Comparison of DFA vs wavelet analysis for estimation of regularity of HR series during the marathon

Imen Kammoun [a], Véronique Billat [b], Jean-Marc Bardet [a,*]

[a] *Université Paris 1, SAMOS-MATISSE-CES, 90 rue de Tolbiac, 75013 Paris, France*

[b] *Université d'Evry, LEPHE, E.A. 3872 Genopole, Boulevard F. Mitterrand, Evry Cedex, France*

Abstract

In order to interpret and explain the physiological signal behaviors, it can be interesting to find some constants among the fluctuations of these data during all the effort or during different stages of the race. These different stages can be detected using a change points detection method. Then, the Hurst parameter of long-range dependence could be a new way for deducing some explanations. Several common estimators of this parameter, so-called scaling behavior exponents, consist in performing a linear regression fit of a scale-dependent quantity versus the scale in a logarithmic representation. This includes the Detrended Fluctuation Analysis (DFA) method and wavelet analysis method. This second method provides more robust results and can be applied to more general models. Then, it permits us the construction of the semi-parametric process which could be more relevant than other for modelling HR data. It also shows an evolution of the Hurst parameter during the race, what confirms results obtained by Peng *et al.* in their study concerning recorded HR time series during the exercise for healthy adults (where the estimated parameter is close to that observed in the race beginning) and heart failure adults (where the estimated parameter is close to that observed in the end of race). So, this evolution, which can not be observed with DFA method, may be associated with fatigue appearing during the last phase of the marathon.

Key words: Wavelet analysis, Detrended fluctuation analysis, Fractional Gaussian noise, Self-similarity, Hurst parameter, Long-range dependence processes, Heart rate time series

* Corresponding author.
Email address: bardet@univ-paris1.fr (Jean-Marc Bardet).

Preprint submitted to Elsevier *17 August 2007*

1 Introduction

Fifty athletes were followed during a marathon (Paris Marathon 2004) and for each one, various physiological parameters are measured. Each heart rates (HR) signal, recorded instantaneously on cardio-frequency meter (CFM), corresponds to an endurance type effort observed on a course of 42km realized, on average, in 3h14mn. For each runner, the periods (in ms) between the successive pulsations (see Fig. 1) are recorded. The HR signal in number of beats per minute (bpm) is then deduced (the HR average for the whole sample is of 162 bpm).

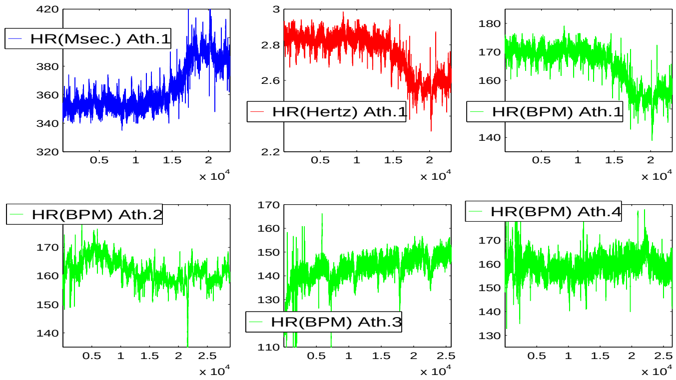

Fig. 1. Heat rate signals of Athlete 1 in ms, Hertz and BPM (up), of Athletes 2, 3 and 4 in BPM (down)

During effort, one or more phases can be observed, which evolve and change differently from an athlete to another. Moreover, data depend in particular on the installation of the CFM. So, as a first step of this study, a pretreatment of these data is proposed for "cleaning" them of outliers and detecting different significative stages during the race. In Section 3, two methods are presented for estimating the regularity parameter : the DFA method and wavelet analysis which is developed for a more general models. The last of section is devoted to applications of both methods to generated data and HR data.

2 Data processing

2.1 Abrupt change detection

HR data of each athlete may show various modes: before the race beginning, during a transition step (recorded between the race beginning and the stage of HR reached during the effort), the main stage during the exercise, an arrival phase until the race end and sometimes a recovery phase. For distinguishing these different steps, a method of change points detection developed by

Lavielle (see for instance [20]) is adapted and applied.
To begin with, a first treatment consists in detecting the beginning and the end of the race. The main idea is to consider that the signal distribution depends on a vector of unknown characteristic parameters in each stage. The different stages (before, during and after the race) and therefore the different vectors of parameters, change at two unknown instants (here the number of change points is known, but the method can be also used even if its number is unknown by adding a penalization term, see above). For instance and it will be our choice, changes in mean and variance can be detected.

General principle of the method of change detection

Assume that a sample of a time series $(Y(i), i = 1, \dots, n)$ is observed. Assume also that it exists $\tau = (\tau_1, \tau_2, \dots, \tau_{K-1})$ with $0 = \tau_0 < \tau_1 < \tau_2 < ... < \tau_{K-1} < n = \tau_K$ and such that for each $j \in \{1, 2, \dots, K\}$, the distribution law of $Y(i)$ is depending on a parameter $\theta_j \in \Theta \subset \mathbb{R}^d$ (with $d \in \mathbb{N}$) for all $\tau_{j-1} < i \leq \tau_j$. Therefore, K is the number of segments to be deduced starting from the series and $\tau = (\tau_1, \tau_2, \dots, \tau_{K-1})$ is the ordered change instants.

Now, define a contrast function

$$U_\theta\Big(Y(\tau_j + 1), Y(\tau_j + 2), \dots, Y(\tau_{j+1})\Big),$$

of $\theta \in \mathbb{R}^d$ applied on each vector $\Big(Y(\tau_j + 1), Y(\tau_j + 2), \dots, Y(\tau_{j+1})\Big)$ for all $j \in \{0, 2, \dots, K - 1\}$. A general example of such a contrast function is

$$U_\theta\Big(Y(\tau_j{+}1), Y(\tau_j{+}2), \dots, Y(\tau_{j+1})\Big) = -2 \log L_\theta\Big(Y(\tau_j{+}1), Y(\tau_j{+}2), \dots, Y(\tau_{j+1})\Big),$$

where L_θ is the likelihood. Then, for all $j \in \{0, 2, \dots, K - 1\}$, define:

$$\widehat{\theta}_j = \operatorname*{Argmin}_{\theta \in \Theta} U_\theta\Big(Y(\tau_j + 1), Y(\tau_j + 2), \dots, Y(\tau_{j+1})\Big).$$

Now, set:

$$\widehat{G}(\tau_1, \dots, \tau_{K-1}) = \sum_{j=0}^{K-1} U_{\widehat{\theta}_j}\Big(Y(\tau_j + 1), Y(\tau_j + 2), \dots, Y(\tau_{j+1})\Big)$$

As a consequence, an estimator $(\widehat{\tau}_1, \dots, \widehat{\tau}_{K-1})$ can be defined as:

$$(\widehat{\tau}_1, \dots, \widehat{\tau}_{K-1}) = \operatorname*{Argmin}_{0<\tau_1<\tau_2<\dots<\tau_{K-1}<n} \widehat{G}(\tau_1, \dots, \tau_{K-1}). \qquad (1)$$

The principle of such method of estimation is very general (it can be also devoted to estimate abrupt change in polynomial trends) and different asymptotic behavior of the estimator $(\widehat{\tau}_1, \dots, \widehat{\tau}_{K-1})$ can be deduced under general

assumption on the time series Y (see for instance Bai Perron, Lavielle Moulines and Lavielle).
For HR data, it is obvious that the beginning and the end of the race implies respectively an increasing (respectively decreasing) of the mean of HR. However, for avoiding all confusion linked for instance to the stress of the runner or other harmful noises, it was chosen to detect a change in mean and variance.

Change detection in mean and variance

Therefore, for all $j \in \{0, 1, \ldots, K-1\}$, consider the following general model:

$$Y(i) = \mu_j + \sigma_j \varepsilon_i \quad \text{for all } i \in \{\tau_j + 1, \ldots, \tau_{j+1}\},$$

where $\theta_j = (m_j, \sigma_j) \in \mathbb{R} \times (0, \infty)$ and (ε_i) is a sequence of zero-mean random variables with unit variance.
In the case of changes in both mean and variance, and it is such a framework we consider for the heart rates series, a "natural" contrast function is defined by:

$$U_{\theta_j}\Big(Y(\tau_j + 1), \ldots, Y(\tau_{j+1})\Big) = \sum_{\ell=\tau_j+1}^{\tau_{j+1}} \frac{(Y(\ell) - m_j)^2}{\sigma_j^2},$$

an therefore the well-known estimator of θ_j is:

$$\widehat{\theta}_j = (\widehat{m}_j, \widehat{\sigma}_j) = \Big(\frac{1}{\tau_{j+1} - \tau_j} \sum_{\ell=\tau_j+1}^{\tau_{j+1}} Y(\ell), \frac{1}{\tau_{j+1} - \tau_j} \sum_{\ell=\tau_j+1}^{\tau_{j+1}} \Big(Y(\ell) - \widehat{m}_j\Big)^2\Big).$$

Now, the estimator $(\widehat{\tau}_1, \ldots, \widehat{\tau}_{K-1})$ can be deduced from (1).

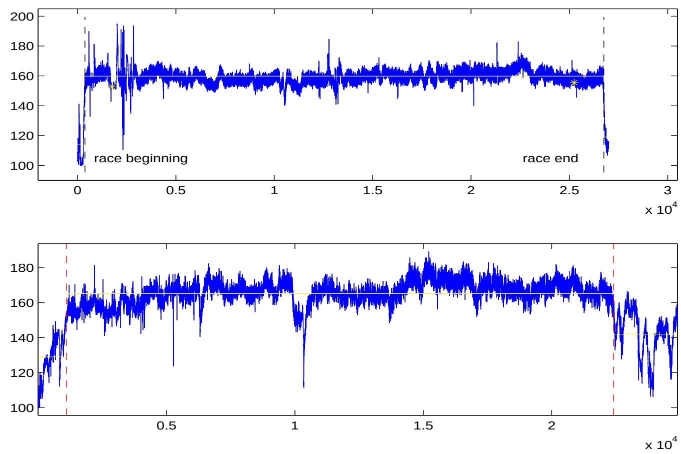

Fig. 2. Detection of the race beginning and end from HR data (in BPM)

This method was applied to the different HR data for detecting the beginning and the end of the race. For avoiding the possibility of estimate an abrupt change during the race (explained for instance by a stop for drinking and

eating), both $\widehat{\tau}_1$ (instant of the beginning) and $\widehat{\tau}_2$ were assumed to satisfied $\widehat{\tau}_1 \leq 1500$ and $n - \widehat{\tau}_2 \leq 1500$, nearly corresponding to less than 10mn after the beginning of data record and before the end of data record. The Fig. 2 exhibits an example of an application of the method to HR data.
The final race time of each athlete (that is known) can be compared to the aggregation of beats periods between $\widehat{\tau}_1$ and $\widehat{\tau}_2$. For all athletes, the difference between those two ways of measuring the same time is very often smaller than $1mn$. However, for several athletes, an important difference appears corresponding to truncated HR series or athlete forgetting to start their ECF at time.

2.2 Data smoothing

After this first step of detection of race beginning and end, a correction of aberrant data (due very often to a temporal bad contact between the athlete skin and the ECF) was needed. During exercise, the variation between two successive beats should not exceed ±10% (see for instance ([5], [26])). This can be also justified by observing the empirical distribution of HR data (see histograms in Fig. 3). Thus, for such an histogram, its form should not be spread out on both sides of ±10%. The detection of aberrant data consists to observe the increments series $(C(i))_{1\leq i<n}$ of the signal $(Y(i))_{1\leq i\leq n}$ as well as the decrements series $(C'(i))_{1<i\leq n}$, with:

$$C(i) = \frac{Y(i+1) - Y(i)}{Y(i)}, \qquad C'(i) = \frac{Y(i-1) - Y(i)}{Y(i)},$$

and to find the observation of which the relative increments exceed ±10%. For example (see Fig. 3) for a HR series of 26380 observations, it was found 32 observations which have to be corrected (186 was found for another HR series with 27077 observations).

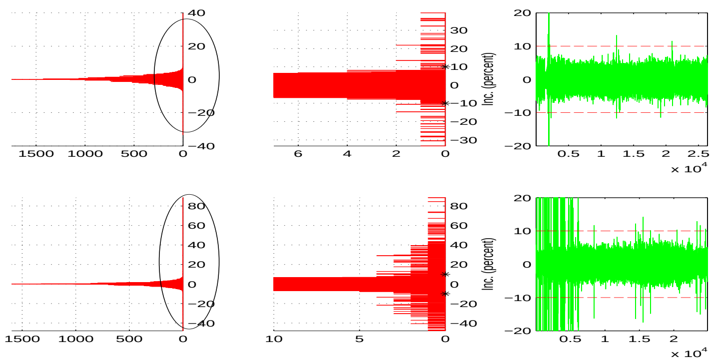

Fig. 3. Plot of the increments of observed HR series for Ath1 (top) and Ath2 (bottom)

For "cleaning" HR data, abnormal observations have to be replaced by suitable others. For determining these new values, various procedures were applied. First, an exponential smoothing is applied. It consists in replacing an abnormal observation by a linear combination of all the past observations affected by decreasing weights. However, this method is not able to correct every abnormal observation and there remain always some increments of frequencies which exceed ±10% (see Fig. 4).

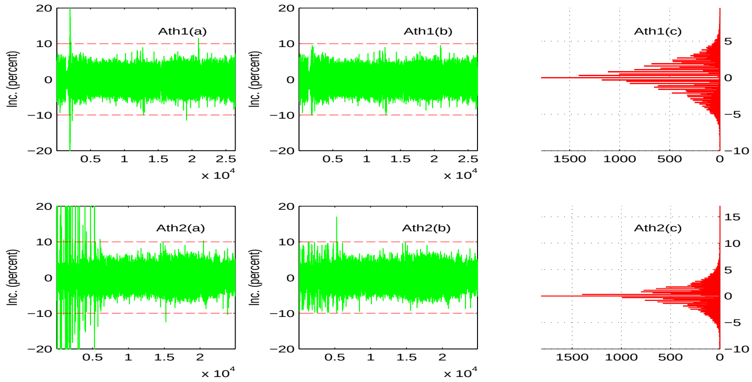

Fig. 4. (a) Increments in HR time series after exponential smoothing (b) Increments in HR time series after Kalman smoothing (c) Histogram of increments after processing

For improving these results, a recursive method was considered: the Kalman smoothing (see for instance [22]). It is a problem of smoothing on a fixed time interval where one seeks to calculate the optimal approximation of a series value knowing the observations in the selected interval. This problem like the filtering and forecast ones is solved recursively [19].

Contrary to a simple exponential smoothing, the Kalman smoothing, applied in this example on an interval of 260 observations, presents a clear improvement of the results. Often, only one iteration of the Kalman smoothing was needed to correct the whole signal (in the sense that increments do not exceed ±10%). In the other cases, the procedure is repeated for different selected intervals. For example (see Fig. 4), one iteration was needed to correct 31 observations and a second one for the remaining observation. In a second HR series, after the first iteration, there remain 10 observations to be corrected and it was done after 6 other iterations.

Detection of the different stages of a race

It is also interesting to distinguish the different stages during the race in order to unveil if a change of behavior was happened. These stages can be detected using the previous method of change points detection (see Fig. 5). The procedure is exactly the same except that the number of changes is unknown and can be also estimated. Thus, a new contrast V is built by adding to the previous contrast U an increasing function depending on the change number K, *i.e.* more precisely,

$$\widehat{V}(\tau_1, \ldots, \tau_{K-1}, K) = \widehat{G}(\tau_1, \ldots, \tau_{K-1}) + \beta \times \mathrm{pen}(K),$$

with $\beta > 0$. As a consequence, by minimizing V in $\tau_1, \ldots, \tau_{K-1}, K$, an estimator $\widehat{K}$ is obtained which varies with the penalization parameter β.
For HR data, the choice of pen(K) was K. Let $\widehat{G}_K = \widehat{G}(\widehat{\tau}_1, \ldots, \widehat{\tau}_{K-1})$, for $K = K_1, \ldots, K_{MAX}$ we define

$$\beta_i = \frac{\widehat{G}_{K_i} - \widehat{G}_{K_{i+1}}}{K_{i+1} - K_i} \text{ and } l_i = \beta_i - \beta_{i+1} \text{ with } i \geq 1.$$

Then the retained K is the greatest value of K_i such that $l_i >> l_j$ for $j > i$. Applied to the whole set HR data, the number of abrupt changes is estimated at 4 or 3. Three phases were selected to be studied, which are located in the beginning of the race, in the middle and in the end. However for certain recorded signals the first or the last phase can not be distinguished probably for measurements reasons.

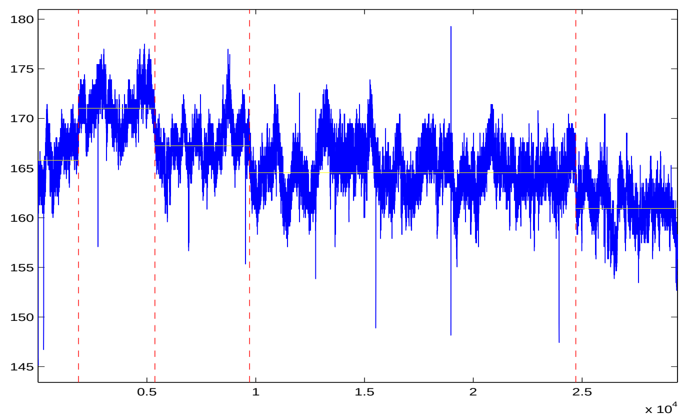

Fig. 5. The estimated configuration of changes in a HR time series of an athlete

3 HR data modelling with a long range dependent process and estimation of the Hurst parameter

In this section, a first model is proposed for modelling HR data. After a statistic study showing a badness-of-fit of this model to the data, a more suitable model is defined. Then, using a wavelet based procedure, some physiologic conclusions can be obtained from HR data.

When we observe entire or partial (during the three phases) HR time series, we remark that it exhibits a certain persistence and the related correlations decays very slowly with time what characterizes trajectories of a long memory Gaussian noise. Moreover, the aggregated signals (see for example Fig. 9) present a certain regularity very close to that of fractional Brownian motion simulated trajectories with a parameter close to 1 (Fig. 7). So, one first model which could correspond to our data is the fractional Gaussian noise.
The following Figure 6 presents a comparison between the graphs of HR data during a stage (detected previously) and a fractional Gaussian noise (FGN in the sequel) with parameter $H = 0.99$ (see the definition above). Before using statistical tools for testing the similarities of both these graphs, let us remind some elements concerning the FGN.

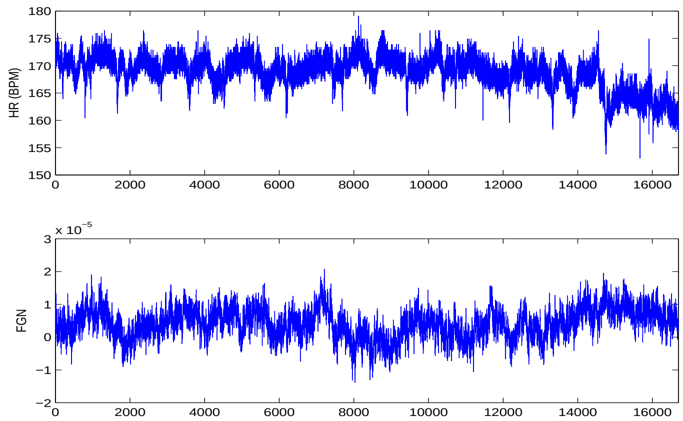

Fig. 6. Comparison of HR data in the middle of race (Ath4) and generated FGN(H=0.99) trajectories

The FGN is one of the most famous example of stationary long range dependent (LRD in the sequel) process. The LRD phenomenon was observed in many fields including telecommunication, hydrology, biomechanic, economy... A stationary second order process $Y = \{Y(k),\, k \in \mathbb{N}\}$ is said to be a LRD process if:

$$\sum_{k\in\mathbb{N}} |r_Y(k)| = \infty \quad \text{with} \quad r_Y(k) = \mathbb{E}\Big[Y(0)Y(k)\Big].$$

Thus $Y(k)$ is depending on $Y(0)$ even if k is a very large lag. Another way for writing the LRD property is the following:

$$r_Y(k) \sim k^{2H-2}L(k) \text{ , as } k \to \infty,$$

with $L(k)$ a slowly varying function (*i.e.* $\forall t > 0$, $L(xt)/L(x) \to 1$ when $x \to \infty$) and the Hurst parameter $H \in (\frac{1}{2}, 1)$.
The LRD is closely related to the self-similarity concept. A process $X = \{X(t),\, t \geq 0\}$ is so called a self-similar process with self-similarity exponent

H, if $\forall c > 0$:

$$\Big(X(ct)\Big)_t \stackrel{\mathcal{L}}{=} c^H \Big(X(t)\Big)_t.$$

Now, if we consider the aggregated process $\{X(t),\, t \geq 0\}$ defined by $X(k) = \sum_{i=1}^k Y(i)$ with Y a LRD process, then under weak conditions (for instance Y is a Gaussian or a causal process), it can be proved that, roughly speaking, for $k \to \infty$, the law of $\{X(t), t \geq k\}$ is a self-similar law (see Doukhan *et al.*, 2003, for more details).

The FGN is an example of a LRD Gaussian process. More precisely, $Y^H = \{Y^H(k),\, k \in \mathbb{N}\}$ is a FGN,

$$r_{Y^H}(k) = \frac{\sigma^2}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}) \quad \forall k \in \mathbb{N},$$

with $H \in (0,1)$ and $\sigma^2 > 0$ (it can be proved that such a Gaussian time series exists, *i.e.* all covariance matrix of any vector is a Toeplitz positive definite matrix, see for instance more details in Samorodnitsky and Taqqu, 1994). As a consequence, for $H \in (\frac{1}{2}, 1)$, a usual Taylor formula implies

$$r_{Y^H}(k) \sim \sigma^2 H(2H-1)k^{2H-2} \text{ , when } k \to \infty.$$

For a zero-mean FGN, the corresponding aggregated process, denoted here X^H, is so-called the fractional Brownian motion (FBM) and X^H is a self-similar Gaussian process with self-similar parameter H and therefore satisfies,

$$\mathrm{Var}(X^H(k)) = \sigma^2 |k|^{2H} \quad \forall k \in \mathbb{N}$$

(it can be even proved that X^H is the only Gaussian self-similar process with stationary increments). It is obvious that $Y^H(k) = X^H(k) - X^H(k-1)$, the sequence of the increments of a FBM, is a FGN.

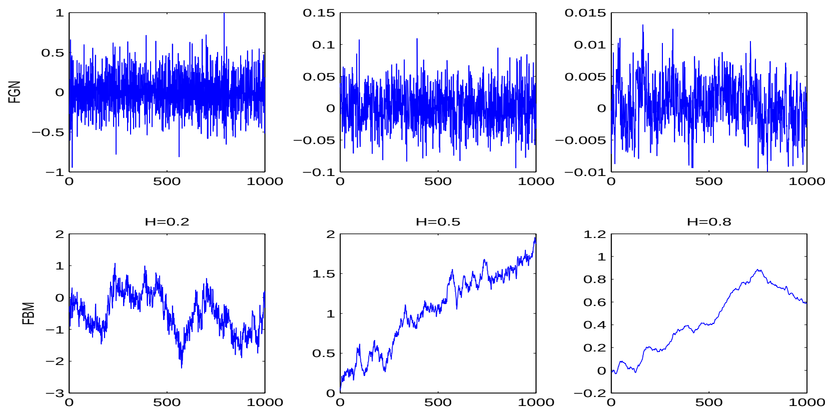

Fig. 7. Generated FGN trajectories and corresponding aggregated series (FBM) for $H = 0.2 < 0.5$ anti-persistant noise (left), $H = 0.5$ white noise (center) and $H = 0.8 > 0.5$ LRD process (right)

Several generated trajectories of FGN and corresponding FBM are presented in Fig. 7 for different values of H.

For testing if a HR path can be suitably model by a FGN, a first step consists in estimating H. Here we chose to use two estimators (but there exist many else, see for instance Doukhan *et al.*, 2003) that are known to be unchanged to the presence of a possible trend.

3.2 *Two estimators of the Hurst parameter: DFA and wavelet based estimators*

For estimating H, a frequently used method in the case of physiological data processing is the Detrended Fluctuation Analysis (DFA). The DFA method was introduced by Peng *et al.* [23]. The DFA method is a version for trended time series of the method of aggregated variance used for long-memory stationary process. It consists briefly on:

(1) Aggregated the process and divided it into windows with fixed length,
(2) Detrended the process from a linear regression in each windows,
(3) Computed the standard deviation of the residual errors (the DFA function) for all data,
(4) Estimated the coefficient of the power law from a log-log regression of the DFA function on the length of the chosen window (see Fig. 8).

After the first stage, the process is supposed to behave like a self-similar process with stationary increments added with a trend (see previously). The second stage is supposed to remove the trend. Finally, the third and fourth stages are the same than those of the aggregated method (for zero-mean stationary process). An example of the DFA method applied to a path of a FGN with different values of H is shown in Fig. 8.

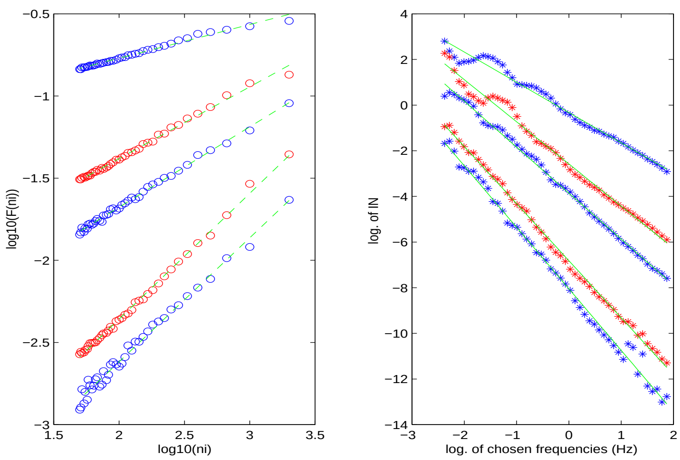

Fig. 8. Results of the DFA method and wavelet analysis applied to a path of a discretized FGN for different values of $H = 0.2, 0.4, 0.5, 0.7, 0.8$, with $N = 10000$

In [10], the asymptotic properties of the DFA function in case of a FGN path $(Y(1), \ldots, Y(N))$ are studied. In such a case the estimator $\widehat{H}_{DFA}$ converges to H with a non-optimal convergence rate ($N^{1/3}$ instead of $N^{1/2}$ reached for instance by maximum likelihood estimator). An extension of these results for a general class of stationary Gaussian LRD processes is also established. In this semi-parametric frame, we showed that the estimator $\widehat{H}_{DFA}$ converges to H with an optimal convergence rate (following the minimax criteria) when an optimal length of windows is known.

The processing of experimental data, and in particular physiological data, exhibits a major problem that is the non-stationarity of the signal. Hu, Chen, Ivanov and stanley (2001) have studied different types of non-stationarity associated with examples of trends and deduced their effect on an added noise and the kind of competition who exists between this two signals. They have also explained (2002) the effects of three other types of non-stationarity, which are often encountered in real data. In [10], we proved that $\widehat{H}_{DFA}$ does not converge to H when a polynomial trend (with degree greater or equal to 1) or a piecewise constant trend is added to a LRD process: the DFA method is clearly a non robust estimation of the Hurst parameter in case of trend.

For improving this estimation at least for polynomial trended LRD process, a wavelet based estimator is now considered. This method has been introduced by Flandrin (1992) and was developed by Abry *et al.* (2002) and Bardet *et al.* (2000). In Wesfreid *et al.* (2005), a multifractal analysis of HR time series is presented for trying to unveil their scaling law behavior using the Wavelet Transform Modulus Maxima (WTMM) method.

Let $\psi : \mathbb{R} \to \mathbb{R}$ a function so-called the mother wavelet. Let $(a, b) \in \mathbb{R}_+^* \times \mathbb{R}$ and denote $\lambda = (a, b)$. Then define the family of functions ψ_λ by

$$\psi_\lambda(t) = \frac{1}{\sqrt{a}}\, \psi\left(\frac{t}{a} - b\right)$$

Parameters a and b are so-called the scale and the shift of the wavelet transform. Let us underline that we consider a continuous wavelet transform. Let $d_Z(a, b)$ be the wavelet coefficient of the process $Z = \{Z(t),\, t \in \mathbb{R}\}$ for the scale a and the shift b, with

$$d_Z(a, b) = \frac{1}{\sqrt{a}} \int_{\mathbb{R}} \psi(\frac{t}{a} - b) Z(t) dt =< \psi_\lambda, Z >_{L^2(\mathbb{R})} .$$

For a time series instead of a continuous time process, a Riemann sum can replace the previous integral for providing a discretized wavelet coefficient $e_Z(a, b)$. The function ψ is supposed to be a function such that it exists $M \in \mathbb{N}^*$ satisfying ,

$$\int_{\mathbb{R}} t^m \psi(t)dt = 0 \quad \text{for all } m \in \{0, 1, \dots, M\}. \tag{2}$$

Therefore, ψ has its M first vanishing moments. Note that it is not necessary to choose ψ to be a *"mother"* wavelet associated to a multiresolution analysis of $\mathbb{L}^2(\mathbb{R})$. The whole theory can be developed without resorting to this assumption. The choice of ψ is then very large.

The wavelet based method can be applied to LRD or self-similar processes for respectively estimating the Hurst or the self-similarity parameter. This method is based on the following properties: for Z a stationary LRD process or a self-similar process having stationary increments, for all $a > 0$, $(d_Z(a,b))_{b \in \mathbb{R}}$ is a zero-mean stationary process and

- If Z is a stationary LRD process,

$$\mathbb{E}(d_Z^2(a,b)) = \mathrm{Var}(d_Z(a,b)) \sim C(\psi, H)a^{2H-1} \quad \text{when } a \to \infty$$

- If Z is a self-similar process having stationary increments,

$$\mathbb{E}(d_Z^2(a,b)) = \mathrm{Var}(d_Z(a,b)) \sim K(\psi, H)a^{2H+1} \quad \text{for all } a > 0$$

with $C(\psi, H)$ and $K(\psi, H)$ two positive constants depending only on ψ and H (those results are proved in Flandrin, 1992, and Abry *et al.*, 1998). Therefore, in both these cases, the variance of wavelet coefficients is a power law of a, and a log-log regression provides an estimator of H. From a path $(Z(1), \dots, Z(N))$, the estimator will be deduced from the log-log regression of the "natural" sample variance of discretized wavelet coefficients, *i.e.*,

$$S_N(a) = \frac{1}{[N/a]} \sum_{i=1}^{[N/a]} e_Z^2(a,i). \tag{3}$$

A graph $(\log a_i, \log S_N(a_i))_{1 \le i \le \ell}$ is drawn from a priori family of scales and the slope of the least square regression line provides the estimator $\widehat{H}_{WAV}$ of H. For a FGN (respectively a FBM), Bardet *et al.* (2000) (respectively Bardet, 2002) proved that the $\widehat{H}_{WAV}$ converges to H with a non-optimal convergence rate ($N^{1/3}$ instead of $N^{1/2}$ reached for instance by maximum likelihood estimator). In the semi-parametric frame of a general class of stationary Gaussian LRD processes, it was established by Moulines *et al.* (2006) that the estimator $\widehat{H}_{WAV}$ converges to H with an optimal convergence rate (following the minimax criteria) when an optimal length of windows is known. The theoretical asymptotic behavior of $\widehat{H}_{DFA}$ and $\widehat{H}_{WAV}$ are thus comparable for a Gaussian LRD process.

This is not true any more when a polynomial trended LRD (or self-similar) processes is considered. Indeed, Abry *et al.* (1998) remarked that every degree

M polynomial trend is without effects on $\widehat{H}_{WAV}$ since ψ has its M first vanishing moments. Therefore, the larger M, the more robust $\widehat{H}_{WAV}$ is.

Finally, Bardet (2002) established a Khi-squared goodness-of-fit test for a path of FBM (therefore for aggregated FGN) using wavelet analysis. This test is based on a (penalized) distance between the points $(\log a_i, \log S_N(a_i))_{1 \leq i \leq \ell}$ and a pseudo-generalized least square regression line (here the scales a_i are selected to behave as $N^{1/3}$).

In the Table 1 appear the different estimations of H computed from the DFA and wavelet analysis methods for 100 realizations of FGN paths with $N = 10000$. We choose for these simulations the concrete procedure of wavelet analysis developed by Abry *et al.* (2002) (a Daubechies wavelet is chosen and a Mallat's fast pyramidal algorithm is used to compute wavelet coefficients).

					p-val		$\sqrt{MSE}$	
H_{fGn}	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$	$\widehat{\sigma}_{DFA}$	$\widehat{\sigma}_{WAV}$	DFA	WAV	DFA	WAV
0.50	0.4936	0.5071	0.0263	0.0427	0.0152	0.0983	0.0187	0.0301
0.60	0.5908	0.6009	0.0290	0.0405	0.0017	0.8289	0.0204	0.0286
0.70	0.6859	0.6985	0.0317	0.0436	0.00001	0.7342	0.0223	0.0304
0.80	0.7875	0.8050	0.0326	0.0388	0.0002	0.1978	0.0230	0.0273
0.90	0.8821	0.8938	0.0366	0.0444	0.000002	0.1030	0.0258	0.0697

Table 1
Comparison of the two samples of estimations of H with 100 realizations of fGn path (N=10000) with DFA and wavelets methods

In one hand, the wavelets method appear slightly more effective than DFA method considering the p-value which is very low for the sample of the DFA estimations compared to wavelet analysis estimations. This is essentially due to the estimator bias which is more important in the case of DFA. In the other hand, if we consider the root of MSE which is the sum of the squared bias and the variance, the DFA estimator seems to be slightly more effective. Note that for FGN processes (without trend), the Whittle maximum likelihood estimator of H gives a "better" results (see for instance Taqqu *et al.*, 1999).

3.3 Application of both the estimators to HR data

Both estimators of H can also be applied to the HR time series of the 9 athletes. The following figures 9 and 10 exhibit examples of applications of both the estimation method to HR data.

For each athlete, it was first done to the whole time series, and then to the different phases of the race (as it was obtained from the detection of abrupt

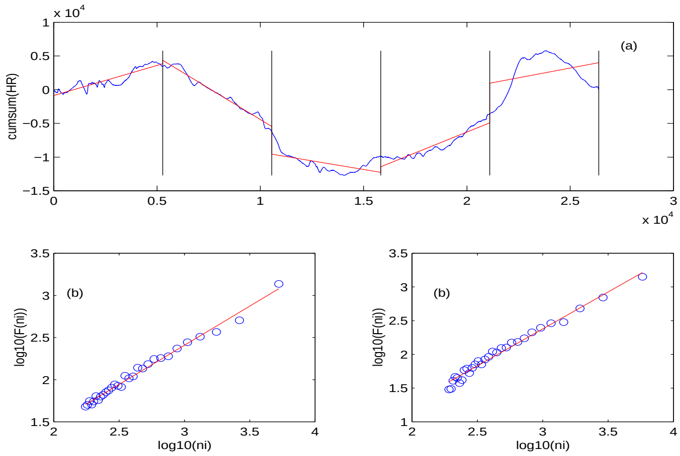

Fig. 9. Two first steps of the DFA method applied to a HR series (up) and results of the DFA method applied to HR series for two different athletes (down)

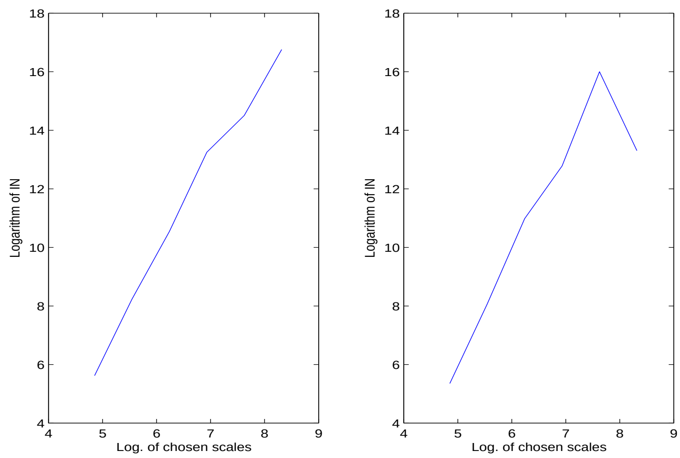

Fig. 10. The log-log graph of the variance of wavelet coefficients relating to the HR series observed during the race and in the end of race (Ath2)

changes, see Section 2). The estimation results of H, for the different signals observed during the three phases of the race, are recapitulated in the Table 2 using wavelets method and in Table 3 using DFA method.

Two main problems resort from these different estimations. First, $\widehat{H}_{DFA}$ and $\widehat{H}_{WAV}$ are often larger than 1. However, the FGN is only defined for $H \in (0,1)$. For defining a process allowing $H > 1$, three main assumptions of FGN have to be changed:

(1) the assumption that the process is a stationary process;
(2) the assumption that the process is a Gaussian process;
(3) the assumption that only two parameters (H and σ^2) are sufficient to define the process.

		Phases		
	HR series	Beginning	Middle	Race end
Ath1	0.8931	1.1268	1.1064	1.2773
Ath2	1.1174	0.7871	1.0916	0.8472
Ath3	1.0208	1.0315	1.1797	-
Ath4	0.9273	-	1.0407	0.7925
Ath5	1.0986	1.3110	1.0113	1.3952
Ath6	1.0769	1.5020	1.1597	1.3673
Ath7	1.0654	1.4237	1.1766	1.0151
Ath8	0.9568	1.6600	0.9699	1.1948
Ath9	0.9379	1.5791	0.9877	0.7263

Table 2
Estimated H with wavelets methods for HR series of different athletes

In the sequel (see above), a new model is proposed. Both the first assumptions are still satisfied and the third one is replaced by a semi-parametric assumption.

The second problem is implied by the results of the goodness-of-fit test (for wavelet analysis method). Indeed, this test is never accepted as well for the whole time series as for the partial times series. An explanation of such a phenomenon can be deduced from Figure 10: for the wavelet analysis, the points $(\log a_i, \log S_N(a_i))_{1\leq i\leq \ell}$ are clearly lined for $a_i \leq a_m$, but not exactly lined for $a_i \geq a_m$. Thus the HR time series seems to nearly behave like a FGN for "small" scales (or high frequencies), but not for "large" scales (or small frequencies). A process following this conclusion can not be the better fit of HR time series...

Remark: this last conclusion leads also to a clear advantage of wavelet based over DFA estimator. Indeed, the DFA algorithm measures only one exponent characterizing the entire signal. Then, this method corresponds rather to the study of "monofractal" signals such as FGN. At the contrary, the wavelet method provides the graph $(\log a_i, \log S_N(a_i))_{1\leq i\leq \ell}$ which can be very interesting for analyze the "fractal" behavior of data (see also Billat *et al.*, 2005).

3.4 A second model: a locally fractional Gaussian noise

In Bardet and Bertrand (2007), a generalization of the FBM, so-called the (M_K)-multiscale FBM, was introduced. The (M_0)-FBM is a FBM with self-similarity parameter H_0. Roughly speaking, the (M_K)-FBM has the same harmonizable representation (and therefore quite the same behavior as the FBM) than a FBM with self-similarity parameter H_i for frequencies $|\xi| \in$

$[\omega_i, \omega_{i+1}[$ for all $i = 0, \ldots, K$ ($K \in \mathbb{N}$). For instance, a (M_1)-FBM behaves as a FBM with self-similarity parameter H_0 for small frequencies and as a FBM with self-similarity parameter H_1 for high frequencies. Such a model was fruitfully used for modeling biomechanical signals (position of the center of pressure on a force platform during quiet postural stance measured at a frequency of 100 Hz for the one minute period).

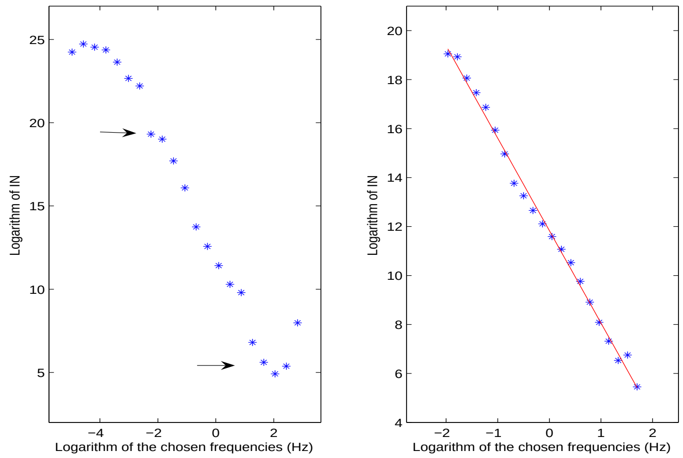

Fig. 11. The log-log graph of the variance of wavelet coefficients relating to the HR series observed during the arrival phase (Ath6) with a frequency band of [0.01 12](right) and of [0.2 4](left).

Here, Fig. 11 suggests than a fitted model for aggregated HR data should behave like a FBM with self-similarity parameter H for low frequencies and differently for high frequencies (and not necessary like a FBM). Thus define a locally fractional Brownian motion $X_\rho = \{X_\rho(t), t \in \mathbb{R}\}$ as the process such that:

$$X_\rho(t) = \int_{\mathbb{R}} \frac{e^{it\xi} - 1}{\rho(\xi)} \widehat{W}(d\xi)$$

where the function $\rho : \mathbb{R} \to [0, \infty)$ is an even continuous function such that:

- $\rho(\xi) = \frac{1}{\sigma} |\xi|^{H+1/2}$ for $|\xi| \in [\omega_0, \omega_1]$ with $H \in \mathbb{R}$, $\sigma > 0$ and $0 < \omega_0 < \omega_1$
- $\int_{\mathbb{R}} \left(1 \wedge |\xi|^2\right) \frac{1}{\rho^2(\xi)} \, d\xi < \infty.$

and $W(d\xi)$ is a Brownian measure and $\widehat{W}(d\xi)$ its Fourier transform in the distribution meaning. Cramér and Leadbetter (1967) proved the existence of such Gaussian process with stationary increments. The main advantages of such process compared to usual FBM are the following:

(1) X_ρ "behaves" like a FBM only for local band of frequencies;
(2) In this band, the parameter H is not restricted to be in $(0, 1)$: it is in $\mathbb{R}$.

From this definition, one deduces a possible model for HR data:

$$Y_\rho(t) = X_\rho(t+1) - X_\rho(t) = 2 \cdot \mathcal{R}e\Big(\int_{\mathbb{R}} \frac{e^{it\xi}\sin(\xi/2)}{\rho(\xi)} \widehat{W}(d\xi) \Big) \quad \text{for } t \in \mathbb{R}.$$

Note that $Y_\rho = \{Y_\rho(t), t \in \mathbb{R}\}$ is a stationary Gaussian process and the function $2\sin(\xi/2)\rho^{-1}(\xi)$ is so-called the spectral density of Y_ρ.

Let $\Delta_N \to 0$ and $N\Delta_N \to \infty$ when $N \to \infty$. The wavelet based estimator can provide a convergent estimation of H when a path

$$(Y_\rho(\Delta_N), Y_\rho(2\Delta_N), \ldots, Y_\rho(N\Delta_N))$$

and therefore a path $(X_\rho(\Delta_N), \ldots, X_\rho(N\Delta_N))$ is observed. Indeed, consider a "mother" wavelet ψ such that $\psi: \mathbb{R} \mapsto \mathbb{R}$ is a $\mathcal{C}^\infty$ function satisfying :

- for all $s \geq 0$, $\displaystyle\int_{\mathbb{R}} |t^s \psi(t)|\, dt < \infty$;
- its Fourier transform $\widehat{\psi}(\xi)$ is an even function compactly supported on $[-\beta, -\alpha] \cup [\alpha, \beta]$ with $0 < \alpha < \beta$.

Then, using results of Bardet and Bertrand (2007), for all $a > 0$ such that $[\frac{\alpha}{a}, \frac{\beta}{a}] \subset [\omega_0, \omega_1]$, *i.e.* $a \in [\frac{\beta}{\omega_1}, \frac{\alpha}{\omega_0}]$, $(d_{X_\rho}(a,b))_{b \in \mathbb{R}}$ is a stationary Gaussian process and

$$\mathbb{E}\Big(d^2_{X_\rho}(a,.)\Big) = K(\psi, H, \sigma) \cdot a^{2H+1},$$

with $K(\psi, H, \sigma) > 0$ only depending on ψ, H and σ. However this property is checked if and only if the function ψ is chosen such that:

$$\frac{\beta}{\alpha} < \frac{\omega_1}{\omega_0}.$$

Moreover, for $a \in [\frac{\beta}{\omega_1}, \frac{\alpha}{\omega_0}]$, the sample variance $S_N(a)$ computed from a path $(X_\rho(\Delta_N), \ldots, X_\rho(N\Delta_N))$ and defined in (3) converges to $\mathbb{E}\Big(d^2_{X_\rho}(a,.)\Big)$ and satisfies a central limit theorem with convergence rate $\sqrt{N\Delta_N}$. Thus, with fixed scales $(a_1, \ldots, a_\ell) \in [\frac{\beta}{\omega_1}, \frac{\alpha}{\omega_0}]^\ell$, a log-log-regression of $(a_i, S_N(a_i))_{1 \leq i \leq \ell}$ provides an estimation of H (and a central limit theorem with convergence rate $N\Delta_N$ satisfied by $\widehat{H}_{WAV}$ can also be established). As previously, we consider also Khi-squared goodness-of-fit test based on the wavelet analysis and defined as a weighted distance between points $(\log(a_i), \log(S_N(a_i)))_{1 \leq i \leq \ell}$ and a pseudo-generalized regression line.

Remark: The main problem with these estimator and test is the localization of the suitable frequency band $[\omega_0, \omega_1]$ (ω_0 and ω_1 are assumed to be unknown parameters). A solution consists in selecting a very large band of scales and determining then graphically the "most" linear part of the set of points $(\log(a_i), \log(S_N(a_i)))_{1\leq i\leq \ell}$. Another possible way may be to compute an adaptive estimator of this band using a quadratic criterion (following a similar procedure than in Bardet and Bertrand, 2007). Here, like 9 different paths of HR data are observed, a common frequency band $[\overline{\omega_0}, \overline{\omega_1}]$ can be graphically obtained and used for whole HR data (see above).

3.5 Application to HR data

First, one considers that a HR time series $(Y(1), \dots, Y(n))$ can be written $(Y_\rho(\Delta_n), Y_\rho(2\Delta_n), \dots, Y_\rho(n\Delta_n))$, $Y_\rho = \{Y_\rho(t), t \in \mathbb{R}\}$ a process defined as previously. Secondly, the wavelet analysis is applied to the 9 (whole or partial) HR time series (the chosen "mother" wavelet is a kind of Lemarié-Meyer wavelet such that $\beta = 2\alpha$). Using first a very large band of scales for all HR time series (for example [0.01, 12] in Fig. 12), one estimation of frequency band is deduced: $[\overline{\omega_0}, \overline{\omega_1}] = [0.2, 4]$ is the chosen frequency band for the whole and partial signals.

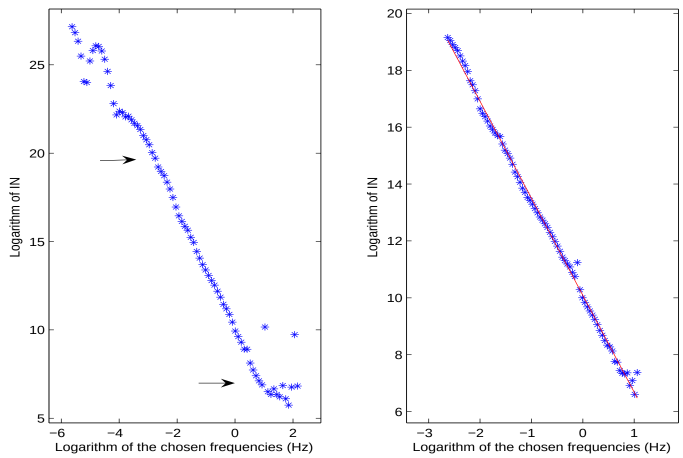

Fig. 12. The log-log graph of the variance of wavelet coefficients relating to the HR series observed in the middle of the exercise (Ath5)

The estimation results of H, for the different signals observed during the three phases of the race, are recapitulated in the Table 3.

Both DFA and wavelet analysis methods provide estimations of Hurst exponent which reflect the possible modeling of HR data with long range dependence time series.
We also note that with a p-value of 0.64, both the samples $(\widehat{H}_{DFA})_{1,\dots,9}$ and $(\widehat{H}_{WAV})_{1,\dots,9}$ obtained from all HR time series are significantly close.
The same comparison can also be done when the three characteristic stages of the race (beginning, middle and end of the race) are distinguished. The result is different. Indeed, the corresponding p-values between $(\widehat{H}_{DFA})_{1,\dots,9}$ and $(\widehat{H}_{WAV})_{1,\dots,9}$ are significatively different in the middle part of the race (and

	HR series		Race beginning		During the race		End of race	
	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$
Ath1	0.928	1.288*	1.032	1.192	1.060	1.214	0.429	1.400
Ath2	1.095	1.268*	0.905	0.973	1.126	1.108	1.240	1.452*
Ath3	1.163	1.048	0.553	0.898	1.130	1.172	-	-
Ath4	1.193	0.916*	-	-	1.098	1.249*	1.172	1.260
Ath5	1.239	1.110	1.267	1.117*	1.133	1.205	1.273	1.348*
Ath6	1.247	1.084*	1.237	1.106	1.091	1.172	1.436	1.338
Ath7	1.155	1.095	0.850	1.295	1.182	1.186*	1.129	1.209
Ath8	1.258	1.011	1.304	1.128*	0.995	1.134	1.122	1.247
Ath9	1.243	1.429*	0.820	1.019	1.127	1.535*	1.250	1.238*
p-value	0.6414		0.3723		0.0225		0.1260	
F-stat	0.23		0.85		6.38		2.65	

Table 3
Estimated $\widehat{H}$, with DFA and wavelets methods, for HR series of different athletes (*) The series for which the test is rejected. Comparison of the two samples $(\widehat{H}_{DFA})_{1,\ldots,9}$ and $(\widehat{H}_{WAV})_{1,\ldots,9}$ for whole and partial series (p-value).

relatively different in the stage of race end).

In spite of values relating to the estimator of H for all the athletes in the different phases which are relatively large, the DFA has sometimes tendency to under estimating this parameter like in the race beginning (Ath3) and the end of race (Ath1). Indeed, these value are clearly due to a certain trend supports by the fact that data points in log-log plot (Fig. 13) have not a straight line form, and we have proved in [10] that the DFA method is not robust in the case of trended long range dependent process. However in both the cases, the wavelets method is more effective since it removes sufficiently this kind of trend.

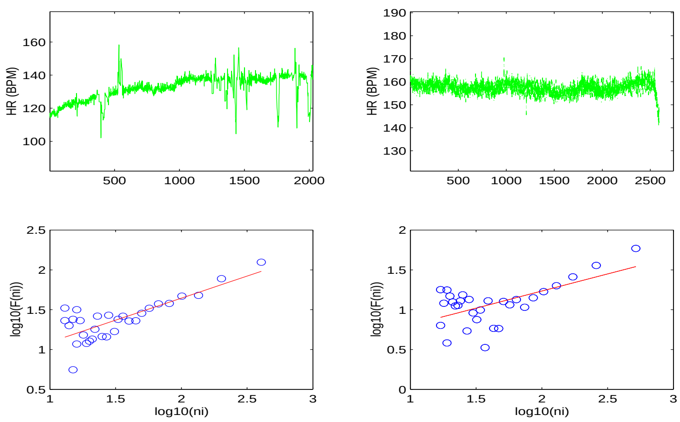

Fig. 13. The results of the DFA method applied to records for race beginning (Ath3) (left) and for end of race (Ath1) (right)

For HR data and when the goodness-of-fit test is accepted, the wavelet method shows a fractal parameter H close to 1. According to the different studies (using DFA method) about physiologic time series for distinguishing healthy from pathologic data sets (see [18], [24], [25]), an exponent $H \simeq 1$ indicate a healthy cardiac HR time series. Indeed, for the study concerning a 24 hours recorded interbeat time series during the exercise for healthy adults and heart failure adults, the following results are obtained: for healthy subjects, $H = 1.01 \pm 0.16$, for the group of heart failure subjects $H = 1.24 \pm 0.22$.

During the different stages of the marathon race, a small increase of the fractal parameter H is observed especially at the end of races. This behavior and this evolution may be associated with fatigue appearing during the last phase of the marathon. This evolution can not be observed with DFA method. Indeed, in one hand, when we observe the three 9-samples of wavelet estimators (related to the 3 phases of the race), the p-value (see Fig. 14) indicates a significantly difference due precisely to this evolution of the fractal parameter. On the other hand, a large p-value (0.85) is obtained for the same test using DFA estimation.

	$\widehat{H}_{DFA}$	$\widehat{H}_{WAV}$
p-value	0.8570	0.0158
F-stat	0.16	5.27

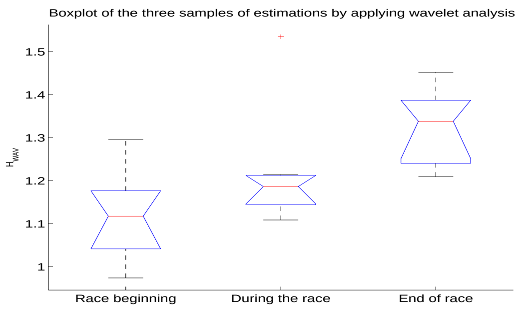

Fig. 14. Comparison of the three samples constituting by estimations in the beginning of race, during the race and then in the race end by the DFA and wavelet methods

The representation given by Fig. 14, highlight a difference in the behaviors of HR series in the beginning of the race and in the end of race. Indeed, the dispersions in the first and last sample are more important than in the middle of race and it seems that each athlete starts and finished the race at his own rhythm but in the middle athletes seems to have the same rate.

4 Conclusion

As indicated in the beginning of the last section, our main goal is to see whether the heart rate time series during the race have specific properties that of scaling law behavior. The wavelet analysis and the DFA methods are applied to 9 HR time series during the whole and also the different three phases of the race (beginning, middle and end of race) obtained by an automatic procedure. Even if their results are not exactly the same, both methods provide Hurst exponents which reflect the possible modeling of HR data by a LRD time series. However, in [10], even if the DFA estimator of Hurst parameter is proved to be convergent with a reasonable convergence rate for LRD stationary Gaussian processes, it is not at all a robust method in case of trend. The wavelet based method provides a more precise and robust estimator of the Hurst parameter. Thus, the results obtained from this wavelet estimator seem to be more valid.

Moreover, a Khi-squared goodness-of-fit test can also be deduced from this method. It seems to show that a classical LRD stationary Gaussian process is not exactly a suitable model for HR data. Graphs obtained with wavelet analysis also show that a locally fractional Gaussian noise, a semi-parametric process defined in Section 3 could be more relevant for modelling these data. A Khi-squared test confirms the goodness-of-fit of such a model. Thus, using the wavelet estimation of a fractal parameter in a specific frequency band, one obtains a conclusion relatively close to those obtained by other studies (conclusion which can not be detected with DFA method): these fractal parameters increase through the race phases, what may be explained with fatigue appearing during the last phase of the marathon. Thus this fractal parameter may be a relevant factor to detect a change during a long-distance race.

Finally, for the 9 athletes and as the test is validated with significance level around 0.65, we can estimate $\widehat{H}_{beginning}$ at 1.1, the $\widehat{H}_{middle}$ at 1.2 and $\widehat{H}_{end}$ at 1.3 with a larger confidence interval at the beginning and the end of the race. This behavior could bring a new way of understanding what is happening during a race.

References

[1] P. Abry, D. Veitch and P. Flandrin, *Long-range dependent: revisiting aggregation with wavelets,* J. Time Ser. Anal., 253-266, 1998.

[2] Abry P., Flandrin P., Taqqu M., Veitch D., *Self-similarity and long-range dependance through the wavelets lens*, Theory and applications of Long-Range Dependance, 527-556, Birkhäuser, 2003.

[3] Abry P., Flandrin P., Taqqu M., Veitch D., *Wavelets for the analysis, estimation and synthesis of scaling data*, Self-similar Network Traffic and Performance Evaluation (K. Park and W. Willinger, eds.), 39-87, Wiley, New York, 2000.

[4] Absil P.A., Sepulchre R., Bilge A., Gérard P., *Nonlinear analysis of cardiac rhythm fluctuations using DFA method*, J. Physica A : Statistical mechanics and its applications, 235-244, 1999.

[5] Bagger M., Petersen P.H., Pedersen P.K., *Biological Variation in variables associated with exercise training*, Int. J. Sports Med, 433-440, 2003.

[6] Bai J. *Least squares estimation of a shift in linear processes*, J. of Time Series Anal. 5, 453-472, 1998.

[7] Bai J. and Perron P. *Estimating and testing linear models with multiple structural changes*, Econometrica 66, 47-78, 1998.

[8] Bardet J.M. *Statistical study of the wavelet analysis of fractional Brownian motion*, IEEE Trans. Inform. Theory. 48, 991-999, 2002.

[9] Bardet J.M., Bertrand P., *Identification of the multiscale fractional Brownian motion with biomechanical applications*, Journal of Time Series Analysis, 1-52, 2007.

[10] Bardet J.M., Kammoun I., *Asymptotic Properties of the Detrended Fluctuation Analysis of Long Range Dependence Processes*, Preprint SAMOS, 2006.

[11] Bardet J.M., Lang G., Moulines E. and Soulier P. *Wavelet estimator of long-range dependent processes*, Statistical Inference for Stochastic Processes 3, 85-99, 2000.

[12] Bardet J.M., Lang G., Oppenheim G., Philippe A. and Taqqu M.S., *Generators of long-range dependent processes : A survey.*, Long-range Dependence : Theory and Applications, Birkhauser, 2002.

[13] Beran J., *Statistics for long memory processes*, Monographs on Statist. and Appl. Probab. 61. Chapman & Hall, 315, 1994.

[14] Billat V., Wesfreid E., Cottin F., Kapfer C., Koralztein J.P., *Fractal analysis of speed and physiological oscillations in long- and middle-distance running: Effect of training*, International Journal of Computer Science in Sport, 16-30, 2003. http://www.billat.net/.

[15] Brockwell P.J., Davis R.A., *Time series: Theory and methods*, Springer Series in Statistics, Second edition, 1990.

[16] Cramér H., Leadbetter M.R., *Stationary and related stochastic processes. Sample function properties and their applications*, Wiley and Sons, 1967.

[17] Flandrin P., *Wavelet analysis and synthesis of fractional Brownian motion*, IEEE Trans. on Inform. Theory 38, 910-917, 1992.

[18] Goldberger A.L., Amaral L.A.N., Hausdorff J.M., Ivanov P.C., Peng C.K., Stanley H.E., *Fractal Dynamics in Physiology: Alterations with Disease and Aging*, PNAS Vol. 99, No. 35, 2466-2472, 2002.

[19] Gourieroux C., Monfort A., *Sries temporelles et modles dynamiques*, Ed. Economica, 1995.

[20] Lavielle M., *Detection of multiple changes in a sequence of random variables*, Stoch. Process Appl, 79-102, 1999. http://www.math.u-psud.fr/ lavielle/.

[21] Lavielle M., Moulines E., *Least-squares estimation of an unknown number of shifts in a time series*, J. of Time Series Anal. 21, 33-59, 2000.

[22] Lemoine M., Pelgrin F., *Introduction aux modèles espace-état et au filtre de Kalman*, Revue OFCE, 2003.

[23] Peng C.K., Buldyrev S.V., Havlin S., Simons M., Stanley H.E., Goldberger A.L., *Mosaic organization of DNA nucleotides*, Physical Review E, Vol. 49, 1685-1689, 1994.

[24] Peng C.K., Havlin S., Stanley H.E., Goldberger A.L., *Quantification of scaling exponents and crossover phenomena in nonstationary heartbeat time series*, Chaos 5, 82, 1995.

[25] Peng C.K., Mietus J., Hausdorff J., Havlin S., Stanley H.E., Goldberger A.L. *Long-Range Anticorrelations and Non-Gaussian Behavior of the Heartbeat*, Phys. Rev. Lett. 70, 1343-1346, 1993.

[26] Pikkujms S.M., Mkikallio T.H., Airaksinen K.E., Huikuri H.V. *Determinants and interindividual variation of R-R interval dynamics in healthy middle-aged subjects.* Am. J. Physiol. Heart Circ. Physiol. 280(3): H1400-H1406, 2001.

[27] Samorodnitsky, G. and Taqqu M.S., *Stable non-Gaussian Random Processes*, Chapman and Hall, 1994.

[28] Stoev S., Taqqu M.S., Park C., Marron J.S., *On the wavelet spectrum diagnostic for Hurst parameter estimation in the analysis of Internet traffic*, Elsevier, Vol. 49, 423-445, 2005.

[29] Taqqu M.S., Teverovsky V., *Robustness of whittle type estimators for time series with long range dependence*, J. Stochastic Model, 13, 723-757, 1997.

[30] Taqqu M.S., Teverovsky V., Willinger W., *Estimators for long-range dependence: an empirical study*, Fractals 3, 785-798, 1995.

[31] Wesfreid E., Billat V., Meyer Y.,*Multifractal analysis of heartbeat time series in human races*, Elsevier Appl. Comput. Harmon. Anal. 18, 329-335, 2005.