J. R. Statist. Soc. B (2006)
68, *Part* 5, *pp.* 837–857

A functional wavelet–kernel approach for time series prediction

Anestis Antoniadis,

Joseph Fourier University, Grenoble, France

and Efstathios Paparoditis and Theofanis Sapatinas

University of Cyprus, Nicosia, Cyprus

[Received April 2005. Final revision June 2006]

Summary. We consider the prediction problem of a time series on a whole time interval in terms of its past. The approach that we adopt is based on functional kernel nonparametric regression estimation techniques where observations are discrete recordings of segments of an underlying stochastic process considered as curves. These curves are assumed to lie within the space of continuous functions, and the discretized time series data set consists of a relatively small, compared with the number of segments, number of measurements made at regular times. We estimate conditional expectations by using appropriate wavelet decompositions of the segmented sample paths. A notion of similarity, based on wavelet decompositions, is used to calibrate the prediction. Asymptotic properties when the number of segments grows to ∞ are investigated under mild conditions, and a nonparametric resampling procedure is used to generate, in a flexible way, valid asymptotic pointwise prediction intervals for the trajectories predicted. We illustrate the usefulness of the proposed functional wavelet–kernel methodology in finite sample situations by means of a simulated example and two real life data sets, and we compare the resulting predictions with those obtained by three other methods in the literature, in particular with a smoothing spline method, with an exponential smoothing procedure and with a seasonal autoregressive integrated moving average model.

Keywords: α-mixing; Besov spaces; Exponential smoothing; Functional kernel regression; Pointwise prediction intervals; Resampling; Seasonal autoregressive integrated moving average models; Smoothing splines; Time series prediction; Wavelets

1. Introduction

In many real life situations we seek information on the evolution of a (real-valued) continuous time stochastic process $X = (X(t); t \in \mathbb{R})$ in the future. Given a trajectory of X observed on the interval $[0, T]$, we would like to predict the behaviour of X on the entire interval $[T, T + \delta]$, where $\delta > 0$, rather than at specific time points. An appropriate approach to this problem is to divide the interval $[0, T]$ into subintervals $[l\delta, (l+1)\delta]$, $l = 0, 1, \ldots, k-1$, with $\delta = T/k$, and to consider the stochastic process $Z = (Z_i; i \in \mathbb{N})$, where $\mathbb{N} = \{1, 2, \ldots\}$, defined by

$$Z_i(t) = X\{t + (i-1)\delta\}, \qquad i \in \mathbb{N}, \quad \forall t \in [0, \delta). \tag{1}$$

Note that δ is not a parameter to be included in the modelling formulation. For some specific examples at hand, where some periodicity is obvious in the observed phenomena, the parameter δ is directly tied to the period. It allows us, in a natural way, to render the discrete

Address for correspondence: Anestis Antoniadis, Laboratoire de Modelisation et Calcul, Université Joseph Fourier, BP 53, 38041 Grenoble Cedex 9, France.
E-mail: Anestis.Antoniadis@imag.fr

 1369–7412/06/68837

(function-valued) time series Z strictly stationary (as a sequence), without imposing any particular stationarity assumption on the behaviour of the time series within a segment. However, δ does not need to be a period. Consider, for example, television audience series recordings. A natural choice is to separate the audience recordings by specific days of the week, given the strong differential of television audience with respect to the day of the week. Looking at the recordings for a specific day of the week allows us in a natural way to model the resulting sequence of days as a (function-valued) strictly stationary time series, with some dependence due to the television habits of the audience. Whatever the case is, it is clear that the choice of δ is suited for data that can be appropriately segmented; this is the approach that is adopted in this paper.

In the recent literature, practically all investigations to date for this prediction problem are for the case where one assumes that (an appropriately centred version of) the stochastic process Z is a (zero-mean) Hilbert-valued *autoregressive (of order 1) process* (ARH(1)); the best prediction of Z_{n+1} given its past history $(Z_n, Z_{n-1}, \ldots, Z_1)$ is then given by

$$\begin{aligned}\tilde{Z}_{n+1} &= \mathbb{E}(Z_{n+1}|Z_n, Z_{n-1}, \ldots, Z_1)\\ &= \rho(Z_n), \qquad\qquad n \in \mathbb{N},\end{aligned}$$

where ρ is a bounded linear operator that is associated with the ARH(1) process. The approaches adopted mainly differ in the way of estimating the 'prediction' operator ρ, or its value $\rho(Z_n)$ given $Z_1, Z_2, \ldots, Z_n$ (see, for example, Bosq (1991), Besse and Cardot (1996), Pumo (1998) and Antoniadis and Sapatinas (2003)).

In many practical situations, however, the discrete time stochastic process $Z = (Z_i; i \in \mathbb{N})$ may not be modelled with such an autoregressive structure. This is the case that we consider in the following development. In particular, we assume that the (real-valued) continuous time stochastic process $X = (X(t); t \in \mathbb{R})$ has a representation of the form (1) with 'blocks' Z_i, for $i \in \mathbb{N}$, that are observed on a discrete sampling grid of fixed size. We then develop a version of prediction via functional kernel nonparametric regression estimation techniques, in which both the predictor and the response variables are discretized functions of time, using a conditioning idea. Under mild assumptions on the observed time series, prediction of the block Z_{n+1} is obtained by kernel regression of the present block Z_n on the past blocks $\{Z_{n-1}, Z_{n-2}, \ldots, Z_1\}$. The resulting predictor will be seen as a weighted average of the past blocks, placing more weight on those blocks the preceding of which is similar to the present one. Hence, the analysis is rooted in the ability to find 'similar blocks'. Considering that blocks can be sampled values of quite irregular curves, similarity matching is based on a distance metric on the wavelet coefficients of a suitable wavelet decomposition of the blocks. A resampling scheme, involving resampling of the original blocks to form 'pseudoblocks' of the same duration, is then used to calculate pointwise prediction intervals for the predicted block.

Unlike traditional forecasting methods for discrete time series (e.g. seasonal autoregressive integrated moving average (SARIMA) models, vector autoregressive models or exponential smoothing), the forecasting methodology suggested gives some protection against neglecting essential characteristics of the stochastic process that could prove useful for prediction. In particular, it avoids the need to treat a possibly complicated dependence structure within a segment by a multivariate forecasting mechanism and allows us to model phenomena naturally with either slow variation (e.g. climatic cycles) or high frequency phenomena (e.g. audit television series) where recordings are made each second and for which stationarity in the classical discrete sense does not hold.

The paper is organized as follows. In Section 2, we first introduce some relevant notation and then discuss the extension of the conditioning approach to the one time interval ahead prediction. Resampling-based pointwise prediction intervals are also derived. In Section 3, we

illustrate the usefulness of the proposed functional wavelet–kernel approach for time series prediction by means of a simulated example and two real life data sets. We also compare the resulting predictions with those obtained by three other methods in the literature, in particular with a smoothing spline method, with an exponential smoothing procedure, and with a SARIMA model. Auxiliary results and proofs are compiled in Appendix A.

2. Functional wavelet–kernel prediction

2.1. Notation

Let $X = (X(t); t \in \mathbb{R})$ be a (real-valued) continuous time stochastic process that is defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Motivated by applications to prediction and forecasting, it is supposed that the time domain of X is divided into intervals of constant width $\delta > 0$. Therefore, from X, a sequence of (function-valued) random variables $(Z_i;\ i \in \mathbb{N})$ is constructed according to representation (1), i.e.

$$Z_i(t) = X\{t + (i-1)\delta\}, \qquad i \in \mathbb{N}, \quad \forall t \in [0, \delta).$$

This approach has become popular in the statistical community because of its ability to aid understanding of the whole evolution of the stochastic process X.

Recall that our aim is a one time interval ahead prediction, i.e. a one-step-ahead prediction for the discrete (function-valued) time series $Z = (Z_i;\ i \in \mathbb{N})$. In what follows, we assume that Z is strictly stationary (see Bosq (2000), chapter 1) with $\mathbb{E}(\|Z_i\|) < \infty$, where $\|\cdot\|$ denotes the (semi-) norm of the corresponding functional space. If the time series Z is not stationary, it is assumed that it has been transformed to a strictly stationary time series by a preprocessing procedure. Using a standard wavelet approach, the random curves Z_i are then expanded into a wavelet basis. We may have used a (fixed) spline or Fourier basis instead but there are some good reasons to prefer wavelet bases. A spline expansion could make sense if the sample paths exhibit a uniformly smooth temporal structure, the same being true for a Fourier basis. In contrast, a wavelet decomposition of the sample paths is local, so if the information that is relevant to our prediction problem is contained in a particular part (or parts) of the sample path Z_i, as is typical in many practical applications, this information will be contained in a small number of wavelet coefficients.

Since in the subsequent development we are dealing with wavelet decompositions, for each $i \in \mathbb{N}$, denote by $\Xi_i = \{\xi_i^{(J,k)} :\ k = 0, 1, \ldots, 2^J - 1\}$ the set of scaling coefficients at scale J of the ith segment Z_i. Because $Z = (Z_i;\ i \in \mathbb{N})$ is a strictly stationary stochastic process, the same is also true for the 2^J-dimensional stochastic process $(\Xi_i;\ i \in \mathbb{N})$. Moreover, if the strict stationarity assumption is too strong, we could calibrate the non-stationarity by considering only J-stationarity, i.e. strict stationarity of the scaling coefficients up to (the finest) scale J, with a possibly different distribution at each scale $j \leqslant J$ (see Cheng and Tong (1998)).

In practice, the random curves Z_i are observed only at discretized equidistant time values in $[0, \delta)$, say $t_1, \ldots, t_P$, with $P = 2^J$ for some fixed positive integer J. For J sufficiently large, and for a sufficiently regular scaling function, we have $\xi_i^{(J,k)} \simeq 2^{-J/2} Z_i(t_{k+1})$, for all $k = 0, 1, \ldots, 2^J - 1$; hence, the above facts still hold for the set of discrete scaling coefficients.

2.2. Finite dimensional kernel prediction

Consider the nonparametric prediction of a (real-valued) stationary discrete time stochastic process. Let $Y_{n,(r)} = (Y_n, Y_{n-1}, \ldots, Y_{n-r+1})' \in \mathbb{R}^r$ be the vector of lagged variables, and let s be the forecast horizon. It is well known that the autoregression function plays a predominant forecasting role in the above time series context. Recall that the autoregression function f is defined by

$$f(\mathbf{y}) = \mathbb{E}(Y_{n+s} | Y_{n,(r)} = \mathbf{y}), \qquad \mathbf{y} \in \mathbb{R}^r.$$

It is clear that the first task is to estimate f. The classical approach to this problem is to find some parametric estimate of f. More specifically, it is assumed that f belongs to a class of functions, only depending on a finite number of parameters to be estimated. This is the case of several well-known parametric models, which have been widely studied in the literature (see, for example, Box and Jenkins (1976) and Brockwell and Davis (1991)).

The above prediction problem can also be undertaken with a nonparametric view, without any assumption on the functional form of f. This is a much more flexible approach that only imposes regularity conditions on f. Nonparametric methods for forecasting time series can be viewed, up to a certain extent, as a particular case of nonparametric regression estimation under dependence (see, for example, Bosq (1991), Härdle and Vieu (1992) and Hart (1996)). A popular nonparametric method for such a task is to use the kernel smoothing ideas because they have good properties in (real-valued) regression problems, from both a theoretical and a computational point of view. The kernel estimator $\hat{f}_n$ (based on $Y_1, \ldots, Y_n$) of f is defined by

$$\hat{f}_n(\mathbf{y}) = \frac{\sum_{t=r}^{n-s} \mathbb{K}\{(\mathbf{y} - Y_{t,(r)})/h_n\} Y_{t+s}}{\sum_{m=r}^{n-s} \mathbb{K}\{(\mathbf{y} - Y_{m,(r)})/h_n\}},$$

or 0 if the denominator is zero. In our development, for simplicity, we consider a product kernel, i.e., for each $\mathbf{y} = (y_1, \ldots, y_r)'$,

$$\mathbb{K}(\mathbf{y}) = \prod_{i=1}^{r} K(y_i);$$

also h_n is a sequence of positive numbers (the bandwidths). The s-ahead prediction is then simply given by $Y_{n+s|n} = \hat{f}_n(Y_{n,(r)})$. Theoretical results show that the detailed choice of the kernel function does not influence strongly the behaviour of the prediction but the choice of the bandwidth values is crucial for the accuracy of prediction (see, for example, Bosq (1998)).

As is readily seen, the prediction is expressed as a locally weighted average of past values, where the weights measure the similarity between $(Y_{t,(r)};\ t = r, \ldots, n-s)$ and $Y_{n,(r)}$, taking into account the process history. Let now $\|\cdot\|$ be a generic notation for a Euclidean norm. If the kernel values decrease to 0 as $\|\mathbf{y}\|$ increases, the smoothing weights have high values when the $(Y_{t,(r)})$ is close to $Y_{n,(r)}$ and are close to 0 otherwise. In other words, the prediction $Y_{n+s|n}$ is obtained as a locally weighted average of blocks of horizon s in all blocks of length r in the past, weighted by similarity coefficients $w_{n,t}$ of these blocks with the current block,

$$\hat{f}_n(Y_{n,(r)}) = \sum_{t=r}^{n-s} w_{n,t}(Y_{n,(r)}) Y_{t+s},$$

where

$$w_{n,t}(\mathbf{y}) = \frac{\mathbb{K}\{(\mathbf{y} - Y_{t,(r)})/h_n\}}{\sum_{m=r}^{n-s} \mathbb{K}\{(\mathbf{y} - Y_{m,(r)})/h_n\}}.$$

2.3. Functional wavelet–kernel prediction

Recall that, in our setting, the strictly stationary time series $Z = (Z_i;\ i \in \mathbb{N})$ is function valued rather than $\mathbb{R}$ valued, i.e. each Z_i is a random curve. In this functional set-up, and to simplify the notation, we address, without loss of generality, the prediction problem for a horizon $s = 1$. We could mimic the above kernel regression ideas and use the estimate

$$Z_{n+1|n}(\cdot) = \sum_{m=1}^{n-1} w_{n,m}\, Z_{m+1}(\cdot), \tag{2}$$

where the triangular array of local weights $\{w_{n,m} : m = 1, 2, \ldots, n-1; n \in \mathbb{N}\}$ increases with the closeness or similarity of the last observed sample path Z_n and the sample paths Z_m in the past, in a (semi)norm sense; this is made more precise in equation (3) later.

The literature on this infinite dimension kernel regression related topic is relatively limited, to our knowledge. Bosq and Delecroix (1985) dealt with general kernel predictors for Hilbert-valued stationary Markovian processes. A similar idea was applied by Besse *et al.* (2000) for ARH(1) processes in the special case of a Sobolev space. Extending and justifying these kernel regression techniques to infinite dimensional stochastic processes with no specific structures (e.g. ARH(1) or more general Markovian processes) will require using measure theoretic assumptions on infinite dimensional spaces (e.g. a probability density function with respect to an invariant measure). This, obviously, restricts the analysis and the applicability of the resulting predictor to a small class of stochastic processes such as diffusion processes, where it is well known that, under some assumptions, the measure corresponding to the probability distribution of the diffusion process has a probability density function with respect to Wiener measure (see, for example, Lipster and Shiryayev (1977)).

Such a kind of assumptions is more natural in finite dimensional spaces such as those which are obtained through orthonormal wavelet decompositions, especially when the discretized sample paths of the observed process are quite coarse. Taking advantage of these remarks, the suggested forecasting methodology relies on a wavelet decomposition of the observed curves and uses the concepts of strict stationarity and α-mixing that are briefly discussed in Appendix A. Moreover, distributional assumptions on the scaling coefficients such as those given in Appendix A are less restrictive than using similar assumptions on the original time series Z.

To summarize, the forecasting methodology proposed is decomposed into two phases:

(a) find within the past sample paths those that are 'similar' to the last observed sample path (this determines the weights);
(b) use the weights and the stochastic multiresolution analysis to forecast by a locally weighted averaging process such as that described by equation (2).

Since we are dealing with a wavelet decomposition, it is worth isolating the first phase by discussing possible ways to measure the similarity of two curves, by means of their wavelet approximation, and then to proceed to the second phase, using again this wavelet approximation. The analysis of the functional wavelet–kernel prediction method proposed is based on finding similar sample paths. Similarity is now defined in terms of a distance metric that is related to the functional space in which the sample paths lie. When the space is a Besov space, it is well known that its norm is characterized by a weighted l_p-norm of the wavelet coefficients of its elements (see, for example, Meyer (1992)). It is therefore natural to address the similarity issue on the wavelet decomposition of the observed sample paths. The wavelet transform is applied to the observed sample paths, and owing to the approximation properties of the wavelet transform only a few coefficients of the transformed data will be used; a kind of contractive property of the wavelet transform.

Applying the discrete wavelet transform to each sample path decomposes the temporal information of the time series into discrete wavelet coefficients that are associated with both time and scale. Discarding scales in the discrete wavelet transform that are associated with high frequency oscillations provides a straightforward data reduction step and decreases the computational burden. We want to use the distributional properties of the wavelet coefficients of the observed series. Imagine first that we are given two observed series, and let $\theta_{j,k}^{(i)}$, $i = 1, 2$, be the discrete wavelet coefficient of the discrete wavelet transform of each signal at scale j ($j = j_0, \ldots, J-1$) and location k ($k = 0, 1, \ldots, 2^j - 1$). At each scale $j \geqslant j_0$, define a measure of discrepancy in terms

of a distance

$$d_j(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) = \left\{ \sum_{k=0}^{2^j-1} (\theta_{jk}^{(1)} - \theta_{j,k}^{(2)})^2 \right\}^{1/2},$$

which measures how effectively the two signals match at scale j. To combine all scales, we then use

$$D(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}) = \sum_{j=j_0}^{J-1} 2^{-j/2} d_j(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}).$$

Remark 1.

(a) Since we have assumed that the time series $Z = (Z_i;\ i \in \mathbb{N})$ is strictly stationary, the scaling coefficients below the scale j_0 do not have any discriminative power; hence we use only discrete wavelet coefficients after j_0 in defining the distance $D(\cdot, \cdot)$ above.

(b) An intuition behind the distance $D(\cdot, \cdot)$ originates from the fact that successive scales (from the finest to coarser scales) consist of only half as many discrete wavelet coefficients as the previous scale. Since each scale-based distance $d_j(\cdot, \cdot)$ is a sum of $n_j = 2^j$ terms, and the n_j are halved as j decreases, the relative magnitude of the scale-based distances $d_j(\cdot, \cdot)$ $(j = J-1, \ldots, j_0)$ varies greatly. This complicates a direct comparison of different scale-based distances $d_j(\cdot, \cdot)$, so a unit weight vector would be less than ideal. The weights that we propose adjust for the differences in magnitude by giving each successive scale-based distance $d_j(\cdot, \cdot)$ twice as much weight as the previous finer scale. This weighting scheme puts all the scale-based distances $d_j(\cdot, \cdot)$ on the same calibre and places a greater emphasis on the discrete wavelet coefficients corresponding to the coarser scale where a stationary signal is best represented.

As for the second phase, recall that, for each $i \in \mathbb{N}$, $\Xi_i = \{\xi_i^{(J,k)} :\ k = 0, 1, \ldots, 2^J - 1\}$ denotes the set of scaling coefficients at scale J of the ith segment Z_i. The kernel prediction of the scaling coefficients at time $n+1$, $\Xi_{n+1|n}$, is given by

$$\Xi_{n+1|n} = \frac{\sum_{m=1}^{n-1} K[D\{\mathcal{C}(\Xi_n), \mathcal{C}(\Xi_m)\}/h_n]\Xi_{m+1}}{1/n + \sum_{m=1}^{n-1} K[D\{\mathcal{C}(\Xi_n), \mathcal{C}(\Xi_m)\}/h_n]}, \tag{3}$$

where the factor $1/n$ in the denominator allows expression (3) to be properly defined and does not affect its rate of convergence. Here, for simplicity, we use the notation $D(x, y)/h_n = D(x/h_n, y/h_n)$, and $\mathcal{C}(\Xi_k)$ is the set of wavelet coefficients that are obtained by applying the 'pyramid algorithm' (see Mallat (1989)) on the set of (finest level) scaling coefficients Ξ_i, for $i = 1, 2, \ldots, n$. This leads to the time domain prediction at time $n+1$,

$$Z_{n+1|n}^J(t) = \sum_{k=0}^{2^J-1} \xi_{n+1|n}^{(J,k)} \phi_{J,k}(t), \qquad \forall\, t \in [0, \delta), \tag{4}$$

where $\xi_{n+1|n}^{(J,k)}$ are the components of the predicted scaling coefficients $\Xi_{n+1|n}$, and with an analogous expression for $Z_{n+1}^J(t)$. The following theorem shows its consistency property.

Theorem 1. Suppose that assumptions (A1)–(A6), which are given in Appendix A, are true. For every $t \in \{t_1, t_2, \ldots, t_P\}$, $P = 2^J$ and, if $h_n = O[\{\log(n)/n\}^{1/(2+2^J)}]$, then, as $n \to \infty$, we have

$$\sup_{\mathbf{x}\in S} |Z^J_{n+1|n}(t) - \mathbb{E}\{Z^J_{n+1}(t)|\mathbf{Z}_n = \mathbf{x}\}| = O\left[\left\{\frac{\log(n)}{n}\right\}^{1/(2+2^J)}\right], \qquad \text{almost surely,}$$

for any compact set S such that the density f of the vector of scaling coefficients at scale J is such that $\min_S(f) > 0$, where $\mathbf{Z}_n = (Z_n(t_1), \ldots, Z_n(t_P))'$.

Remark 2.

(a) In the assertion of theorem 1, the size $P = 2^J$ of the sampling grid over each segment affects the rate of convergence of the predictor. When an asymptotically non-increasing number P of measurements is available for each portion of the time series, which is the most usual in practice, the rate of convergence becomes slower as the size P of the sampling grid increases (the curse of dimensionality) but we obtain consistency as the number of segments increases to ∞.

(b) By considering a fixed number P of wavelet coefficients, the rates in theorem 1 remain the same, up to some constants, whatever distance we use; this is why the particular form of D does not appear in its proof. However, the distance that we have chosen is particularly suited for measuring the proximity between two segments and seems well adapted to the prediction.

We may now summarize the suggested forecasting algorithm as follows.

(a) Each segment Z_i, $i = 1, \ldots, n$, of the original time series X is sampled on a fixed equidistant sampling grid of size $P = 2^J$, giving a P-dimensional vector $\mathbf{Z}_i = (Z_i(t_1), \ldots, Z_i(t_P))'$, $i = 1, \ldots, n$.

(b) Apply the discrete wavelet transform to each of these $\mathbf{Z}_i$ to obtain a P-dimensional scaling coefficient vector Ξ_i in the scale–location space.

(c) Compute the kernel-predicted scaling coefficients by using equation (3), and use them in equation (4) to obtain the one time interval ahead prediction.

2.4. Resampling-based pointwise prediction intervals

Apart from the prediction $Z^J_{n+1|n}(t)$ that was discussed above, we also construct resampling-based pointwise prediction intervals for $Z_{n+1}(t)$. A pointwise prediction interval for $Z_{n+1}(t)$ is defined to be a set of lower and upper points $L_{n+1,\alpha}(t_i)$ and $U_{n+1,\alpha}(t_i)$ respectively, such that, for every t_i, $i = 1, 2, \ldots, P$, and a given $\alpha \in (0, 1)$,

$$\mathbb{P}\{L_{n+1,\alpha}(t_i) \leqslant Z_{n+1}(t_i) \leqslant U_{n+1,\alpha}(t_i)\} = 1 - 2\alpha.$$

Since we are looking at the one-step prediction of $Z_{n+1}(t)$ given Z_n, we are in fact interested in the conditional distribution of $Z_{n+1}(t)$ given Z_n, i.e. $L_{n+1,\alpha}(t_i)$ and $U_{n+1,\alpha}(t_i)$ are the lower and upper α-percentage points of the conditional distribution of $Z_{n+1}(t_i)$ given Z_n.

To construct such a prediction interval the following simple resampling procedure is proposed. Given Z_n, i.e. given $C(\Xi_n)$, define the weights

$$w_{n,m} = \frac{K[D\{\mathcal{C}(\Xi_n), \mathcal{C}(\Xi_m)\}/h_n]}{n^{-1} + \sum_{m=1}^{n-1} K[D\{\mathcal{C}(\Xi_n), \mathcal{C}(\Xi_m)\}/h_n]} + \frac{(n-1)^{-1}}{1 + n\sum_{m=1}^{n-1} K[D\{\mathcal{C}(\Xi_n), \mathcal{C}(\Xi_m)\}/h_n]}.$$

Note that the weights have been selected appropriately so that

$$0 \leqslant w_{n,m} \leqslant 1 \quad \text{and} \quad \sum_{m=1}^{n-1} w_{n,m} = 1.$$

Now, given Z_n, generate pseudorealizations $Z^*_{n+1}(t)$ such that, for $m = 1, 2, \ldots, n-1$,

$$\mathbb{P}\left\{Z^*_{n+1}(t) = Z_{m+1}(t) | Z_n\right\} = w_{n,m},$$

i.e. $Z^*_{n+1}(t)$ is generated by choosing randomly a segment from the whole set of observed segments $Z_{m+1}(t)$, where the probability that the $(m+1)$th segment is chosen depends on how 'similar' is the preceding segment Z_m to Z_n. This 'similarity' is measured by the resampling probability $w_{n,m}$.

Given pseudoreplicates $Z^*_{n+1}(t)$, calculate $R^*_{n+1}(t_i) = Z^*_{n+1}(t_i) - Z^J_{n+1|n}(t_i)$, where $Z^J_{n+1|n}(t_i)$ is our time domain conditional mean predictor. Let $R^*_{n+1,\alpha}(t_i)$ and $R^*_{n+1,1-\alpha}(t_i)$ be the lower and upper α percentage points of $R^*_{n+1}(t_i)$. Note that these percentage points can be consistently estimated by the corresponding empirical percentage points over B realizations $Z^{*(b)}_{n+1}(t_i)$, $b = 1, 2, \ldots, B$, of $Z^*_{n+1}(t_i)$. A $100(1-2\alpha)\%$ pointwise prediction interval for $Z_{n+1}(t_i)$ is then obtained by

$$\{[L^*_{n+1,\alpha}(t_i),\ U^*_{n+1,\alpha}(t_i)],\ i = 1, 2, \ldots, P\},$$

where $L^*_{n+1,\alpha}(t_i) = R^*_{n+1,\alpha}(t_i) + Z^J_{n+1|n}(t_i)$ and $U^*_{n+1,\alpha}(t_i) = R^*_{n+1,1-\alpha}(t_i) + Z^J_{n+1|n}(t_i)$.

The following theorem shows that the method proposed is asymptotically valid, i.e. the so-constructed resampling-based prediction interval achieves the desired pointwise coverage probability.

Theorem 2. Suppose that assumptions (A1)–(A6), which are given in the Appendix A, are true. Then, for every $i = 1, 2, \ldots, P$ and a given $\alpha \in (0, 1)$, we have

$$\lim_{n\to\infty} [\mathbb{P}\{L^*_{n+1,\alpha}(t_i) \leqslant Z_{n+1}(t_i) \leqslant U^*_{n+1,\alpha}(t_i) | Z_1, \ldots, Z_n\}] = 1 - 2\alpha.$$

We conclude this section by pointing out that, as in any nonparametric smoothing approach, the choice of the smoothing parameter h_n (the bandwidth) is of great importance. Once h_n has been specified, only time segments that lie within a similarity distance from the segment Z_n within h_n will be used to estimate the prediction at time $n+1$. Intuitively, a large value of h_n will lead to an estimator that incurs large bias, whereas a small value might reduce the bias but the variability of the predicted curve could be large since only few segments are used in the estimation. A good choice of h_n should balance the bias–variance trade-off. In our implementation, we have used leave-one-out cross-validation for time series data as suggested by Hart (1996). The principle of the cross-validation criterion is to select the bandwidth which, for our given prediction horizon $s = 1$, minimizes the mean-squared prediction errors of the $(i+1)$th segment using all segments in the past except the ith, i.e. the value of h_n that minimizes

$$\mathrm{CV}(h) = \frac{1}{n-1} \sum_{i=1}^{n-1} \| Z_{i+1} - Z^{(-i)}_{i+1|i} \|^2,$$

where $Z^{(-i)}_{i+1|i}$ is the kernel regression estimate with bandwidth h that is obtained by using the series without its ith segment. This is the method for choosing the bandwidth that is adopted in the numerical results that are presented in Section 3.

3. Applications

We now illustrate the usefulness of the proposed functional wavelet–kernel approach WK for time series prediction in finite sample situations by means of a simulated example and two real life data sets, in particular with

(a) the 1-day-ahead prediction of Paris electrical power consumption from half-hour daily recordings and
(b) the 1-week-ahead prediction of the average total audience television rates per day of the week from daily recordings.

For the wavelet–kernel approach, the *Symmlet 6* wavelet filter (see Daubechies (1992), page 195) was used. Preliminary simulations show that the analysis is robust with respect to the wavelet filter, e.g. using *Coiflet 3* (see Daubechies (1992), page 258). In the case where the number of time points (P) in each segment is not a power of 2, each segment is extended by periodicity at the right to a length that is closest to the nearest power of 2. The Gaussian kernel K was adopted in our analysis. Again, preliminary simulations show that our analysis is robust with respect to kernels with unbounded support (e.g. Laplace). The bandwidth (h_n) was chosen by leave-one-out cross-validation for time series data as suggested by Hart (1996). For the associated 95% resampling-based pointwise prediction intervals, the number of resampling samples (B) was taken equal to 500.

We compare the resulting predictions with those which are obtained by three well-established methods in the literature, in particular with a smoothing spline method SS, with the classical SARIMA model and with the Holt–Winters forecasting procedure HW. Method SS, which was introduced by Besse and Cardot (1996), assumes an ARH(1) structure for the time series $Z = (Z_i;\ i \in \mathbb{N})$ and handles the discretization problem of the observed curves by simultaneously estimating the sample paths and projecting the data on a q-dimensional subspace (that the predictable part of Z is assumed to belong to) by using smoothing splines (by solving an appropriate variational problem). The corresponding smoothing parameter λ and dimensionality q are chosen by a cross-validation criterion. Following the Box–Jenkins methodology (see Box and Jenkins (1976), chapter 9), a suitable SARIMA model is also fitted to the time series $Z = (Z_i;\ i \in \mathbb{N})$. Finally, the HW forecasting procedure (see, for example, Chatfield (1980), chapter 5), which is a variant of exponential smoothing dealing with time series containing trend and seasonal variation, is also applied to the time series $Z = (Z_i;\ i \in \mathbb{N})$.

The quality of the prediction methods was measured by the following often-used criteria (see, for example, Besse *et al.* (2000) and Antoniadis and Sapatinas (2003)):

(a) *the mean-square error* MSE, which is defined by

$$\mathrm{MSE} = \frac{1}{P}\sum_{i=1}^{P}\{\hat{Z}_{n_0}(t_i) - Z_{n_0}(t_i)\}^2, \tag{5}$$

(b) *the relative mean absolute error* RMAE, which is defined by

$$\mathrm{RMAE} = \frac{1}{P}\sum_{i=1}^{P}\frac{|\hat{Z}_{n_0}(t_i) - Z_{n_0}(t_i)|}{|Z_{n_0}(t_i)|}, \tag{6}$$

where Z_{n_0} is the n_0th element of the time series Z and $\hat{Z}_{n_0}$ is the prediction of Z_{n_0} given the past.

The computational algorithms that are related to wavelet analysis were performed using version 8.02 of the free software WaveLab. The overall numerical study has been carried out in the following programming environments: `Matlab 7.0` for the wavelet–kernel and smoothing spline estimators, and `SAS/ETS 6.0` for the SARIMA model `(proc arima)` and the HW `(proc forecast)` forecasting procedure.

3.1. A simulated example

We generated a series of observations as the superposition of two deterministic signals with different periods and first-order moving average noise. More specifically, we considered the following structure for X:

$$X(t) = \beta_1 m_1(t) + \beta_2 m_2(t) + \varepsilon(t), \tag{7}$$

where

$$m_1(t) = \cos(2\pi t/64) + \sin(2\pi t/64),$$
$$m_2(t) = \cos(2\pi t/6) + \sin(2\pi t/6),$$
$$\varepsilon(t) = u(t) + \theta\, u(t-1), \qquad u(t) \overset{\text{IID}}{\sim} N(0, \sigma^2).$$

The parameters were chosen as follows: $\beta_1 = 0.8$, $\beta_2 = 0.18$, $\theta = 0.8$ and $\sigma^2 = 0.005$. The motivation beyond this choice is to generate realizations containing a dominant component with a period of 64 observations, a less pronounced and more irregular component with a period of six observations, contaminated with an additive and correlated random component. The time period that we have analysed runs 30 segments, each containing 64 observations (i.e. $n = 1920$), and the last segment is displayed in Fig. 1(a), showing the above marked long periodicity together with the short, and randomly corrupted, pseudoperiodicity.

The bandwidth h_n for the method WK was chosen by cross-validation and was found to be equal to 0.8. We have compared our results with those obtained by using the method SS, with smoothing parameter λ and dimensionality q chosen by cross-validation and found to be equal to 10.1×10^{-3} and 2 respectively. A suitable ARIMA model, including a seasonality of 64 observations, has also been fitted to the time series, and the most parsimonious SARIMA model, containing a significant AR(6) component, validated through a portmanteau test for serial correlation of the fitted residuals, was selected. To complete the comparison, the HW forecasting procedure was also applied.

Fig. 1(a) also displays the various predictions that were obtained by the WK, SS, SARIMA and HW methods, whereas Fig. 1(b) displays the 95% resampling-based pointwise prediction interval for the simulated signal corresponding to the prediction that was obtained by method WK. MSE and RMAE for each prediction method are displayed in Table 1 (we have taken $n_0 = 30$ and $P = 64$). As observed in both Fig. 1 and Table 1, the predictions that are obtained by method WK are reasonably close to the true points, whereas the predictions that are made by method SS fail to capture the short pseudoperiodicities. Although the HW forecasting procedure totally ignores the small pseudoperiodicities, the predictions that are obtained by the SARIMA model gradually die out and converge to the 64-period cycle owing to the very long forecasting horizon. This example clearly illustrates the effect of the functional wavelet–kernel approach proposed. By treating all future observations that we would like to forecast as a segment, and using a notion of similarity which is based on a distance metric on the wavelet coefficients of a suitable decomposition of previous segments, the functional wavelet–kernel prediction method proposed can satisfactorily capture not only global but also local characteristics.

3.2. Electrical power consumption

The electrical load application concerns the prediction of electrical power consumption in Paris from half-hour daily recordings. The short-term predictions are based on data sampled over 30 min, which were obtained after eliminating certain components that are linked to weather conditions, calendar effects, outliers and known external actions. The data set analysed is part of a

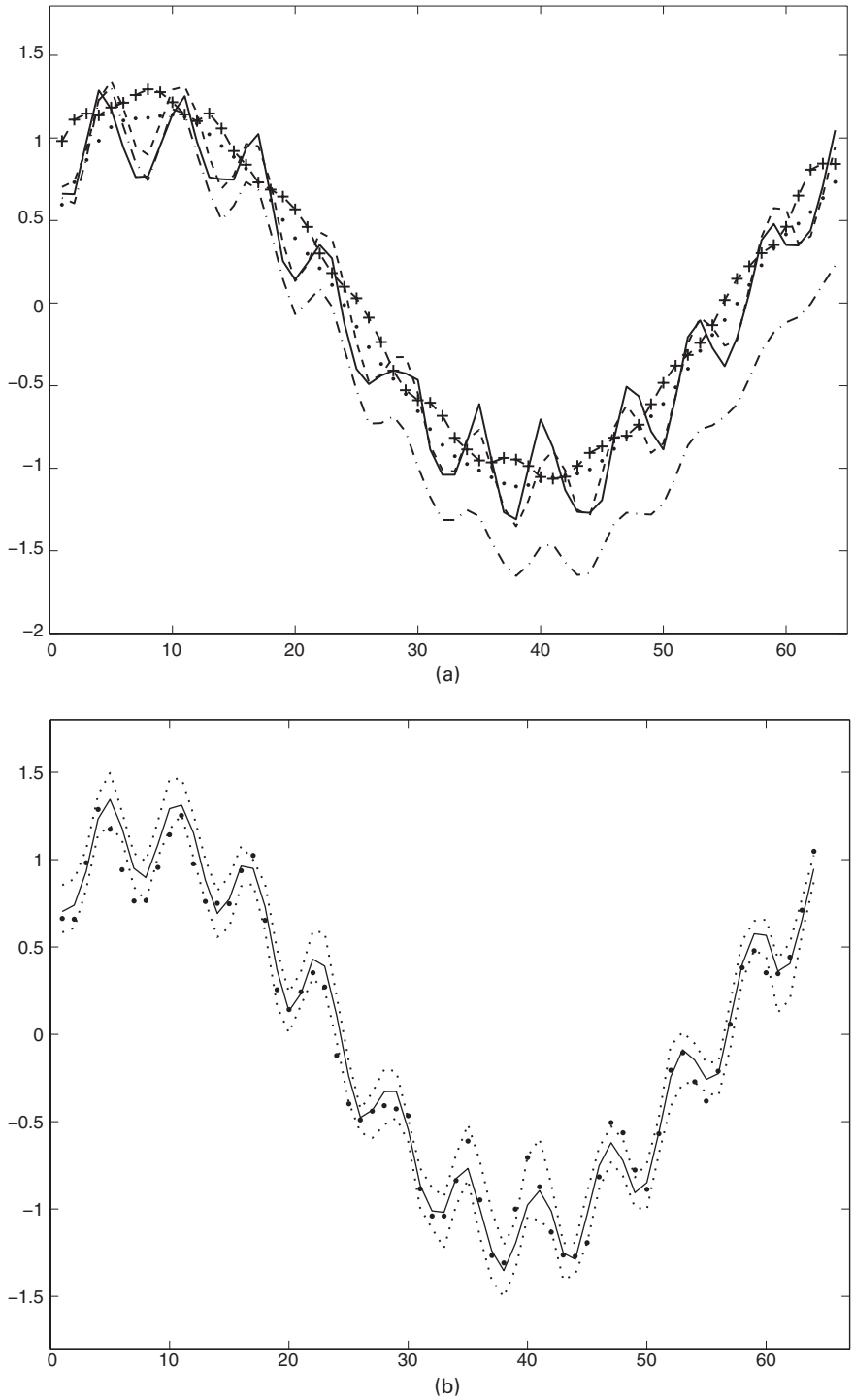

Fig. 1. (a) Simulated signal (———) and its various predictions by using methods WK (- - - - - - -), SS (· · · · · · ·), SARIMA (· - · - · - ·) and HW (+-+-+) and (b) 95% resampling-based pointwise prediction interval (· · · · · · ·) for the simulated signal, based on the corresponding prediction obtained by method WK (———) (•, true points)

Table 1. MSE and RMAE for the prediction of the simulated signal based on methods WK, SS, SARIMA and HW

Prediction method	*MSE*	*RMAE (%)*
WK	17.5×10^{-2}	1.16
SS	35.3×10^{-2}	3.67
SARIMA	87.3×10^{-2}	16.5
HW	50.3×10^{-2}	6.94

larger series that was recorded from the French national electricity company during the period running from August 1st, 1985, to July 4th, 1992. The time period that we have analysed runs for 35 days, starting from July 24th, 1991, to August 27th, 1991, and is displayed in Fig. 2. We note quite a regularity in this time series and a marked periodicity of 7 days (linked to economic rhythms) together with a pseudodaily periodicity. However, daily consumption patterns due to holidays, week-ends and discounts in electricity charges (e.g. relay-switched water heaters to benefit from special night rates) make the use of SARIMA modelling for forecasting problematic for about 10% of the days when working with half-hour data (see Misiti *et al.* (1994)).

To apply the forecasting methodology proposed we must choose the segmentation parameter δ which is strongly connected to the structure of the time series. A possibility would be to choose δ to exploit both the week periodicity and the daily pseudoperiodicity in the time series; another

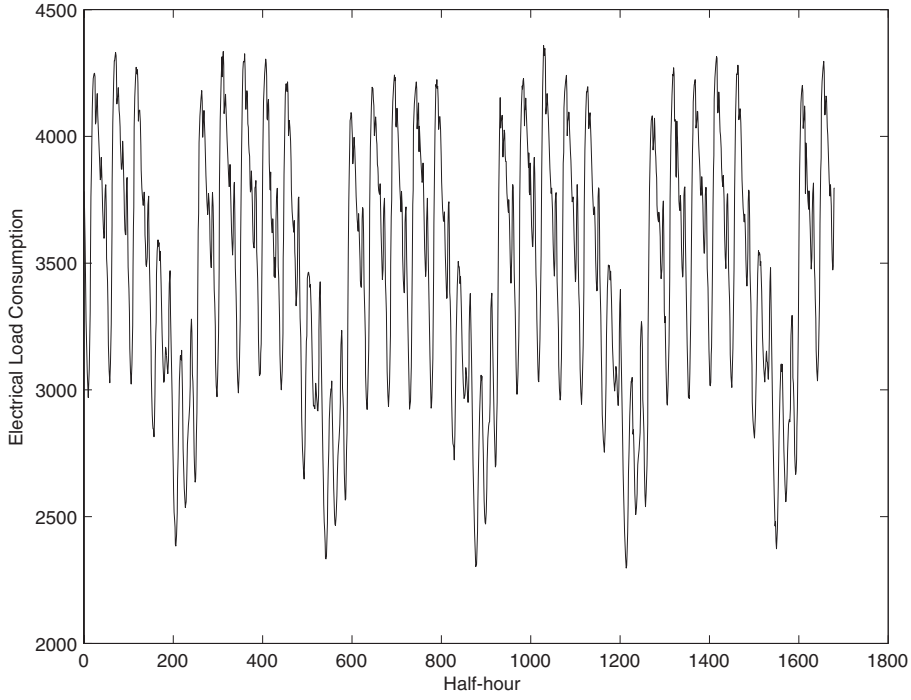

Fig. 2. Half-hour electrical power consumption in Paris from July 24th to August 27th, 1991

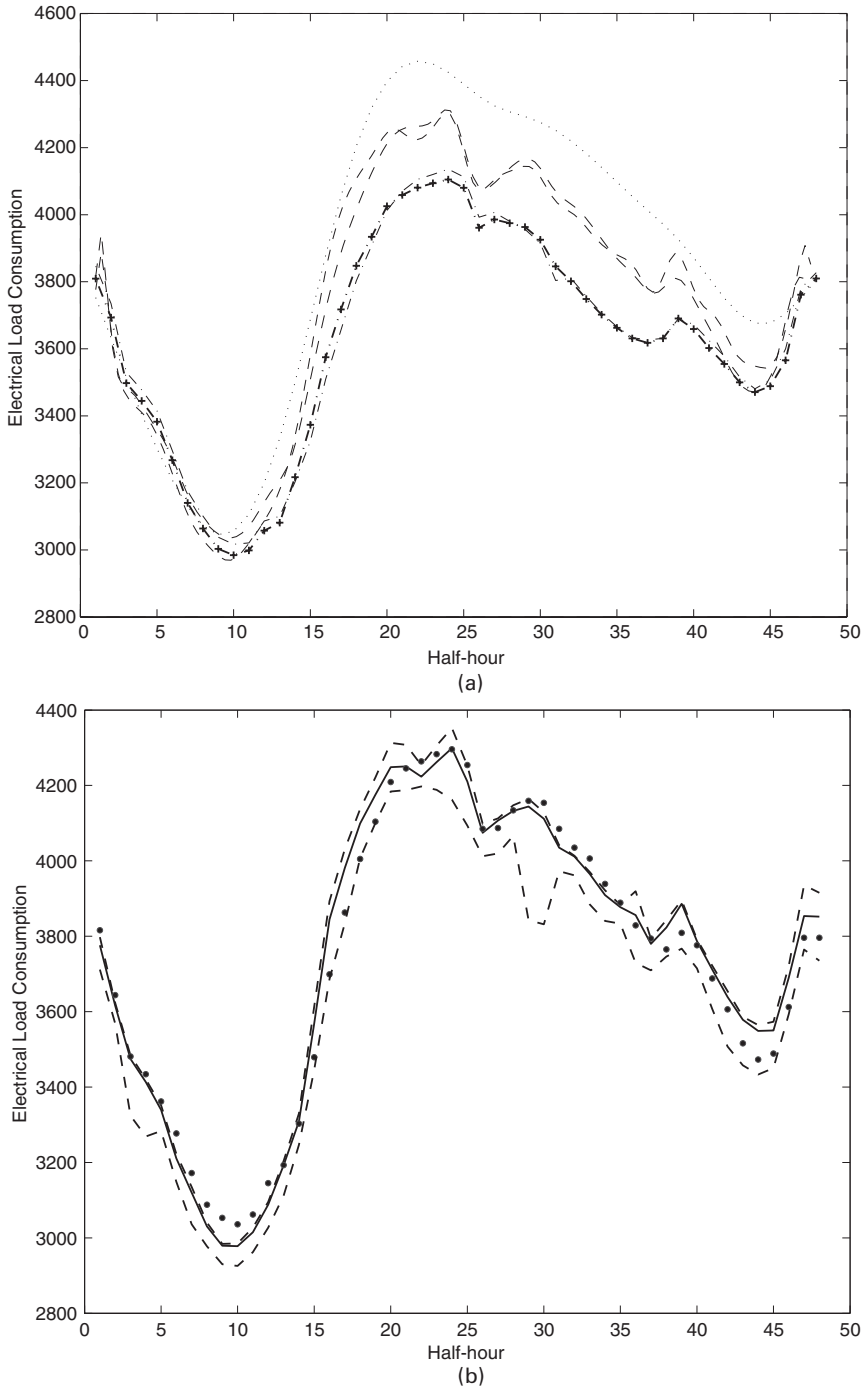

Fig. 3. (a) Half-hour electrical power consumption in Paris during August 27th, 1991 (———), and its various predictions by using methods WK (-------), SS (·····), SARIMA (·-·-·-) and HW (+-+-+-) and (b) 95% resampling-based pointwise prediction interval (-------) for the half-hour electrical power consumption in Paris during August 27th, 1991, based on the corresponding prediction obtained by method WK (———) (•, true points)

Table 2. MSE and RMAE for the prediction of half-hour electrical power consumption in Paris during August 27th, 1991, based on methods WK, SS, SARIMA and HW

Prediction method	*MSE*	*RMAE (%)*
WK	2.86×10^3	1.2
SS	2.38×10^4	3.6
SARIMA	1.88×10^4	2.9
HW	1.95×10^4	2.9

possibility could be to examine the similarity index based on a segmentation done by days. The above remarks suggest taking δ as a multiple of 48, between 48 (1 day) and 336 (1 week). However, segmentation by week should be avoided because the presence of a special day during a week (e.g. daily patterns due to holidays) would probably exclude it from the prediction of a 'normal' week. A better solution would be to choose $\delta = 48k$ with k being the minimal number of days that are necessary to induce a calendar homogeneity among the segments, say $k = 3$ or $k = 4$ according to the status that we give on Fridays. Such a choice would force the predictor to 'discover' the weekly periodicity. We have therefore chosen not to do this by taking $\delta = 48$, which is suitable for a 1-day-ahead prediction.

The bandwidth (h_n) for the wavelet–kernel method was chosen by cross-validation and was found to be equal to 0.01. We have compared our results with those obtained by using the method SS, with smoothing parameter λ and dimensionality q chosen by cross-validation and found to be equal to 5.56×10^4 and 4 respectively. A suitable ARIMA model, including 48 half-hour seasonality, has also been fitted to the time series from July 24th, 1991, to August 26th, 1991, and the most parsimonious SARIMA model, validated through a portmanteau test for serial correlation of the fitted residuals, was selected. To complete the comparison, the HW forecasting procedure was also applied.

Fig. 3(a) displays the observed data for August 27th, 1991, and its predictions obtained by the wavelet–kernel, SS, SARIMA and HW methods, whereas Fig. 3(b) displays the 95% resampling-based pointwise prediction interval corresponding to the prediction that was obtained by the WK method. MSE and RMAE for each prediction method are displayed in Table 2 (we have taken $n_0 = 35$ and $P = 48$). As observed in both Fig. 3 and Table 2, the prediction that is obtained by method WK is reasonably close to the true points, whereas the prediction that is made by method SS falls far from them. The predictions that are made by the SARIMA and HW methods are very similar but, although they are better than the predictions made by method SS, still fall far from the true points. This example clearly illustrates the effect of the proposed functional WK prediction method since the trajectory to be predicted seems not regular with some peculiar peaks.

3.3. Television audience rates

In France, Médiametrie has become the *de facto* national measurement service for the television industry. Among Médiamétrie's ratings calculations, we shall be interested in the one called 'cumulative rating', TTV, which measures the number of unique viewers (age 4 years and up) of a national television channel in a particular time period of the evening.

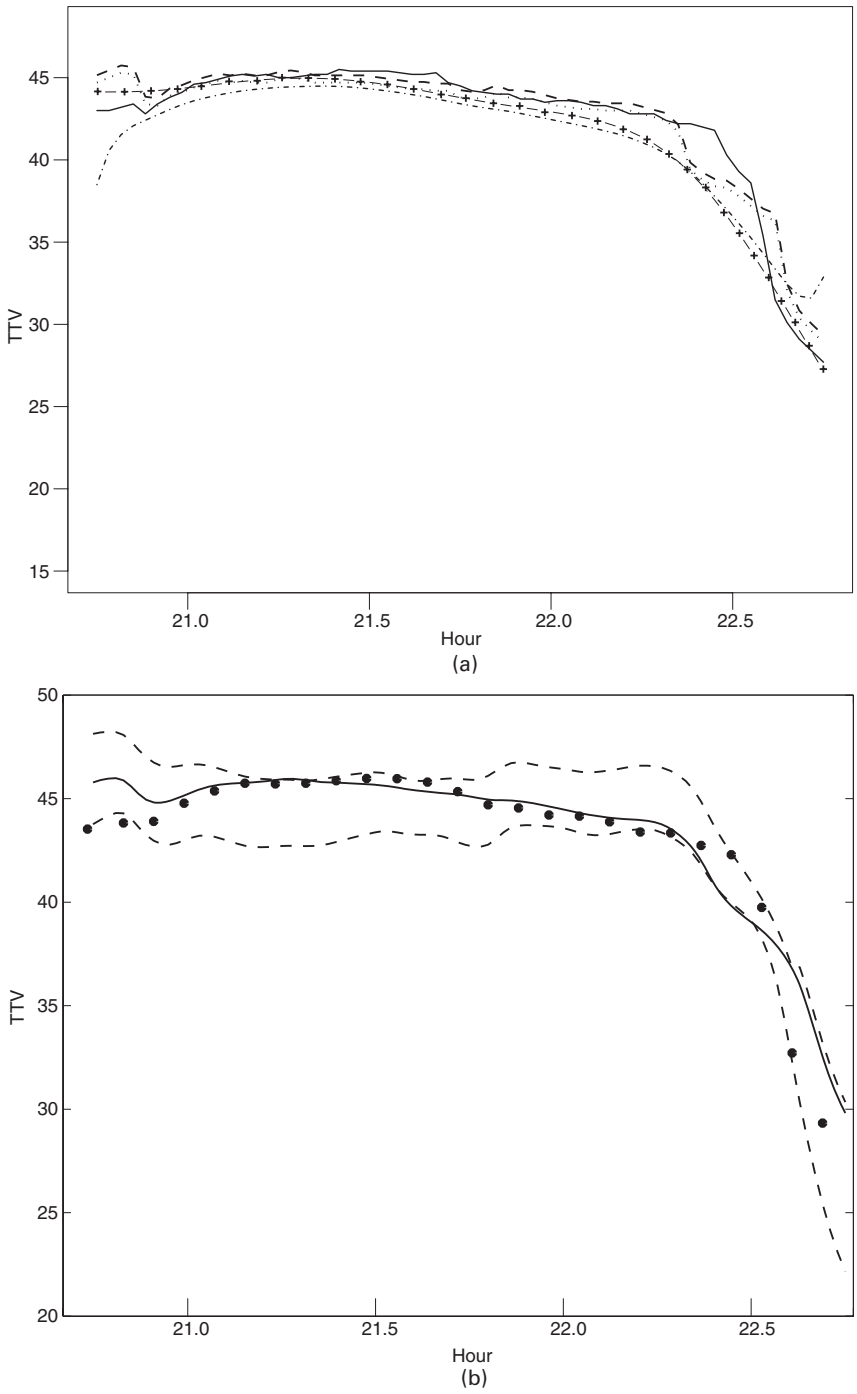

Fig. 4. TTV-rates for Monday, April 17th, 2000, in France during the evening time and their various predictions by using methods WK (-------), SS (·····), SARIMA (·-·-·-) and HW (+-+-+-) and (b) 95% resampling-based pointwise prediction interval (-------) for the TTV-rates for Monday, April 17th, 2000, in France during the evening time period based on the corresponding prediction obtained by method WK (———) (•, a subset of true points)

Table 3. MSE and RMAE for the prediction of the TTV-rates for Monday, April 17th, 2000, in France during the evening time period based on methods WK, SS, SARIMA and HW

Prediction method	*MSE*	*RMAE (%)*
WK	1.99	3.79
SS	1.91	3.73
SARIMA	2.13	4.08
HW	3.66	6.44

To illustrate the suggested forecasting methodology, we shall be interested here in predicting the rate TTV for a particular day of the week and for the particular evening time period from 8.45 p.m. to 10.45 p.m. The reason for segmenting the rates per day of the week is because the offer of television programmes differs considerably from one day to another. The data that are analysed are Monday series of TTV-recordings (averaged every 2 min) in the time period 8.45–10.45 p.m. from October 1st, 1998, to May 31st, 2000 (88 weeks). Each TTV-curve is therefore composed of 61 observations. We have used the first 87 weeks as a training sample, and the remaining last week for testing our procedures and for computing the error rates.

The bandwidth h_n for the method WK was chosen by cross-validation and found to be equal to 5.4. We have compared our results with those obtained by using method SS, with smoothing parameter λ and dimensionality q chosen by cross-validation and found to be equal to 89.3 and 2 respectively. A suitable ARIMA model, including weekly seasonality, has also been fitted to the time series from October 1st, 1998, to April 12th, 2000, and the most parsimonious SARIMA model, validated through a portmanteau test for serial correlation of the fitted residuals, was selected. To complete the comparison, the HW forecasting procedure was also applied.

Fig. 4(a) displays the observed data for Monday, April 17th, 2000, and its various predictions obtained by methods WK, SS, SARIMA and HW. MSE and RMAE rates of each prediction method are displayed in Table 3 (we have taken $n_0 = 88$ and $P = 61$). As observed in both Fig. 4 and Table 3, the predictions that are made by methods WK and SS are almost identical, with SS performing slightly better in terms of error rates. In contrast, although the SARIMA method performs better than method HW, both are inferior to the predictions that are made by the functional-based methods. Also observe that, at the beginning of the time period and also at the end after 10.00 p.m., all predictions are biased and hence not very close to the true points. This difficulty in prediction is captured in Fig. 4(b), which displays the corresponding 95% resampling-based pointwise prediction interval for the Monday TTV-rate based on the corresponding prediction obtained by the method WK. The main reason for this bias is that the programmes that are scheduled during this time period do not start exactly at their scheduled time or finish abruptly (e.g. at the end of a film).

Acknowledgements

Anestis Antoniadis was supported by Interuniversity Attraction Poles research network P5/24 and the 'Cyprus–France CY-FR/0204/04 Zenon program'. Efstathios Paparoditis and Theofanis Sapatinas were supported by the 'Cyprus–France CY-FR/0204/04 Zenon program'. We thank Jean-Michel Poggi (Université Paris-Sud, France) for providing us with the Paris electrical

power consumption data and Médiamétrie for providing us with the Monday TTV-rates data. The authors are grateful to the Joint Editor, the Associate Editor and the two referees, whose valuable and constructive comments led to a significant improvement of this paper.

Appendix A

Our theoretical results are derived under α-mixing assumptions on the time series $Z=(Z_i;\ i\in\mathbb{N})$. For a strictly stationary series $Z=(Z_i;\ i\in\mathbb{N})$, the α-mixing coefficient (see Rosenblatt (1956)) is defined by

$$\alpha_Z(m)=\sup_{A\in\mathcal{D}_l,B\in\mathcal{D}_{l+m}}|\mathbb{P}(A\cap B)-\mathbb{P}(A)\,\mathbb{P}(B)|,$$

where $\mathcal{D}_l=\sigma(Z_i,i\leqslant l)$ and $\mathcal{D}_{l+m}=\sigma(Z_i,i\geqslant l+m)$ are the σ-fields that are generated by $(Z_i;\ i\leqslant l)$ and $(Z_i;\ i\geqslant l+m)$ respectively, for any $m\geqslant 1$. The stationary sequence $Z=(Z_i;\ i\in\mathbb{N})$ is said to be α mixing if $\alpha_Z(m)\to 0$ as $m\to\infty$. Among various mixing conditions that have been used in the literature, α-mixing is reasonably weak (see, for example, Doukhan (1994)).

Recall that $\Xi_i=\{\xi_i^{(J,k)}:\ k=0,1,\ldots,2^J-1\}$ denotes the set of scaling coefficients at scale J of the ith segment Z_i and let $\mathcal{A}_{J,l}=\sigma(\xi_i^{J,k},i\leqslant l)$ and $\mathcal{A}_{J,l+m}=\sigma(\xi_i^{J,k},i\geqslant l+m)$ be the σ-fields that are generated by $(\xi_i^{(J,k)};\ i\leqslant l)$ and $(\xi_i^{(J,k)};\ i\geqslant l+m)$ respectively. Because $\sigma(\xi_i^{(J,k)},i\in I)\subseteq\sigma(Z_i,i\in I)$ for any $I\subset\mathbb{N}$, we obtain

$$\begin{aligned}\alpha_{J,k}(m)&=\sup_{A\in\mathcal{A}_{J,l},B\in\mathcal{A}_{J,l+m}}|\mathbb{P}(A\cap B)-\mathbb{P}(A)\,\mathbb{P}(B)|\\&\leqslant\sup_{A\in\mathcal{D}_l,B\in\mathcal{D}_{l+m}}|\mathbb{P}(A\cap B)-\mathbb{P}(A)\,\mathbb{P}(B)|\\&=\alpha_Z(m).\end{aligned}$$

Our asymptotic results will be based on the following set of assumptions, which we detail below before proceeding to the proofs.

A.1. Main assumptions

We first impose assumptions on the sample paths of the underlying stochastic process.

A.1.1. Assumption (A1)

When we observe a fixed number P of sampled values in each sample path, we assume that the sample paths of the strictly stationary process $Z=(Z_i;\ i\in\mathbb{N})$ are continuous on $[0,\delta)$, and that the scaling function ϕ of the wavelet basis has an exponential decay (see expression (4.1) in Meyer (1992)).

A.1.2. Assumption (A2)

The α_Z-mixing coefficient of the strictly stationary process $Z=(Z_i;\ i\in\mathbb{N})$ satisfies

$$\sum_{m=N}^{\infty}\alpha_Z(m)^{1-2/l}=O(N^{-1})\qquad\text{for some } l>4. \tag{8}$$

We next impose some assumptions on the joint and conditional probability density functions of the scaling coefficients $\xi_i^{(J,k)}$.

A.1.3. Assumption (A3)

$E|\xi_i^{(J,k)}|^l<\infty$, for $l>4$ and every $k=0,1,\ldots,2^J-1$.

A.1.4. Assumption (A4)

The probability density function f_{Ξ_i} of Ξ_i exists, is absolutely continuous with respect to Lebesgue measure and satisfies the conditions

(a) f_{Ξ_i} is Lipschitz continuous, i.e.

$$|f_{\Xi_i}(x)-f_{\Xi_i}(y)|\leqslant C\|x-y\|.$$

(b) For any compact subset S of $\mathbb{R}^{2^J}$, $\min_s(f_{\Xi_i}) > 0$.
(c) The conditional probability density function of $\xi_{i+1}^{(J,k)}$ given Ξ_i is bounded, i.e. $f_{\xi_{i+1}^{(J,k)}|\Xi_i}(\cdot|x) \leqslant C < \infty$.

We also impose some conditions on the kernel function and the bandwidth that is associated with it.

A.1.5. Assumption (A5)

The (univariate) kernel K is a bounded symmetric density on $\mathbb{R}$ satisfying $|K(x) - K(y)| \leqslant C|x - y|$ for all $x, y \in \mathbb{R}$. Furthermore, $\int x\,K(x)\,\mathrm{d}x = 0$ and $\int x^2\,K(x)\,\mathrm{d}x < \infty$.

A.1.6. Assumption (A6)

The bandwidth h_n satisfies $h_n \to 0$ and $nh_n^{2^J}/\log(n) \to \infty$ as $n \to \infty$.

Let us now explain the meaning of these assumptions. Assumptions (A1) and (A2) are quite common in time series prediction (see Bosq (1998)). Assumptions (A3) and (A4) are essentially made on the distributional behaviour of the scaling coefficients at scale J and, therefore, are less restrictive. They are moreover natural in nonparametric regression. Assumption (A4), part (b), is needed for obtaining consistency results. We have used assumption (A4), part (c), to make the presentation clearer. However, it can be relaxed into the existence of absolute moments by means of conventional truncation techniques that are used in the $\mathbb{R}$-valued situation (see, for example, Mack and Silverman (1982)). Conditions (A5) and (A6) are classical for kernel regression estimation.

A.2. Proof of theorem 1

Since each observed segment is a time series with fixed (finite) length, the use of a wavelet transform at the appropriate resolution J makes the approximation error negligible, i.e.

$$\mathbb{E}(Z_{n+1}^J|Z_n) \simeq \mathbb{E}(Z_{n+1}|Z_n).$$

Hence, we proceed by deriving the appropriate convergence rate for

$$\|Z_{n+1|n}^J - \mathbb{E}(Z_{n+1}^J|Z_n)\|.$$

We first show that, as $n \to \infty$,

$$\|\Xi_{n+1|n} - \mathbb{E}(\Xi_{n+1}|\Xi_n)\| \to 0, \qquad \text{almost surely.} \tag{9}$$

For this, it suffices to show that, for every $k = 0, 1, \ldots, 2^J - 1$, as $n \to \infty$,

$$\xi_{n+1|n}^{(J,k)} \to \mathbb{E}(\xi_{n+1}^{(J,k)}|\Xi_n), \qquad \text{almost surely.}$$

Let $x \in \mathbb{R}^{2^J}$, let $\Xi_{n+1|n}(x)$ be the value of $\Xi_{n+1|n}$ in equation (3) for $\Xi_n = x$ and denote by $\xi_{n+1|n}^{(J,k)}(x)$ the kth component of $\Xi_{n+1|n}(x)$. Consider the 2^J-dimensional random variable $W_l = \mathcal{C}(\Xi_l)$, and denote by $f_{\xi_{l+1}^{(J,k)},W_l}$ and f_{W_l} the joint and marginal densities of $(\xi_{l+1}^{(J,k)}, W_l)$ and W_l respectively. Because of condition (A4), and the fact that W_l is a linear transformation of Ξ_l, $f_{\xi_{l+1}^{(J,k)},W_l}$ and f_{W_l} exist with respect to Lebesgue measure for every $k = 0, 1, \ldots, 2^J - 1$. Let

$$\hat{f}_{W_l}(x) = (nh_n^{2^J})^{-1}\left[\sum_{m=1}^{n-1} K\left\{\frac{D(x, W_m)}{h_n}\right\} + \frac{1}{n}\right]$$

and note that $\hat{f}_{W_l}(x)$ is a kernel estimator of the 2^J-dimensional density $f_{W_l}(x)$. The added factor $1/n$ does not affect the rate of convergence of $\hat{f}_{W_l}$ but ensures that it is strictly positive for any n. For notational convenience, in what follows, let $\Phi_{n,k}(x) = \mathbb{E}(\xi_{n+1}^{(J,k)}|\Xi_n = x)$ and

$$\hat{g}_{n,k}(x) = (nh_n^{2^J})^{-1}\sum_{m=1}^{n-1} K\left\{\frac{D(x, W_m)}{h_n}\right\}\xi_{m+1}^{(J,k)}.$$

We then have

$$\xi_{n+1|n}^{(J,k)}(x) - \mathbb{E}(\xi_{n+1}^{(J,k)}|\Xi_n = x) = \frac{1}{\hat{f}_{W_l}(x)}\{\hat{g}_{n,k}(x) - \Phi_{n,k}(x)\,f_{W_l}(x)\} - \frac{\Phi_{n,k}(x)}{\hat{f}_{W_l}(x)}\{\hat{f}_{W_l}(x) - f_{W_l}(x)\}. \tag{10}$$

Using now the assumptions of theorem 1, this decomposition, and remarks 4.1 and 4.2 in Ferraty *et al.* (2002), it follows that, as $n \to \infty$,

$$\max_k \{\sup_{x \in S} |\xi^{(J,k)}_{n+1|n}(x) - \mathbb{E}(\xi^{(J,k)}_{n+1} | \Xi_n = x)|\} = O\left[\left\{ \frac{\log(n)}{n} \right\}^{1/(2+2^J)} \right], \qquad \text{almost surely.} \tag{11}$$

Recalling now that our estimator is defined as

$$Z^J_{n+1|n}(t) = \sum_{k=0}^{2^J-1} \xi^{(J,k)}_{n+1|n}\, \phi_{J,k}(t),$$

using the convergence rate that is given in expression (11), and the fact that we have used a regular multiresolution analysis, we have, for $Z_n = x$, as $n \to \infty$,

$$\begin{aligned} \sup_t |Z^J_{n+1,n}(t) - \mathbb{E}\left\{ Z^J_{n+1}(t) | Z_n = x \right\}| &\leqslant 2^{J/2} \max_k \left[\sup_{x \in S} |\xi^{(J,k)}_{n+1|n} - \mathbb{E}(\xi^{(J,k)}_{n+1} | \Xi_n = x)| \sup_t \left\{ \sum_{k=0}^{2^J-1} |\phi(2^J t - k)| \right\} \right] \\ &= O\left[\left\{ \frac{\log(n)}{n} \right\}^{1/(2+2^J)} \right], \qquad \text{almost surely.} \end{aligned} \tag{12}$$

Bound (12) ensures the validity of the assertion. This completes the proof of theorem 1.

A.3. Proof of theorem 2

For every $t_i \in \{t_1, t_2, \ldots, t_P\}$, note that $Z_{n+1}(t_i) = \xi^{(J,i)}_{n+1}$. Since

$$\begin{aligned} L^*_{n+1,\alpha}(t_i) &= R^*_{n+1,\alpha}(t_i) + Z^J_{n+1|n}(t_i) \\ &= \mathbb{E}\{Z_{n+1}(t_i) | Z_n\} + [Z^*_{n+1}(t_i) - \mathbb{E}\{Z_{n+1}(t_i) | Z_n\}], \end{aligned}$$

it suffices to show that the distribution of $Z^*_{n+1}(t_i) - \mathbb{E}\{Z_{n+1}(t_i) | Z_n\}$ approximates correctly the conditional distribution of $Z_{n+1}(t_i) - \mathbb{E}\{Z_{n+1}(t_i) | Z_n\}$ given Z_n.

Now, given $Z_n = x$, i.e. given $\Xi_n = \tilde{x}$, we have

$$\begin{aligned} \mathbb{P}[Z^*_{n+1}(t_i) - \mathbb{E}\{Z_{n+1}(t_i) | Z_n = x\} \leqslant y] &= \sum_{m=1}^{n-1} \mathbf{1}_{(-\infty, y]}[Z_{m+1}(t_i) - \mathbb{E}\{Z_{n+1}(t_i) | Z_n = x\}] w_{n,m} \\ &= \sum_{m=1}^{n-1} \mathbf{1}_{(-\infty, \tilde{y}]}\, Z_{m+1}(t_i)\, w_{n,m} \\ &= \sum_{m=1}^{n-1} \mathbf{1}_{(-\infty, \tilde{y}]}(\xi^{(J,i)}_{m+1}) w_{n,m} \\ &= \frac{\sum_{m=1}^{n-1} \mathbf{1}_{(-\infty, \tilde{y})}(\xi^{(J,i)}_{m+1}) K[D\{\mathcal{C}(\tilde{x}), \mathcal{C}(\Xi_m)\}/h_n]}{n^{-1} + \sum_{m=1}^{n-1} K[D\{\mathcal{C}(\tilde{x}), \mathcal{C}(\Xi_m)\}/h_n)] } \\ &\quad + O(n^{-1}), \end{aligned} \tag{13}$$

where $\tilde{y} = y + \mathbb{E}\{Z_{n+1}(t_i) | Z_n = x\}$. Note that expression (13) is a kernel estimator of the conditional mean $\mathbb{E}\{\mathbf{1}_{(-\infty,\tilde{y}]}(\xi^{(J,i)}_{n+1}) | \Xi_n = \tilde{x}\} = \mathbb{P}(\xi^{(J,i)}_{n+1} \leqslant \tilde{y} | \Xi_n = \tilde{x})$, i.e. of the conditional distribution of $\xi^{(J,i)}_{n+1}$ given that $\Xi_n = \tilde{x}$. Denote now the conditional distribution of $\xi^{(J,i)}_{n+1}$ given Ξ_n by $F_{\xi^{(J,i)}_{n+1}|\Xi_n}(\cdot|\Xi_n)$ and its kernel estimator given in expression (13) by $\hat{F}_{\xi^{(J,i)}_{n+1}|\Xi_n}(\cdot|\Xi_n)$. Then, by the same arguments as in theorem 1, we obtain that, for every $y \in \mathbb{R}$, as $n \to \infty$,

$$\sup_{x \in S} |\hat{F}_{\xi^{(J,i)}_{n+1}|\Xi_n}(y|x) - F_{\xi^{(J,i)}_{n+1}|\Xi_n}(y|x)| \to 0, \qquad \text{in probability.}$$

It remains to show that this convergence is also uniform over y. Fix now an x in the support S of Ξ_n, and let $\varepsilon > 0$ be arbitrary. Since $F_{\xi^{(J,i)}_{n+1}|\Xi_n}(y|x)$ is continuous we have that, for every $k \in \mathbb{N}$, points

$-\infty = y_0 < y_1 < \ldots < y_{k-1} < y_k = \infty$ exist such that $F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x) = i/k$. For $y_{i-1} \leqslant y \leqslant y_i$, and using the monotonicity of $\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}$ and $F_{\xi_{n+1}^{(J,i)}|\Xi_n}$, we have

$$\begin{aligned}\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_{i-1}|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x) &\leqslant \hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y|x) \\ &\leqslant \hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x).\end{aligned}$$

From this, we obtain

$$|\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y|x)| \leqslant \sup_i |\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x)| + \frac{1}{k},$$

and, therefore,

$$\begin{aligned}\mathbb{P}\{|\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y|x)| > \varepsilon\} &\leqslant \mathbb{P}\{\sup_i |\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x)| + k^{-1} > \varepsilon\} \\ &\leqslant \mathbb{P}\{\sup_i \sup_x |\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x)| + k^{-1} > \varepsilon\}.\end{aligned}$$

Now, choose k sufficiently large that $1/k < \varepsilon/2$. For such a fixed k, and because, for every $y \in \mathbb{R}$, as $n \to \infty$,

$$\sup_x |\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y|x)| \to 0, \qquad \text{in probability,}$$

we can choose n sufficiently large that

$$\mathbb{P}\{\sup_i \sup_x |\hat{F}_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x) - F_{\xi_{n+1}^{(J,i)}|\Xi_n}(y_i|x)| > \varepsilon/2\} < \tau,$$

for any desired τ. Since τ is independent on y and x, the desired convergence follows. This completes the proof of theorem 2.

References

Antoniadis, A. and Sapatinas, T. (2003) Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. *J. Multiv. Anal.*, **87**, 133–158.

Besse, P. C. and Cardot, H. (1996) Approximation spline de la prévision d'un processus fonctionnel autorégressif d'ordre 1. *Can. J. Statist.*, **24**, 467–487.

Besse, P. C., Cardot, H. and Stephenson, D. B. (2000) Autoregressive forecasting of some functional climatic variations. *Scand. J. Statist.*, **27**, 673–687.

Bosq, D. (1991) Modelization, nonparametric estimation and prediction for continuous time processes. In *Nonparametric Functional Estimation and Related Topics* (ed. G. Roussas), pp. 509–529. Dordrecht: Kluwer.

Bosq, D. (1998) Nonparametric statistics for stochastic processes. *Lect. Notes Statist.*, **110**.

Bosq, D. (2000) Linear processes in function spaces. *Lect. Notes Statist.*, **149**.

Bosq, D. and Delecroix, M. (1985) Nonparametric prediction of a Hilbert-space valued random variable. *Stochast. Processes Appl.*, **19**, 271–280.

Box, G. E. and Jenkins, G. M. (1976) *Time Series Analysis and Control*. San Francisco: Holden-Day.

Brockwell, P. J. and Davis, R. A. (1991) *Time Series: Theory and Methods*, 2nd edn. New York: Springer.

Chatfield, C. (1980) *The Analysis of Time Series*, 4th edn. London: Chapman and Hall.

Cheng, B. and Tong, H. (1998) *k*-stationarity and wavelets. *J. Statist. Planng Inf.*, **68**, 129–144.

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: Society for Industrial and Applied Mathematics.

Doukhan, P. (1994) *Mixing: Properties and Examples*. New York: Springer.

Ferraty, F., Goia, A. and Vieu, P. (2002) Functional nonparametric model for time series: a fractal approach for dimension reduction. *Test*, **11**, 317–344.

Härdle, W. and Vieu, P. (1992) Kernel regression smoothing of time series. *J. Time Ser. Anal.*, **13**, 209–232.

Hart, J. D. (1996) Some automated methods of smoothing time-dependent data. *J. Nonparam. Statist.*, **6**, 115–142.

Lipster, R. S. and Shiryayev, A. N. (1977) *Statistics of Random Processes*, vol. I, *General Theory*. New York: Springer.

Mack, Y. P. and Silverman, B. W. (1982) Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Ver. Geb.*, **62**, 405–415.

Mallat, S. G. (1989) A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn Anal. Mach. Intell.*, **11**, 674–693.

Meyer, Y. (1992) *Wavelets and Operators*. Cambridge: Cambridge University Press.
Misiti, M., Misiti, Y., Oppenheim, G. and Poggi, J.-M. (1994) Décomposition en ondelettes et méthodes comparatives: étude d'une courbe de charge électrique. *Rev. Statist. Appl.*, **42**, 57–77.
Pumo, B. (1998) Prediction of continuous time processes by $C[0, 1]$-valued autoregressive processes. *Statist. Inf. Stochast. Processes*, **3**, 297–309.
Rosenblatt, M. (1956) A central limit theorem and a strong mixing condition. *Proc. Natn. Acad. Sci. USA*, **42**, 43–47.