives.

One way to reconcile the practical setup with our theory is to assume that for an observed process $X_0, \ldots, X_{t-1}$, there exists a doubly-indexed LSW process $\mathbf{Y}$ such that $X_k = Y_{k,T}$ for $k = 0, \ldots, t-1$. When a new observation X_t arrives, the underlying LSW process changes, i.e. there exists another LSW process $\mathbf{Z}$ such that $X_k = Z_{k,T+1}$ for $k = 0, \ldots, t$. An essential point underlying our adaptive algorithm of the next subsection is that the spectra of $\mathbf{Y}$ and $\mathbf{Z}$ are close to each other, due to the above construction and the regularity assumptions imposed by Definition 1 (in particular, the Lipschitz continuity of $S_j(z)$).

The objective of our algorithm is to choose appropriate values of certain nuisance parameters (see the next subsection) in order to forecast X_t from $X_0, \ldots, X_{t-1}$. Assume that these parameters have been selected well, i.e. that the forecasting has been successful. The closeness of the two spectra implies that we can also expect to successfully forecast X_{t+1} from $X_0, \ldots, X_t$ using the same, or possibly "neighbouring", values of the nuisance parameters.

Bearing in mind the above discussion, we introduce our algorithm with a slight abuse of notation: we drop the second subscript when referring to the observed time series.

4.3 Data-driven choice of parameters

In theory, the best one-step-ahead linear predictor of $X_{t,T}$ is given by (3.1), where $\boldsymbol{b}_t = (b^{(1)}_{t-1-s;T})_{s=0,\ldots,t-1}$ solves the prediction equations (3.9). In practice, each of the t components of the vector $\boldsymbol{b}_t$ is estimated using our estimator of the local autocovariance function based on observations $X_{0,T}, \ldots, X_{t-1,T}$. Hence, we have to find a balance between the estimation

error, potentially increasing with t, and the prediction error which is a decreasing function of t.

As a natural balancing rule which works well in practice, we suggest to choose a number p such that the "clipped" predictor

$$\hat{X}_{t,T}^{(p)} = \sum_{s=t-p}^{t-1} b_{t-1-s;T}^{(1)} X_{s,T} \tag{4.5}$$

gives a good compromise between the theoretical prediction error and the estimation error. The construction (4.5) is reminiscent of the classical idea of AR(p) approximation for stationary processes.

We propose an automatic procedure for selecting the two nusiance parameters: the order p in (4.5) and the bandwidth g, necessary to smooth the inconsistent estimator $\tilde{c}(z,\tau)$ using a kernel method. The idea of this procedure is to start with some initial values of p and g and to gradually update these parameters using a criterion which measures how well the series gets predicted using a given pair of parameters. This type of approach is in the spirit of *adaptive forecasting* (Ledolter, 1980).

Suppose that we observe the series up to X_{t-1} and want to predict X_t, using an appropriate pair (p,g). The idea of our method is as follows. First, we move backwards by s observations and choose some initial parameters (p_0,g_0) for predicting X_{t-s} from the observed series up to X_{t-s-1}. Next, we compute the prediction of X_{t-s} using the pairs of parameters around our preselected pair (i.e. $(p_0-1,g_0-\delta)$, $(p_0,g_0-\delta)$, $\ldots$, $(p_0+1,g_0+\delta)$ for a fixed constant δ). As the true value of X_{t-s} is known, we are able to use a preset criterion to compare the 9 obtained prediction results, and we choose the pair corresponding to the best predictor (according to this preset criterion). This step is called the *update of the parameters* by predicting X_{t-s}. In the next step, the updated pair is used as the initial parameters, and itself updated by predicting X_{t-s+1} from $X_0,\ldots,X_{t-s}$. By applying this procedure to predict $X_{t-s+2}, X_{t-s+3},\ldots,X_{t-1}$, we finally obtain an updated pair (p_1,g_1) which is selected to perform the actual prediction.

Many different criteria can be used to compare the quality of the pairs of parameters at each step. Denote by $\hat{X}_{t-i}(p,g)$ the predictor of X_{t-i} computed using pair (p,g), and by $I_{t-i}(p,g)$ the corresponding 95% *prediction interval* based on the assumption of Gaussianity:

$$I_{t-i}(p,g) = \left[-1.96\hat{\sigma}_{t-i}(p,g) + \hat{X}_{t-i}(p,g)\ , 1.96\hat{\sigma}_{t-i}(p,g) + \hat{X}_{t-i}(p,g)\right], \tag{4.6}$$

where $\hat{\sigma}^2_{t-i}(p,g)$ is the estimate of $\mathrm{MSPE}(\hat{X}_{t-i}(p,g), X_{t-i})$ computed using formula (3.8) with the remainder neglected. The criterion which we use in the simulations reported in the next section is to compute

$$\frac{\left|X_{t-i} - \hat{X}_{t-i}(p,g)\right|}{\mathrm{length}(I_{t-i}(p,g))}$$

for each of the 9 pairs at each step of the procedure and select the updated pair as the one that minimises this ratio.

We also need to choose the initial parameters (p_0,g_0) and the number s of data points at the end of the series which are used in the procedure. We suggest that s should be set to the length of the largest segment at the end of the series which does not contain any apparent breakpoints observed after a visual inspection. To avoid dependence on the initial values

(p_0, g_0), we suggest to iterate the algorithm a few times, using (p_1, g_1) as the initial value for each iteration. We propose to stop when the parameters (p_1, g_1) are such that at least 95% of the observations fall into the prediction intervals.

In order to be able to use our procedure completely on-line, we do not have to repeat the whole algorithm. Indeed, when observation X_t becomes available, we only have to update the pair (p_1, g_1) by predicting X_t, and we directly obtain the "optimal" pair for predicting X_{t+1}.

There are, obviously, many possible variants of our algorithm. Possible modifications include, for example, using a different criterion, restricting the allowed parameter space for (p, g), penalising certain regions of the parameter space, or allowing more than one parameter update at each time point.

We have tested our algorithm on numerous examples, and the following section presents an application to a real data set. A more theoretical study of this algorithm is left for future work.

5 Application of the general predictor to real data

El Niño is a disruption of the ocean atmosphere system in the tropical Pacific which has important consequences for the weather around the globe. Even though the effect of El Niño is not avoidable, research on its forecast and its impacts allows specialists to attenuate or prevent its harmful consequences (see Philander (1990) for a detailed overview). The effect of the equatorial Pacific meridional reheating may be measured by the deviation of the wind speed on the ocean surface from its average. It is worth mentioning that this effect is produced by conduction, and thus we expect the wind speed variation to be smooth. This legitimates the use of LSW processes to model the speed. In this section, we study the wind speed anomaly index, i.e. its standardised deviation from the mean, in a specific region of the Pacific (12-2N, 160E-70W). Modelling this anomaly helps to understand the effect of El Niño effect in that region. The time series composed of $T = 910$ monthly observations is available free of charge at `http://tao.atmos.washington.edu/data_sets/eqpacmeridwindts`. Figure 2(a) shows the plot of the series.

Figure 2 here

Throughout this section, we use Haar wavelets to estimate the local (co)variance. Having provisionally made a safe assumption of the possible non-stationarity of the data, we first attempt to find a suitable pair of parameters (p, g) which will be used for forecasting the series. By inspecting the acf of the series, and by trying different values of the bandwidth, we have found that the pair $(7, 70/T)$ works well for many segments of the data; indeed, the segment of 100 observations from June 1928 to October 1936 gets predicted very accurately in one-step-ahead prediction: 96% of the actual observations are contained in the corresponding 95% prediction intervals (formula (4.6)).

However, the pair $(7, 70/T)$ does not appear to be uniformly well suited for forecasting the whole series. For example, in the segment of 40 observations between November 1986 and February 1990, only 5% of the observations fall into the corresponding one-step-ahead prediction intervals computed using the above pair of parameters. This provides strong evidence that the series is non-stationary (indeed, if it was stationary, we could expect to

obtain a similar percentage of accurately predicted values in both segments). This further justifies our approach of modelling and forecasting the series as an LSW process.

Motivated by the above observation, we now apply our algorithm, described in the previous section, to the segment of 40 observations mentioned above, setting the initial parameters to $(7, 70/T)$. After the first iteration along the segment, the parameters drift up to $(14, 90/T)$, and 85% of the observations fall within the prediction intervals, which is indeed a dramatic improvement over the 5% obtained without applying our adaptive algorithm. In the second pass, we set the initial values to $(14, 90/T)$, and obtain a 92.5% coverage by the one-step-ahead prediction intervals, with the parameters drifting up to $(14, 104/T)$. In the last iteration, we finally obtain a 95% coverage, and the parameters get updated to $(14, 114/T)$. We now have every reason to believe that this pair of parameters is well suited for one-step-ahead prediction within a short distance of February 1990. Without performing any further updates, we apply the one-step-ahead forecasting procedure to predict, one by one, the eight observations which follow February 1990, the prediction parameters being fixed at $(14, 114/T)$. The results are plotted in Figure 2(b), which also compares our results to those obtained by means of AR modelling. At each time point, the order of the AR process is chosen as the one that minimises the AIC criterion, and then the parameters are estimated by means of the standard S-Plus routine. We observe that for both models, all of the true observed values fall within the corresponding one-step-ahead prediction intervals. However, the main gain obtained using our procedure is that the prediction intervals are on average 17.45% narrower in the case of our algorithm. This result is not peculiar to AR modelling as this percentage is also similar in comparison with other stationary models, like ARMA(2,10), believed to accurately fit the series. A similar phenomenon has been observed at several other points of the series.

Figure 3 here

We end this section by applying our general prediction method to compute multi-step-ahead forecasts. Figure 3 shows the 1- up to 9-step-ahead forecasts of the series, along with the corresponding prediction intervals, computed at the end of the series (December 1995). In Figure 3(a), the LSW model is used to construct the forecast values, with parameters $(10, 2.18)$ chosen automatically by our adaptive algorithm described above. Figure 3(b) shows the 9-step-ahead prediction based on AR modelling (here, AR(2)). The prediction in Figure 3(a) looks "smoother" because it uses the information from the whole series. This information is averaged out, whereas in the LSW forecast, local information is picked up at the end of the series, and the forecasts look more "jagged".

6 Conclusion

In this paper, we have given an answer to the pertinent question, asked by time series analysts over the past few years, of whether and how wavelet methods can help in forecasting non-stationary time series. To develop the forecasting methodology, we have considered the Locally Stationary Wavelet (LSW) model, which is based on the idea of a localised time-scale representation of a time-changing autocovariance function. This model includes the class of second-order stationary processes and has several attractive features, not only for modelling, but also for estimation and prediction purposes. Its linearity and the fact that the time-varying second order quantities are modelled as smooth functions, have enabled

us to formally extend the classical theory of linear prediction to the whole class of LSW processes. These results are a generalisation of the Yule-Walker equations and, in particular, of Kolmogorov's formula for the one-step-ahead prediction error.

In the empirical prediction equations the second-order quantities have to be estimated, and this is where the LSW model proves most useful. The rescaled time, one of the main ingredients of the model, makes it possible to develop a rigorous estimation theory. Moreover, by using well-localised non-decimated wavelets instead of a Fourier based approach, our estimators are able to capture the local time-scale features of the observed non-stationary data very well (Nason and von Sachs, 1999).

In practice, our new prediction methodology depends on two nuisance parameters which arise in the estimation of the local covariance and the mean-square prediction error. More specifically, we need to smooth our inconsistent estimators over time, and to do so, we have to choose the bandwidth of the smoothing kernel. Moreover, we need to reduce the dimension of the prediction equations to avoid too much inaccuracy of the resulting prediction coefficients due to estimation errors. We have proposed an automatic computational procedure for selecting these two parameters. Our algorithm is in the spirit of adaptive forecasting as it gradually updates the two parameters basing on the success of prediction. This new method is not only essential for the success of our whole prediction methodology, it also seems to be promising in a much wider context of choosing nuisance parameters in non-parametric methods in general.

We have applied our new algorithm to a meteorological data set. Our non-parametric forecasting algorithm shows interesting advantages over the classical parametric alternative (AR forecasting). Moreover, we believe that one of the biggest advantages of our new algorithm is that it can be successfully applied to a variety of data sets, ranging from financial log-returns (Fryźlewicz (2002), Van Bellegem and von Sachs (2002)) to series traditionally modelled as ARMA processes, including in particular data sets which are not, or do not appear to be, second-order stationary. The S-Plus routines implementing our algorithm, as well as the data set, can be downloaded from the associated web page

`http://www.stats.bris.ac.uk/~mapzf/flsw/flsw.html`

In the future, we intend to derive the theoretical properties of our automatic algorithm for choosing the nuisance parameters of the adaptive predictor. Finally, our approach offers the attractive possibility to use the prediction error for model selection purposes. LSW processes are constructed using a fixed wavelet system, e.g. Haar or another Daubechies' system. It is clear that we can compare the fitting quality of each such model by comparing its prediction performance on the observed data. In the future, we intend to investigate this in more detail in order to answer the question, left open by Nason *et al.* (2000), of which wavelet basis to use to model a given series.

7 Acknowledgements

The authors would like to thank Rainer Dahlhaus, Christine De Mol, Christian Hafner, Guy Nason and two anonymous referees for stimulating discussions and suggestions which helped to improve the presentation of this article. They are also grateful to Peter Brockwell and Brandon Whitcher for their expertise and advice in analysing the wind data of Section 5. Sébastien Van Bellegem and Rainer von Sachs would like to express their gratitude to the

Department of Mathematics, University of Bristol, and Piotr Fryźlewicz — to the Institut de statistique, Université catholique de Louvain, and SVB, for their hospitality during mutual visits in 2001 and 2002. RvS and SVB were funded from the National Fund for Scientific Research – Wallonie, Belgium (F.N.R.S.), by the contract "Projet d'Actions de Recherche Concertées" no. 98/03-217 of the Belgian Government and by the IAP research network No. P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs). PF was funded by the Department of Mathematics at the University of Bristol, Universities UK, Unilever Research, and Guy Nason.

A Theoretical properties of the predictor

Let $\boldsymbol{X}_{t;T} = (X_{0;T}, \ldots, X_{t-1;T})'$ be a realisation of an LSW process. In this appendix, we study the theoretical properties of the covariance matrix $\boldsymbol{\Sigma}_{t;T} = \mathrm{E}(\boldsymbol{X}_{t;T}\boldsymbol{X}'_{t;T})$. As we need upper bounds for the spectral norms $\|\boldsymbol{\Sigma}_{t;T}\|$ and $\|\boldsymbol{\Sigma}_{t;T}^{-1}\|$, we base the following results and their proofs on methods developed in Dahlhaus (1996b, Section 4) for approximating covariance matrices of locally stationary Fourier processes. However, in our setting these methods need important modifications. The idea is to approximate $\boldsymbol{\Sigma}_{t;T}$ by overlapping block Toeplitz matrices along the diagonal.

The approximating matrix is constructed as follows. First, we construct a coverage of the time axis $[0, T)$. Let L be a divisor of T such that $L/T \to 0$, and consider the following partition of the time axis:

$$\mathcal{P}_0 = \Big\{[0, L), [L, 2L), \ldots, [T - L, T)\Big\}.$$

Then, consider another partition of the time axis, which is a shift of $\mathcal{P}_0$ by $\delta < L$:

$$\mathcal{P}_1 = \Big\{[0, \delta), [\delta, L + \delta), [L + \delta, 2L + \delta), \ldots, [T - L + \delta, T)\Big\}.$$

In what follows, assume that L is a multiple of δ and that $\delta/L \to 0$ as T tends to infinity. Also, consider the partition of the time axis which is a shift of $\mathcal{P}_1$ by δ:

$$\mathcal{P}_2 = \Big\{[0, 2\delta), [2\delta, L + 2\delta), [L + 2\delta, 2L + 2\delta), \ldots, [T - L + 2\delta, T)\Big\}$$

and, analogously, define $\mathcal{P}_3, \mathcal{P}_4, \ldots$ up to $\mathcal{P}_M$ where $M = (L/\delta) - 1$. Consider the union of all these partitions $\mathcal{P} = \{\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_M\}$, which is a highly redundant coverage of the time axis. Denote by P the number of intervals in $\mathcal{P}$, and denote the elements of $\mathcal{P}$ by M_p, $p = 1, \ldots, P$.

For each p, we fix a point ν_p in M_p and consider matrix $\boldsymbol{D}^{(p)}$ defined by:

$$D^{(p)}_{nm} = \sum_{j<0} S_j\left(\frac{\nu_p}{T}\right)\Psi_j(n - m)\mathbb{I}_{n,m\in M_p}$$

where $\mathbb{I}_{n,m\in M_p}$ means that we only include those n, m that are in M_p. Observe that each ν_p is contained exactly in L/δ segments. The following lemma concerns the approximation of

$\boldsymbol{\Sigma}_{t;T}$ by matrix $\boldsymbol{D}$ defined by

$$D_{nm} = \frac{\delta}{L} \sum_{p=1}^{P} D_{nm}^{(p)}.$$

Lemma 1. *Assume that (3.5) holds. If $L \to \infty$, $\delta/L \to 0$ and $L^2/T \to 0$ as $T \to \infty$, then*

$$\boldsymbol{x}' \left(\boldsymbol{\Sigma}_{t;T} - \boldsymbol{D} \right) \boldsymbol{x} = \boldsymbol{x}' \boldsymbol{x} o_T(1).$$

Proof. Define matrix $\boldsymbol{\Sigma}_{t;T}^{(p)}$ by $\left(\Sigma_{t;T}^{(p)}\right)_{nm} = \left(\Sigma_{t;T}\right)_{nm} \mathbb{I}_{n,m \in M_p}$. Straightforward calculations yield

$$\boldsymbol{x}' \left(\boldsymbol{\Sigma}_{t;T} - \boldsymbol{D} \right) \boldsymbol{x} = \boldsymbol{x}' \left[\frac{\delta}{L} \sum_{p=1}^{P} (\boldsymbol{\Sigma}_{t;T}^{(p)} - \boldsymbol{D}^{(p)}) \right] \boldsymbol{x} + \mathrm{Rest}_T \tag{A.1}$$

where

$$\mathrm{Rest}_T = \sum_{n,m=0}^{\frac{T}{\delta}-1} \min\left(|n-m| \frac{\delta}{L}, 1 \right) \sum_{u,s=0}^{\delta-1} x_{n\delta+u} \left(\Sigma_{t;T}\right)_{n\delta+u, m\delta+s} x_{m\delta+s}.$$

Let us first bound this remainder. Replace $\left(\Sigma_{t;T}\right)_{nm}$ by $\sum_j S_j((n+m)/2T)\Psi_j(n-m)$ and denote $b(k) := \sup_z |\sum_j S_j(z)\Psi_j(k)| = \sup_z |c(z,k)|$. We have

$$\begin{aligned} |\,\mathrm{Rest}_T\,| &\leqslant 2\boldsymbol{x}'\boldsymbol{x} \sum_{d=1}^{\frac{T}{\delta}-1} \min\left(d\frac{\delta}{L}, 1 \right) \sum_{k=(d-1)\delta+1}^{d\delta} b(k) + \mathrm{Rest}'_T \\ &\leqslant 2\boldsymbol{x}'\boldsymbol{x} \left(\frac{\delta+\sqrt{L}}{L} \sum_{k=1}^{\infty} b(k) + \sum_{k>\sqrt{L}} b(k) \right) + \mathrm{Rest}'_T \end{aligned}$$

and the main term in the above is $o_T(1)$ since $L \to \infty$ and $\delta/L \to 0$ as $T \to \infty$, and by assumption (3.5). Let us now turn to the remainder Rest'_T. We have

$$\mathrm{Rest}'_T \leqslant \sum_{n,m=0}^{T-1} \left| x_n x_m \sum_{j,k} \left(w_{jk;T}^2 - S_j\left(\frac{n+m}{2T} \right) \right) \psi_{j,k}(m) \psi_{j,k}(n) \right|$$

which may be bounded as follows using the definition of an LSW process, and the Lipschitz property of S_j:

$$\begin{aligned} \mathrm{Rest}'_T &\leqslant O(T^{-1}) \sum_j (C_j + \mathcal{L}_j L_j) \sum_k \left(\sum_{n=k-\mathcal{L}_j+1}^{k} |x_n \psi_{j,k}(n)| \right)^2 \\ &\leqslant O(T^{-1}) \boldsymbol{x}'\boldsymbol{x} \sum_j (C_j + \mathcal{L}_j L_j) \mathcal{L}_j \leqslant O(T^{-1}) \boldsymbol{x}'\boldsymbol{x} \end{aligned}$$

by assumption of the Lemma.

Let us finally consider the main term in (A.1). We have

$$\begin{aligned}\boldsymbol{x}'\left(\frac{\delta}{L}\sum_{p=1}^{P}\boldsymbol{\Sigma}_{t;T}^{(p)}-\boldsymbol{D}^{(p)}\right)\boldsymbol{x} &\leqslant \frac{\delta}{L}\sum_{p=1}^{P}\sum_{jk}\left|w_{jk;T}^2-S_j\left(\frac{\nu_p}{T}\right)\right|\left(\sum_u \psi_{j,k}(u)x_u\mathbb{I}_{u\in M_p}\right)^2\\ &\leqslant O(T^{-1})\frac{\delta}{L}\sum_{p=1}^{P}\sum_{jk}\left(\sum_n x_n^2\mathbb{I}_{n\in M_p}\right)\sum_j C_j+L_j(\mathcal{L}_j+L)\\ &= O(T^{-1})\boldsymbol{x}'\boldsymbol{x}\sum_j(C_j+L_j(\mathcal{L}_j+L))(\mathcal{L}_j+L) \qquad \text{(A.2)}\end{aligned}$$

where the last equality holds because, by construction, each x_n is contained in exactly L/δ segments of the coverage. Since we assumed that $L^2/T\to 0$ as $T\to\infty$, we obtain the result. □

Lemma 2. *Assume that (3.5) holds and there exists a t^* such that $x_u = 0$ for all $u \notin \{t^*,\dots,t^*+L\}$. Then for each $t_0\in\{t^*,\dots,t^*+L\}$,*

$$\boldsymbol{x}'\Sigma_{t,T}\boldsymbol{x}=\sum_j S_j\left(\frac{t_0}{T}\right)\sum_k\left(\sum_{u=t^*}^{t^*+L}x_u\psi_{j,k}(u)\right)^2+\boldsymbol{x}'\boldsymbol{x}O\left(\frac{L^2}{T}\right). \qquad \text{(A.3)}$$

Proof. Identical to the part of the proof of Lemma 1 leading to the bound for the main term, i.e. formula (A.2). □

In what follows, the matrix norm $\|\boldsymbol{M}\|$ denotes the spectral norm of the matrix $\boldsymbol{M}$, i.e. $\max\{\sqrt{\lambda} : \lambda \text{ is the eigenvalue of } \boldsymbol{M}'\boldsymbol{M}\}$. If $\boldsymbol{M}$ is symmetric and nonnegative definite, by standard theory we have

$$\|\boldsymbol{M}\|=\sup_{\|\boldsymbol{x}\|_2^2=1}\boldsymbol{x}'\boldsymbol{M}\boldsymbol{x} \qquad \|\boldsymbol{M}^{-1}\|=\left(\inf_{\|\boldsymbol{x}\|_2^2=1}\boldsymbol{x}'\boldsymbol{M}\boldsymbol{x}\right)^{-1}. \qquad \text{(A.4)}$$

Lemma 3. *Assume that (3.5) holds. The spectral norm $\|\boldsymbol{\Sigma}_{t;T}\|$ is bounded in t. Also, if (3.6) holds, then the spectral norm $\|\boldsymbol{\Sigma}_{t;T}^{-1}\|$ is bounded in t.*

Proof. Lemma 1 implies

$$\|\boldsymbol{\Sigma}_{t;T}\|=\sup_{\|\boldsymbol{x}\|_2^2=1}\frac{\delta}{L}\sum_{p=1}^{P}\sum_{j<0}S_j\left(\frac{\nu_p}{T}\right)\sum_k\left(\sum_n x_n\psi_{j,k-n}\mathbb{I}_{n\in M_p}\right)^2+o_T(1)$$

using Parseval formula, we have

$$\begin{aligned}&=\sup_{\|\boldsymbol{x}\|_2^2=1}\frac{\delta}{2\pi L}\sum_{p=1}^{P}\int_{-\pi}^{\pi}d\omega\ \sum_{j<0}S_j\left(\frac{\nu_p}{T}\right)\left|\hat{\psi}_j(\omega)\right|^2\left|\sum_n x_n\exp(-i\omega n)\mathbb{I}_{n\in M_p}\right|^2+o_T(1)\\ &\leqslant \operatorname*{ess\,sup}_{z,\omega}\sum_j S_j(z)\left|\hat{\psi}_j(\omega)\right|^2\sup_{\|\boldsymbol{x}\|_2^2=1}\|\boldsymbol{x}\|_2^2+o_T(1)=\operatorname*{ess\,sup}_{z,\omega}\sum_j S_j(z)\left|\hat{\psi}_j(\omega)\right|^2+o_T(1)\end{aligned}$$

which is bounded by (3.5) (as (3.5) implies (3.7)). Using (A.4) with $\boldsymbol{M}=\boldsymbol{\Sigma}_{t;T}$, the boundedness of $\|\boldsymbol{\Sigma}_{t;T}^{-1}\|$ is shown in exactly the same way. □

Proof of Proposition 3. The proof uses Lemmas 1 to 3 and is along the lines of Dahlhaus (1996b, Theorem 3.2(i)). The idea is to reduce the problem to a stationary situation by fixing the local time at ν_p. Then, the key point is to use the following relation between the wavelet spectrum of a *stationary* process and its classical Fourier spectrum. If X_t is a stationary process with an absolutely summable autocovariance and with Fourier spectrum $f(\cdot)$, then its wavelet spectrum is given by

$$S_j = \sum_\ell A_{j\ell}^{-1} \int d\lambda \; f(\lambda) |\hat{\psi}_\ell(\lambda)|^2 \tag{A.5}$$

for any fixed non-decimated system of compactly supported wavelets $\{\psi_{jk}\}$. We refer to Dahlhaus (1996b, Theorem 3.2(i)) for details. □

We will now study the approximation of $\boldsymbol{\Sigma}_{t;T}$ by $\boldsymbol{B}_{t;T}$.

Lemma 4. *Under the assumptions of Proposition 2 and 4,*

$$\mathrm{MSPE}(\hat{X}_{t+h-1;T}, X_{t+h-1;T}) = \boldsymbol{b}'_{t+h-1} \boldsymbol{B}_{t+h-1;T} \boldsymbol{b}_{t+h-1} + \boldsymbol{b}'_{t+h-1} \boldsymbol{b}_{t+h-1} \; o_T(1)$$

and, in particular,

$$\mathrm{MSPE}(\hat{X}_{t;T}, X_{t;T}) = \boldsymbol{b}'_{t} \boldsymbol{B}_{t;T} \boldsymbol{b}_{t} + \boldsymbol{b}'_{t} \boldsymbol{b}_{t} \; o_T(1)$$

Proof. By the definition of an LSW process, we have $|w_{jk;T}|^2 = S_j((n+m)/T) + (C_j + L_j|k - n - m|)O(T^{-1})$. Therefore,

$$\begin{aligned} \boldsymbol{b}'_{t+h-1} \boldsymbol{\Sigma}_{t+h-1;T} \boldsymbol{b}_{t+h-1} &= \sum_{jk} \sum_{n,m=0}^{t+h-1} b_n b_m \psi_{jk}(n) \psi_{jk}(m) |w_{jk;T}|^2 \\ &= \sum_{jk} \sum_{n,m=0}^{t+h-1} b_n b_m \Psi_j(n-m) S_j\left(\frac{n+m}{2T}\right) + \mathrm{Rest}_1 \end{aligned} \tag{A.6}$$

We bound Rest_1 as follows:

$$|\,\mathrm{Rest}_1\,| \leqslant O(T^{-1}) \sum_{jk} \sum_{n,m=0}^{t+h-1} \left(\left| k - \left(\frac{n+m}{2}\right) \right| L_j + C_j \right) |b_n b_m \psi_{jk}(n) \psi_{jk}(m)|.$$

If $\mathcal{L}_j$ denotes the length of support of ψ_j, we have $0 \leqslant k-n, k-m \leqslant \mathcal{L}_j$ and so $k-(n+m)/2 \leqslant \mathcal{L}_j$ such that

$$\begin{aligned} |\,\mathrm{Rest}_1\,| &\leqslant O(T^{-1}) \sum_{jk} \sum_{n,m=0}^{t+h-1} (\mathcal{L}_j L_j + C_j) \, |b_n b_m \psi_{jk}(n) \psi_{jk}(m)| \\ &\leqslant O(T^{-1}) \boldsymbol{b}'_{t+h-1} \boldsymbol{b}_{t+h-1} \sum_j \mathcal{L}_j \, (\mathcal{L}_j L_j + C_j) = \boldsymbol{b}'_{t+h-1} \boldsymbol{b}_{t+h-1} o_T(1) \qquad \text{by assumption.} \end{aligned}$$

Finally, by Assumption (3.5), (A.6) yields the result. □

Lemma 5. *Under the assumptions of Proposition 4, we have*

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} = \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\,(1+o_T(1))$$

Proof of Lemma 5. By Lemma 4, we have $\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} = \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1} + \boldsymbol{b}'_{t+h-1}\boldsymbol{b}_{t+h-1}\ o_T(1)$ By Lemma 3, the inverse of $\boldsymbol{\Sigma}_{t;T}$ is bounded in T and, by standard properties of the spectral norm, we have

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{b}_{t+h-1} \leqslant \boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\ \|\boldsymbol{\Sigma}^{-1}_{t+h-1;T}\|$$

for all sequences $\boldsymbol{b}_{t+h-1}$. The above gives

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} \leqslant \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1} + \boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\ \|\boldsymbol{\Sigma}^{-1}_{t+h-1;T}\|\ o_T(1)$$

which is equivalent to

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} \leqslant \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\Big(1-\|\boldsymbol{\Sigma}^{-1}_{t+h-1;T}\|\ o_T(1)\Big)^{-1}$$

for large T. On the other hand, we have

$$\boldsymbol{b}'_{t+h-1}\boldsymbol{\Sigma}_{t+h-1;T}\boldsymbol{b}_{t+h-1} \geqslant \boldsymbol{b}'_{t+h-1}\boldsymbol{B}_{t+h-1;T}\boldsymbol{b}_{t+h-1}\Big(1+\|\boldsymbol{\Sigma}^{-1}_{t;T}\|\ o_T(1)\Big)^{-1}$$

which implies the result. □

B Estimation of the local autocovariance function

In this section, we study the properties of the estimator of the local autocovariance. We first show some relevant properties of the autocorrelation function $\Psi_j(\tau)$ and the matrix $\boldsymbol{A}$ defined in (4.4).

Lemma 6. *1. The system $\{\Psi_j(\tau), j=-1,-2,\ldots\}$ is linearly independent.*

2. Denote by $\Psi(\tau)$ the wavelet autocorrelation function of a continuous *wavelet ψ, i.e.*

$$\Psi(\tau) = \int du\psi(u)\psi(u-\tau), \qquad \tau\in\mathbb{Z}.$$

We have

$$\Psi_j(\tau) = \Psi\left(2^j|\tau|\right)$$

for all $j=-1,-2,\ldots$ and $\tau\in\mathbb{Z}$.

The proof of the first result can be found in Nason *et al.* (2000, Theorem 1). For the proof of the second result, see, for example, Berkner and Wells (2002, Lemma 4.2).

Lemma 7. $\sum_{j=-\infty}^{-1} 2^j\ \Psi_j(\tau) = \delta_0(\tau)$.

Proof. Using Lemma 6 and Parseval's formula,

$$\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = \sum_{j=-\infty}^{-1} 2^j \Psi\left(2^j |\tau|\right) = \sum_{j=-\infty}^{-1} \int_{-\infty}^{\infty} d\omega \, |\hat{\psi}(2^{-j}\omega)|^2 \exp(i\omega\tau)$$

$$= \sum_{j=-\infty}^{-1} \int_0^{2\pi} d\omega \, \sum_{k\in\mathbb{Z}} \left|\hat{\psi}\left(2^{-j}(\omega + 2k\pi)\right)\right|^2 \exp(i\omega\tau). \tag{B.1}$$

Denote by $m_0(\xi)$ the trigonometric polynomial which corresponds to the construction of wavelet ψ and its corresponding scaling function ϕ (Daubechies, 1992, Theorem 6.3.6). We may write

$$\sum_{k\in\mathbb{Z}} \left|\widehat{\psi}\left(2^{-j}\left(\omega + 2k\pi\right)\right)\right|^2 = \sum_{k\in\mathbb{Z}} \left|m_0\left(2^{-j-1}\omega + 2^{-j-1}k2\pi + \pi\right)\right|^2 \left|\widehat{\phi}\left(2^{-j-1}\omega + 2^{-j-1}k2\pi\right)\right|^2$$

and, using the $2\pi k$-periodicity of m_0,

$$= \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \sum_{k\in\mathbb{Z}} \left|\widehat{\phi}\left(2^{-j-1}\omega + 2^{-j-1}k2\pi\right)\right|^2$$

$$= \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \sum_{k\in\mathbb{Z}} \left|m_0\left(2^{-j-2}\omega + 2^{-j-2}k2\pi\right)\right|^2 \left|\widehat{\phi}\left(2^{-j-2}\omega + 2^{-j-2}k2\pi\right)\right|^2$$

$$= \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \left|m_0\left(2^{-j-2}\omega\right)\right|^2 \sum_{k\in\mathbb{Z}} \left|\widehat{\phi}\left(2^{-j-2}\omega + 2^{-j-2}k2\pi\right)\right|^2.$$

By similar transformations, we finally arrive at

$$= \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \prod_{n=2}^{-j} \left|m_0\left(2^{-j-n}\omega\right)\right|^2 \sum_{k\in\mathbb{Z}} \left|\widehat{\phi}\left(\omega + k2\pi\right)\right|^2$$

$$= (2\pi)^{-1} \left|m_0\left(2^{-j-1}\omega + \pi\right)\right|^2 \prod_{n=2}^{-j} \left|m_0\left(2^{-j-n}\omega\right)\right|^2$$

$$= (2\pi)^{-1} \left|1 - m_0\left(2^{-j-1}\omega\right)\right|^2 \prod_{\ell=0}^{-j-2} \left|m_0\left(2^{\ell}\omega\right)\right|^2.$$

Using (B.1), we obtain

$$\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = (2\pi)^{-1} \int_0^{2\pi} \sum_{j=-\infty}^{-1} d\omega \, \exp(i\tau\omega) \left|1 - m_0\left(2^{-j-1}\omega\right)\right|^2 \prod_{\ell=0}^{-j-2} \left|m_0\left(2^{\ell}\omega\right)\right|^2.$$

Expanding the telescopic sum over j, we get

$$\sum_{j=-\infty}^{-1} \left|1 - m_0\left(2^{-j-1}\omega\right)\right|^2 \prod_{l=0}^{-j-2} \left|m_0\left(2^l\omega\right)\right|^2 = 1 - \lim_{j\to-\infty} \prod_{l=0}^{-j-1} \left|m_0\left(2^l\omega\right)\right|^2$$
$$= 1 - \prod_{l=0}^{+\infty} \left|m_0\left(2^l\omega\right)\right|^2.$$

Thus, we obtain

$$\sum_{j=-\infty}^{-1} 2^j \Psi_j(\tau) = \frac{1}{2\pi}\int_0^{2\pi} d\omega\ \exp\left(i\tau\omega\right)\left\{1 - \prod_{\ell=0}^{+\infty}\left|m_0(2^\ell\omega)\right|^2\right\}$$
$$= \delta_0(\tau) - \frac{1}{2\pi}\int_0^{2\pi} d\omega\ \exp\left(i\tau\omega\right)\prod_{\ell=0}^{+\infty}\left|m_0(2^\ell\omega)\right|^2. \tag{B.2}$$

Now, it remains to prove that the second term in (B.2) is equal to zero. By definition, $m_0(\omega) = 2^{-1/2}\sum_{n=0}^{2N-1} h_n e^{-in\omega}$, where $\{h_k\}_{k\in\mathbb{Z}}$ is the low pass quadrature mirror filter used in the construction of Daubechies' compactly supported continuous time wavelet ψ (Daubechies, 1992, Section 6.4). We have

$$\frac{1}{2\pi}\int_0^{2\pi} d\omega\ \exp\left(i\tau\omega\right)\prod_{\ell=0}^{L}\left|m_0(2^\ell\omega)\right|^2 = \prod_{\ell=0}^{L} 2^{-\ell} \sum_{n,m=0}^{2N-1} h_n\overline{h_m}\delta_0(n-m)$$

which clearly tends to 0 as L tends to infinity. □

Lemma 8. *Matrix $\boldsymbol{A}$ defined in (4.4) has the following properties:*

$$\sum_{j=-\infty}^{-1} 2^j A_{j\ell} = 1. \tag{B.3}$$

If, in addition, $\boldsymbol{A}$ is constructed using Haar wavelets, then

$$\sum_{\ell=-\infty}^{-1} |A^{-1}_{j\ell}| \leqslant C\cdot 2^{j/2} \tag{B.4}$$

$$\sum_{\ell=-\infty}^{-1} A^{-1}_{j\ell} = 2^j \tag{B.5}$$

for all $j < 0$, where C is a constant.

Proof. (B.3) is a straightforward corollary of Lemma 7. To prove (B.4), we introduce the auxiliary matrix $\boldsymbol{\Gamma} = \boldsymbol{D}'\boldsymbol{A}\boldsymbol{D}$, where $\boldsymbol{D} = \mathrm{diag}(2^{j/2})_{j<0}$ is diagonal, i.e. $\Gamma_{j\ell} = 2^{j/2}A_{j\ell}2^{\ell/2}$. Nason *et al.* (2000, Theorem 2) show that the spectral norm of $\boldsymbol{\Gamma}^{-1}$ is bounded for Haar wavelets. Therefore, we obtain (B.4) as $\sum_{\ell=-\infty}^{-1}|A^{-1}_{j\ell}| = \sum_{\ell=-\infty}^{-1} 2^{j/2}2^{\ell/2}|\Gamma^{-1}_{j\ell}| \leqslant C\cdot 2^{j/2}$. To prove (B.5), observe that if $X_{t,T}$ is a white noise, then its classical Fourier spectrum is $f(\lambda) = (2\pi)^{-1}$. On the other hand, white noise is an LSW process such that $\sum_j S_j\Psi_j(\tau) = \delta_0(\tau)$

which implies that $S_j = 2^j$ (Lemma 7). (B.5) then follows from the following property: If X_t is the wavelet spectrum of a *stationary* process with absolute summable autocovariance and with Fourier spectrum f, then its wavelet spectrum is given by $S_j = \sum_\ell A_{j\ell}^{-1} \int d\lambda f(\lambda)|\psi_\ell(\lambda)|^2$ and, moreover, $\int d\lambda |\hat{\psi}_\ell(\lambda)|^2 = 2\pi$. □

Proof of Proposition 5. We will first show
(B.6)

$$\operatorname{cov}\left(\sum_s X_{s,T}\psi_{i,k}(s), \sum_s X_{s,T}\psi_{j,k}(s)\right) = \sum_\tau c(k/T,\tau)\sum_n \psi_{i,n}(\tau)\psi_{j,n}(0) + O(2^{-(i+j)/2}T^{-1}).$$

We have

$$\operatorname{cov}\left(\sum_s X_{s,T}\psi_{i,k}(s), \sum_s X_{s,T}\psi_{j,k}(s)\right) = \sum_{l,u}\left(S_l\left(\frac{k}{T}\right) + O\left(\frac{C_l + L_l(u-k)}{T}\right)\right)\sum_{s,t}\psi_{l,s}(u)\psi_{j,k}(s)\psi_{l,t}(u)\psi_{i,k}(t).$$

Using $\mathcal{L}_j = O(M2^{-j})$ in the first step, and the Cauchy inequality in the second one, we bound the reminder as follows:

$$\begin{aligned}
&\left|\sum_{l,u} O\left(\frac{C_l + L_l(u-k)}{T}\right)\sum_{s,t}\psi_{l,s}(u)\psi_{j,k}(s)\psi_{l,t}(u)\psi_{i,k}(t)\right| \leq \\
&\sum_l \frac{C_l + ML_l(2^{-l} + \min(2^{-i}, 2^{-j}))}{T}\sum_u\left|\sum_{s,t}\psi_{l,s}(u)\psi_{j,k}(s)\psi_{l,t}(u)\psi_{i,k}(t)\right| \leq \\
&\sum_l \frac{C_l + ML_l(2^{-l} + 2^{-i/2}2^{-j/2})}{T}(A_{lj})^{1/2}(A_{li})^{1/2} = \\
&\frac{2^{-(i+j)/2}}{T}\left\{\sum_l (C_l + ML_l 2^{-l})2^{(i+j)/2}(A_{lj})^{1/2}(A_{li})^{1/2} + \sum_l ML_l(A_{lj})^{1/2}(A_{li})^{1/2}\right\} = \\
&\frac{2^{-(i+j)/2}}{T}\{I + II\}.
\end{aligned}$$

By formula (B.3),

$$I \leq \sum_l (C_l + ML_l 2^{-l})(2^i A_{li} + 2^j A_{lj}) \leq \sum_l (C_l + ML_l 2^{-l})2\sum_i 2^i A_{li} \leq D_1.$$

As $\sum_i L_i 2^{-i} < \infty$, we must have $L_i \leq C2^i$ so $\sum_i L_i A_{ij} \leq C$ again by (B.3). This and the Cauchy inequality give

$$II \leq 2M\left(\sum_l L_l A_{li}\right)^{1/2}\left(\sum_l L_l A_{lj}\right)^{1/2} \leq D_2.$$

The bound for the reminder is therefore $O(2^{-(i+j)/2}T^{-1})$. For the main term, straightforward

computation gives

$$\sum_{l,u} S_l\left(\frac{k}{T}\right)\sum_{s,t}\psi_{l,s}(u)\psi_{j,k}(s)\psi_{l,t}(u)\psi_{i,k}(t) = \sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{i,n}(\tau)\psi_{j,n}(0),$$

which yields formula (B.6). Using Lemma 7 and (B.6) with $i = j$, we obtain

$$\begin{aligned}\mathrm{E}(\tilde{c}(k/T,0)) &= \sum_{j=-J}^{-1} 2^j\left\{\sum_{\tau} c(k/T,\tau)\Psi_j\tau + O(2^{-j}/T)\right\}\\ &= \sum_{\tau} c(k/T,\tau)\delta_0(\tau) + O(\log(T)/T) = c(k/T,0) + O(\log(T)/T),\end{aligned}$$

which proves the expectation. For the variance, observe that, using Gaussianity, we have

$$\begin{aligned}\mathrm{cov}\left(I_i\left(\frac{k}{T}\right), I_j\left(\frac{k}{T}\right)\right) &= 2\left(\sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{i,n}(\tau)\psi_{j,n}(0) + O(2^{-(i+j)/2}T^{-1})\right)^2\\ &= 2\left(\sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{i,n}(\tau)\psi_{j,n}(0)\right)^2 + O(2^{-(i+j)/2}T^{-1}),\end{aligned} \tag{B.7}$$

provided that (3.5) holds. Using (B.7), we finally obtain

$$\mathrm{Var}(\tilde{c}(k/T,0)) = 2\sum_{i,j=-J}^{-1} 2^{i+j}\left(\sum_{\tau} c(k/T,\tau)\sum_{n}\psi_{i,n}(\tau)\psi_{j,n}(0)\right)^2 + O(T^{-1}). \tag{B.8}$$

□

References

Antoniadis, A. and Sapatinas, T. (2002). Wavelet methods for continuous-time prediction using representations of autoregressive processes in Hilbert spaces. *J. Multivariate Anal.* (Under revision)

Berkner, K. and Wells, R. (2002). Smoothness estimates for soft-threshold denoising via translation-invariant wavelet transforms. *Appl. Comput. Harmon. Anal.*, *12*, 1–24.

Brockwell, P. J. and Davis, R. A. (1991). *Time series: Theory and methods* (Second ed.). Springer, New York.

Calvet, L. and Fisher, A. (2001). Forecasting multifractal volatility. *J. Econometrics*, *105*, 27–58.

Coifman, R. and Donoho, D. (1995). Time-invariant de-noising. In A. Antoniadis and G. Oppenheim (Eds.), *Wavelets and Statistics* (Vol. 103, pp. 125–150). New York: Springer-Verlag.

Dahlhaus, R. (1996a). Asymptotic statistical inference for nonstationary processes with evolutionary spectra. In P. Robinson and M. Rosenblatt (Eds.), *Athens conference on applied probability and time series analysis* (Vol. 2). Springer, New York.

Dahlhaus, R. (1996b). On the Kullback-Leibler information divergence of locally stationary processes. *Stochastic Process. Appl.*, *62*, 139–168.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann. Statist.*, *25*, 1–37.

Dahlhaus, R., Neumann, M. H. and von Sachs, R. (1999). Non-linear wavelet estimation of time-varying autoregressive processes. *Bernoulli*, *5*, 873–906.

Daubechies, I. (1992). *Ten lectures on wavelets.* Philadelphia: SIAM.

Fryźlewicz, P. (2002). *Modelling and forecasting financial log-returns as locally stationary wavelet processes* (Research Report). Department of Mathematics, University of Bristol. (`http://www.stats.bris.ac.uk/pub/ResRept/2002.html`)

Grillenzoni, C. (2000). Time-varying parameters prediction. *Ann. Inst. Statist. Math.*, *52*, 108–122.

Kress, R. (1991). *Numerical analysis.* New York: Springer.

Ledolter, J. (1980). Recursive estimation and adaptive forecasting in ARIMA models with time varying coefficients. In *Applied Time Series Analysis, II (Tulsa, Okla.)* (pp. 449–471). New York-London: Academic Press.

Mallat, S., Papanicolaou, G. and Zhang, Z. (1998). Adaptive covariance estimation of locally stationary processes. *Ann. Statist.*, *26*, 1–47.

Mélard, G. and Herteleer-De Schutter, A. (1989). Contributions to the evolutionary spectral theory. *J. Time Ser. Anal.*, *10*, 41–63.

Nason, G. P. and von Sachs, R. (1999). Wavelets in time series analysis. *Phil. Trans. Roy. Soc. Lond.* A, *357*, 2511–2526.

Nason, G. P., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of evolutionary wavelet spectra. *J. Roy. Statist. Soc. Ser.* B, *62*, 271–292.

Ombao, H., Raz, J., von Sachs, R. and Guo, W. (2002). The SLEX model of a non-stationary random process. *Ann. Inst. Statist. Math.*, *54*, 171–200.

Ombao, H., Raz, J., von Sachs, R. and Malow, B. (2001). Automatic statistical analysis of bivariate nonstationary time series. *J. Amer. Statist. Assoc.*, *96*, 543–560.

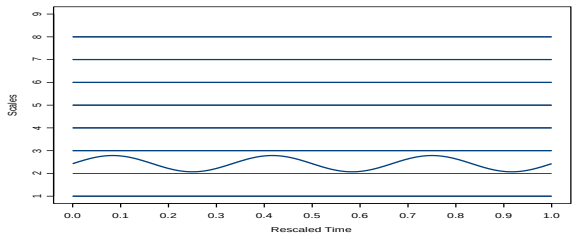
Philander, S. (1990). *El Niño, La Niña and the southern oscillation.* San Diego: Academic Press.

Priestley, M. (1965). Evolutionary spectra and non-stationary processes. *J. Roy. Statist. Soc. Ser.* B, *27*, 204–237.

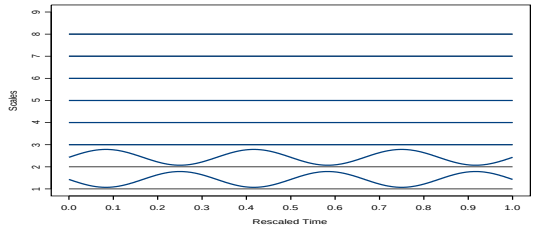
Van Bellegem, S. and von Sachs, R. (2002). *Forecasting economic time series using models of nonstationarity* (Discussion paper No. 0227). Institut de statistique, UCL. (`ftp://www.stat.ucl.ac.be/pub/papers/dp/dp02/dp0227.ps`)

List of Figures

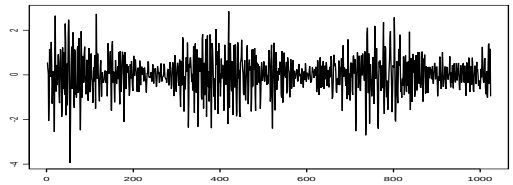
Figures



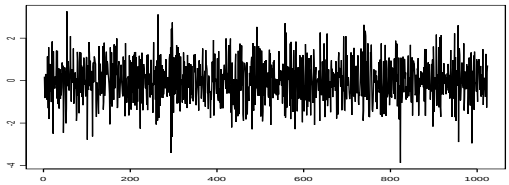
(a) Theoretical wavelet spectrum equal to zero everywhere except scale -2 where $S_{-2}(z) = 0.1 + \cos^2(3\pi z + 0.25\pi)$.



(b) Theoretical wavelet spectrum $S_{-2}(z) = 0.1 + \cos^2(3\pi z + 0.25\pi)$, $S_{-1}(z) = 0.1 + \sin^2(3\pi z + 0.25\pi)$ and $S_j(z) = 0$ for $j \neq -1, -2$.

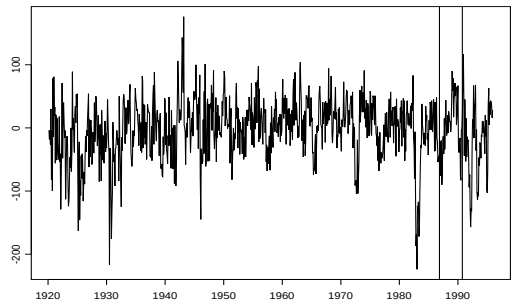


(c) A sample path of length 1024 simulated from the wavelet spectrum defined in (a).

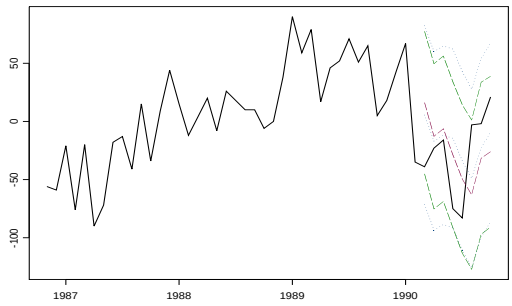


(d) A sample path of length 1024 simulated from the wavelet spectrum defined in (b).

Figure 1: These simulated examples demonstrate the idea of a sparse representation of the local (co)variance. The left-hand column shows an example of a smooth time-varying variance function of a TM process. The example on the right hand side is constructed in such a way that the local variance function $c(z, 0)$ is constant over time. In this example, the only deviation from stationarity is in the covariance structure. The simulations, like all throughout the article, use Gaussian innovations ξ_{jk} and Haar wavelets.

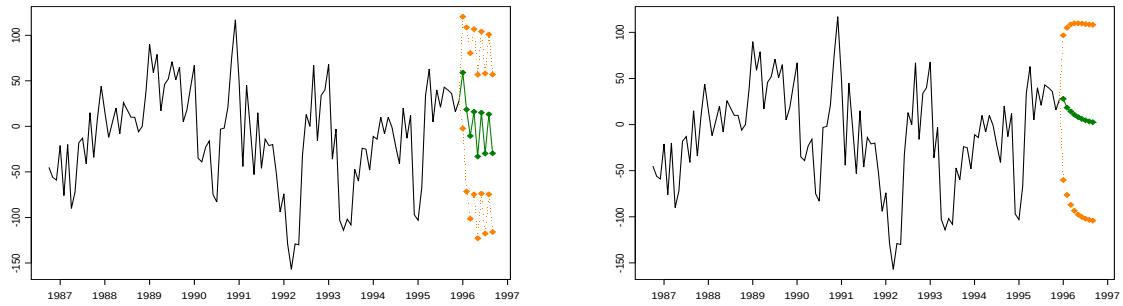


(a) The wind anomaly index (in cm/s). The two vertical lines indicate the segment shown in Figure 2(b).



(b) Comparison between the one-step-ahead prediction in our model (dashed lines) and AR (dotted lines).

Figure 2: The wind anomaly data (910 observations from March 1920 to December 1995).



(a) 9-step-ahead prediction using LSW modelling (b) 9-step-ahead prediction using AR modelling

Figure 3: The last observations of the wind anomaly series and its 1- up to 9-step-ahead forecasts (in cm/s).