Centrum voor Wiskunde en Informatica
REPORT*RAPPORT*

Wavelet Transform in Similarity Paradigm I

Z.R. Struzik, A.P.J.M. Siebes

Information Systems (INS)

INS-R9802 January 31, 1998

Report INS-R9802
ISSN 1386-3681

CWI
P.O. Box 94079
1090 GB Amsterdam
The Netherlands

Wavelet Transform in Similarity Paradigm I

Zbigniew R. Struzik, Arno Siebes

CWI
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
email: Zbigniew.Struzik@cwi.nl

ABSTRACT

Searching for similarity in time series finds still broader applications in data mining. However, due to the very broad spectrum of data involved, there is no possibility of defining one single notion of similarity suitable to serve all applications. We present a powerful framework based on wavelet decomposition, which allows designing and implementing a variety of criteria for the evaluation of similarity between time series. As an example, two main classes of similarity measures are considered. One is the global, statistical similarity which uses the wavelet transform derived Hurst exponent to classify time series according to their global scaling properties. The second measure estimates similarity locally using the scale-position bifurcation representation derived from the wavelet transform modulus maxima representation of the time series. A variety of generic or custom designed matching criteria can be incorporated into the detail similarity measure. We demonstrate the ability of the technique to deal with the presence of scaling, translation and polynomial bias and we also test sensitivity to the addition of random noise. Other criteria can be designed and this flexibility can be built into the data mining system to allow for specific user requirements.

1991 Mathematics Subject Classification: 28A80, 65U05, 68T10, 68P10
1991 Computing Reviews Classification System: H1, I5, Jm, J2, E2
Keywords and Phrases: wavelet transform, global similarity, local similarity.
Note: This work has been carried out under the Impact project.

1. Introduction

Many data mining algorithms exist for (more or less) standard, relational data, see, e.g. [1]. However, in practice there is much non-relational data. The most important example is time series data. For example, banks have standard data on their clients, e.g. their names, where they live et cetera, but also a time series giving the status of their account over time.

To use existing data mining technology on such data means that the time series data has to be reduced to a fixed number of characteristics. A very simple idea would be to use the current status of the account as an extra field in the table. However, if we are going to use the data for credit scoring, the current status of the account is likely to miss out on important information. For example, two clients A and B could have both \$10.000 in their account now, which is the normal status for A, whereas it is a one-time record for B. In such a case, the credit rating would be (much) higher for A than for B.

In other words, the behaviour of a time series over time is among the important characteristics of that time series. This means that we have to represent the behaviour of a time series with a finite number of characteristics. Of course, this representation should be such that two time series which show similar behaviour should be close to one another in the representation space, and vice versa.

Crucial in this statement is, of course, what similarity actualy means. The precise meaning of similarity is strongly dependent on the intended usage of the representation. Sometimes the trend of the series is the important factor in determining similarity, whereas in other cases it is everything but

the trend. Sometimes it is global (statistical) behaviour which is important, whereas in other cases it is highly localised behaviour.

It is, therefore, not possible to define one specific measure of similarity between time-series that is useful under all circumstances. Rather, a flexible toolbox, in which the user can indicate what is important in this specific case, is necessary. In this paper, we introduce a framework for such similarity measures based on wavelets. Moreover, we study the two extremes in this spectrum of possibilities in depth. The framework for similarity as developed in this paper is based on the fractal analysis of time-series [2, 3, 4] [1].

The topic of the similarity of time-series for data mining is not new. Important papers in this area are [5] and [6]. As an aside, note that these papers have other motives in showing why this topic is important. The most important difference between [5] and [6] on the one hand and our paper on the other is the framework. The core criterium for similarity used in [5], de facto requires a priori determining of what the time-series is and what are the outliers or noise. Only then can the actual distance, ϵ sausage criterium, work.

Earlier work by the same authors [7] suggested matching in the space of the Discrete Fourier Transform representation. However, DFT in itself provides only global information. Moreover, as is also concluded in [5], this approach fails in the presence of linear bias and is rather sensitive to local outliers.

The work reported on in [6] is based on local transformations of the time-series. Since there is a choice of the allowed set of transformations, this approach is closer to our approach. Our approach, however, does not rely on an (implicit) underlying model, and thus is not sensitive to outliers, noise, and translations of the data.

In other words, we are not building our similarity framework using a particular similarity model. Rather we utilise a flexible hierarchical representation of a time-series (up to a certain resolution). This representation in turn can be tailored to fit matching criteria required. In this approach, one can build in insensitivity to factors other than matching criteria, these often being outliers, noise, translation, scaling or polynomial bias.

We will continue with the discussion of similarity in the next section, introducing appropriate exponents for both global and local characterisation. In section 3, we will introduce the Wavelet Transform with the appropriate representations. We will elaborate on the global and local similarity measures in sections 4 and 5 respectively. In both sections examples we will given to account for the some of the most powerful abilities of the methods. Finally, a closing word will be given in section 6.

2. Global Versus Local Similarity

The global parameter that we will use for characterising (the roughness of) the time series should not change if we estimate it for the first or the second half or any arbitrary part of the time series, provided the characteristics of the time series do not change in time (stationarity) or with the length of the sample. The former requires that parameters remain stable with respect to scaling within a considerable range of scales (scaling, self-affinity). A good parameter indication of the similarity of the time series with its parts is the exponent with which one has to re-scale the height of the (sliding) window with the part of the time series in order to obtain a time series similar to the one compared. In figure 1 below, we illustrate this concept for the case of part of the time series compared to the complete time series itself.

This concept of self-affinity and the related Hurst exponent H has been developed within the domain of fractal geometry and is broadly applicable for time series from sources in both natural and computer sciences. In particular, it can be shown that the exponent $H = 0.5$ corresponds to the Brownian path (or trail) - a random process with independent increments - the integral of random noise. $H > 0.5$ is evidence of a long range positive correlation in the time series, visually effecting a time series with

[1] In this paper we refer to *time-series* rather than *signals*. However, the term 'signals' is exclusively used in the signal processing literature to which reference is made. Both mean one and the same thing - the sample (not necessarily uniform) record of data.

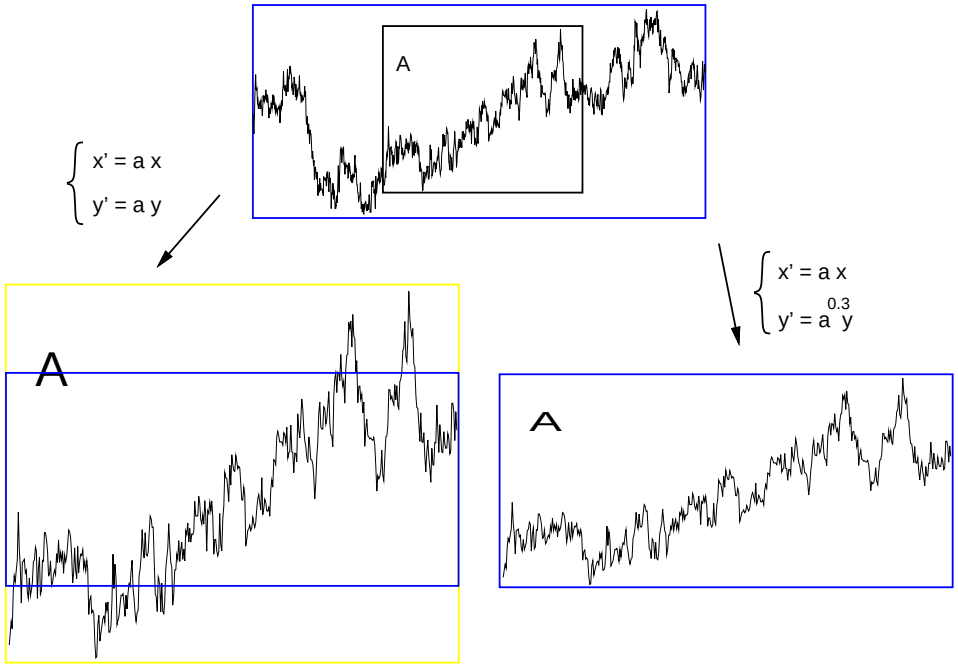

Figure 1: Shows the horizontal versus vertical rescaling argument that the exponent characterises the time series globally. *Similar* rescaling in the bottom left figure versus *affine* rescaling, bottom right, of the fractional Brownian motion of $H = 0.3$. The rescaling factor used for the affine rescaling of (x, y) axis is $(a, a^{0.3})$, while for the similar case both axis were rescaled using the a factor.

tempered jumps. On the contrary, $H < 0.5$ gives evidence of a negative correlation, a so-called anti-correlation, which is displayed by more 'wild' behaviour.

As a global measure, H can be successfully used to compare time series which are *statistically* similar, provided it is estimated in a trustworthy manner (including the removal of non-stationarities in the time series and the provision of limits on scaling range and realistic error bounds). For this purpose, we will use the wavelet transform which has been shown to be a particularly successful tool in assessing the scaling behaviour of time series [3].

Two arguments for the generalisation of the global notion of similarity suggest themselves: one is that we could allow the Hurst exponent to vary with position; the other is that we could be interested in local rather than global similarities between time series. Both require making the characteristics of the time series local in position. The relevant concept is known as the Hölder exponent h of the function in x_0 - if there exists a polynomial $P_n(x)$ of the degree n such that:

$$|f(x) - P_n(x - x_0)| \leq C|x - x_0|^h \ . \tag{2.1}$$

Then h is said to be the local Hölder exponent of the function and it characterises the scaling of the function locally for $n < h \leq n + 1$. The polynomial P_n corresponds to the Taylor series expansion of f around x_0 up to the order n.

The Wavelet Transform (which we will describe below) has been demonstrated to be a tool exceptionally well suited to the estimation of this exponent and in fact, as we will see later, global estimates like the Hurst exponent are obtained through this local Hölder exponent by means of taking an ensemble average in an appropriate partition function.

In conclusion, we have been able to identify two major approaches to identifying the similarity of

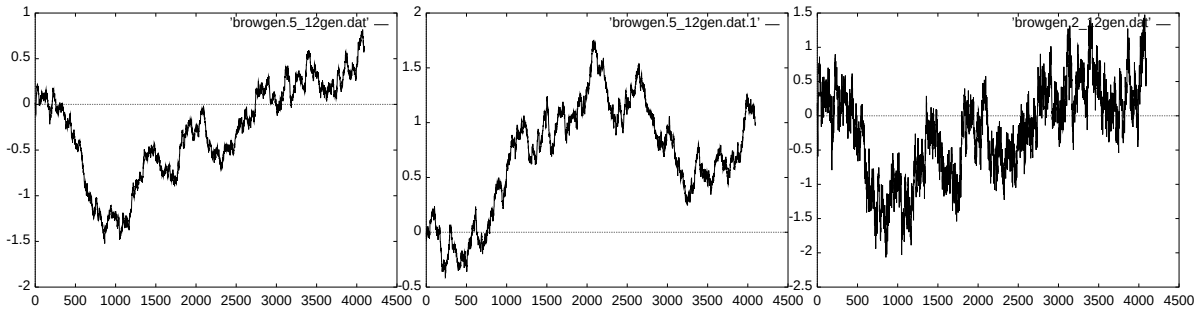

Figure 2: Example time series, $H = 0.5$ for first two, $H = 0.2$ for the most right. Only the first two time series from the left are statistically similar. On the other hand, the first and the third time series are almost identical in detail if different scaling is neglected.

time series; the global (statistical) similarity and the local, feature based similarity. In the following, we will demonstrate how to approach both classes with a common formalism based on the wavelet transform decomposition of time series.

3. Continuous Wavelet Transform and its Maxima Used to Reveal the Structure of the Time Series

As already mentioned above, the recently introduced Wavelet Transform (WT), see e.g. Ref. [8], provides a way of analysing local behaviour of functions. In this, it fundamentally differs from global transforms like the Fourier Transform. In addition to locality, it possesses the often very desirable ability of filtering the polynomial behaviour of some predefined degree.

Conceptually, the wavelet transform is a convolution product of the time series with the scaled and translated kernel - the wavelet $\psi(x)$, usually a $n-th$ derivative of a smoothing kernel $\theta(x)$. This will also be the approach taken in this work; we chose the Gaussian $\theta(x) = \exp(-x^2/2)$ as the smoothing kernel.

$$(W f)(s,b) = \frac{1}{s} \int_{-\infty}^{\infty} dx\ f(x)\ U(s,b)\ \psi(x)\ . \tag{3.1}$$

The scaling and translation actions are incorporated as the operator $U(s,b)$; the scale parameter s 'adapts' the width of the wavelet kernel to the *microscopic resolution* required, thus changing its frequency contents, and the location of the analysing wavelet is determined by the parameter b

$$U(s,b)\psi(x) = \psi(\frac{x-b}{s})\ , \tag{3.2}$$

where $s, b \in \mathbf{R}$ and $s > 0$ for the continuous version (CWT). The power given to the normalising factor s is often chosen to serve a particular purpose. Throughout this work, we will mainly use a default factor s^{-1}, which conserves the integral $\int dx\,|\psi(x)|$ and thus leaves the L^1 measure invariant. This choice allows proper definitions for the Hölder exponent. For the purpose of the decomposition of time series, more appropriate is the factor $s^{-1/2}$, which leads to an invariant L^2 measure, and thus conserves the integral square (energy).

Figure 3 shows how the wavelet transform reveals more and more detail while going towards smaller scales. The wavelet transform is sometimes referred to as the 'mathematical microscope', due to its ability to focus on weak transient frequencies and singularities in the time series. The wavelet used determines the optics of the microscope; its magnification varies with the scale factor s.

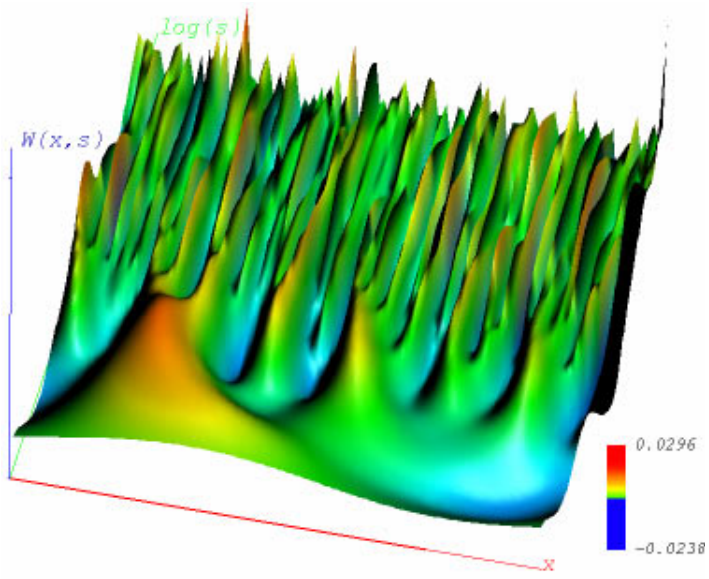

Figure 3: WT representation of the time series from figure 2 leftmost. The wavelet used is the Mexican hat.

Wavelet transformation is an *isometry*; it does not add or remove any information (in its default form - i.e. if no additional processing or restriction, e.g. on the scale range, is involved). Formally, this property can be expressed in the so-called *resolution of identity* for the inner product of the function f and g and their wavelet transforms Wf and Wg:

$$\int_0^\infty \int_{-\infty}^\infty \frac{ds\ db}{s}(Wf)(s,b)(Wg)(s,b) = C_\psi < f,g >$$

where

$$C_\psi = \int_0^\infty \frac{|\hat{\psi}(\omega)|^2}{\omega}\ .$$

Therefore, reconstruction is possible with the same wavelet $\psi^r = \psi$,

$$f(x) = C_\psi^{-1}\ \frac{1}{s^2}\ \int_{-0}^\infty \int_{-\infty}^\infty , ds db\ W f(s,b)\ \psi(\frac{x-b}{s})$$

assuming the wavelet is *admissible* i.e. $C_\psi^{-1} < \infty$. This corresponds to the requirement that the wavelet has zero mean - it is a wave function, hence *wavelet* name

$$\int_{-\infty}^\infty \psi(x)\ dx = 0\ .$$

In fact, the admissibility requirement is so weak that one can use a different wavelet for reconstruction, in particular a Dirac delta $\psi^r(x) = \delta(x)$:

$$f(x) = C_{\psi,\delta}^{-1} \frac{1}{s} \int_0^\infty ds\, W f(s,x) \ .$$

Concluding, as indicated above, the continuous wavelet transformation completely represents the function. One can therefore reconstruct the original function from its transform. According to the definition this should be done over an infinite range of scales. Restricting the range of scales from above and below can, however, be useful. The upper and lower scale component contain respectively low and high frequency components of the time series. Depending on the application, one can tune the range of reconstruction.

Let us show below, figure 4 left, the reconstruction up to large-scale/low magnification level. The low frequency detail neglected in the reconstruction is close to a constant bias; therefore the original signal and the reconstructed version are merely shifted with respect to one another. If now the small scale - high frequency components are skipped in the reconstruction, the original function can thus be reconstructed to a lesser degree of detail, as shown in figure 4 right.

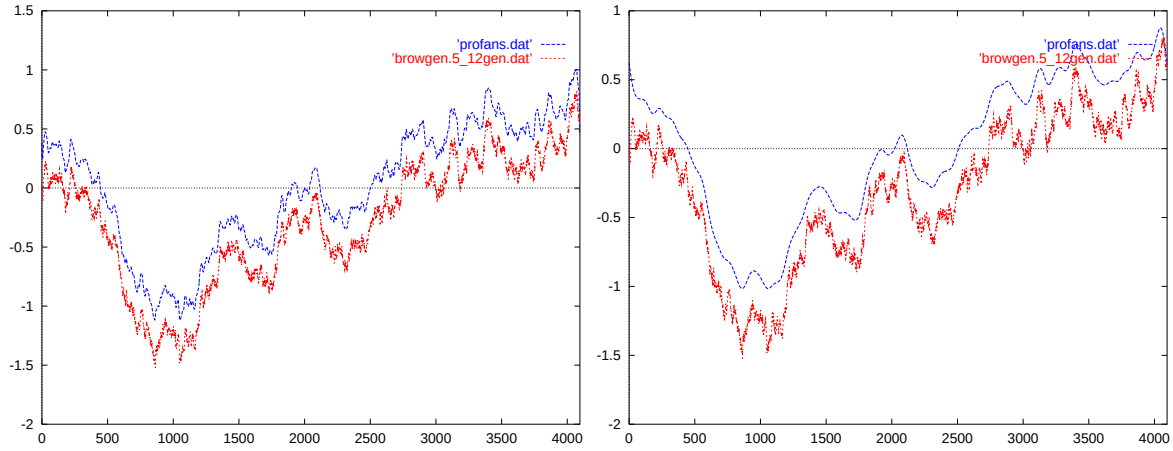

Figure 4: Left: The original and the reconstruction less low frequency bias. Right: The original and the reconstruction less low frequency bias and high frequency detail.

Let us note that quite frequently it is the singularities, the rapid changes, discontinuities and frequency transients, and not the smooth, regular behaviour which are interesting in the time series. While in the following we will largely aim at exploring this point, in the follow up to this work [12] we are going to address the issue of decomposition of time series into time-frequency atoms, preserving regular components of time series.

3.1 Accessing Singular Behaviour with the Wavelet Transformation

First, let us demonstrate the wavelet's excellent suitability to address singular aspects of the analysed time series in a *local* fashion. As already mentioned, the singularity strength is often characterised by the so-called Hölder exponent, compare Eq.2.1. If we represent the function f through its Taylor expansion around $x = x_0$:

$$f(x)_{x_0} = c_0 + c_1(x - x_0) + \cdots + c_n(x - x_0)^n + C|x - x_0|^{h(x_0)} \ .$$

It follows directly that if $h(x_0)$ is equal to a positive integer n, the function f is n times continuously differentiable in x_0. Alternatively, if $n < h(x_0) < n + 1$ the function f is continuous and singular in x_0. In that case, f is n times differentiable, but its n^{th} derivative is singular in x_0 and the exponent h

characterises this singularity. The exponent h, therefore, gives an indication of how regular the function f is in x_0, that is the higher the h, the more regular the function f.

The wavelet transform of the function f in $x = x_0$ with the wavelet of at least n vanishing moments, i.e. orthogonal to polynomials up to (maximum possible) degree n:

$$\int_{-\infty}^{+\infty} x^m \, \psi(x) \, dx = 0 \qquad \forall m, \; 0 \leq m < n \; ,$$

reduces to

$$W^{(n)} f(s, x_0) \sim C \int \psi(x) |s \; x|^{h(x_0)} \; dx \sim C \; |s|^{h(x_0)} \int \psi(x') |x'|^{h(x_0)} \; dx' \; .$$

Therefore, we have the following proportionality of the wavelet transform of the singularity $n \leq h \leq n+1$, with the wavelet with n vanishing moments:

$$W^{(n)} f(s, x_0) \sim |s|^{h(x_0)} \; .$$

Thus the continuous wavelet transform can be used for detecting and representing the singularities in the time series even if masked by the polynomial bias. This ability is inherited by the more efficient representation based on modulus maxima of CWT, which we will introduce before illustrating the ability of the WT maxima method to estimate the singularity exponent on examples.

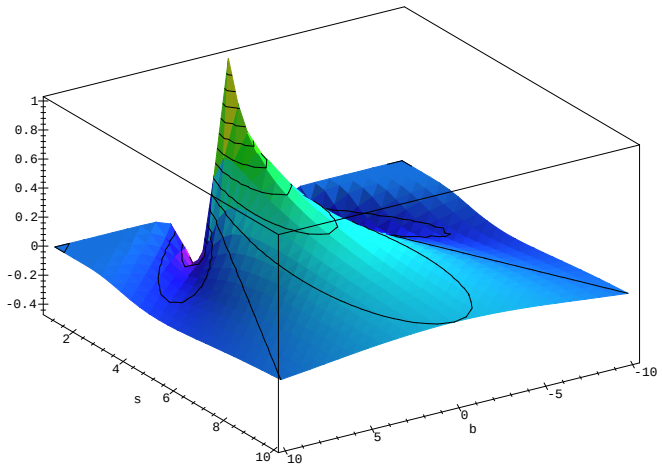

Figure 5: CWT representation of the Dirac delta $D(x_0)$, showing the cone of influence $(x_0 - s) \leq \sigma$. Due to linear scale used the local maximum line centered at $x_0 = 0$ follows $1/s$ which in log-log scale gives -1 slope. Mexican hat wavelet.

In figure 5 we show the influence of the Dirac pulse on the wavelet transform. The range of influence spreads within the entire *cone of influence*, $(x_0 - s) \leq \sigma$, which can be characterised with the standard deviation σ of the wavelet used and therefore increases linearly with the scale.

While the Hölder exponent of the singularity can be evaluated from the entire cone of influence it is much more convenient to consider the maximum of the wavelet transform only. It can be shown that such maximum converges to the singularity and that it can be used for the evaluation of the Hölder exponent of the singularity. Let us consider the following set of examples, see figure 6 left; a single

Dirac pulse at $D(1024)$, the saw tooth consisting of an integrated step function at $I(2048)$ and the (triangular) step function for $S(3072^-)$ from the right. The Hölder exponent of a Dirac pulse is -1, and each step of integration results in an increase of this exponent by 1. We, therefore, have $h = 0$ for the right sided step function $S(3072^-)$ and $h = 1$ for the integrated step $I(2048)$. In the maxima of the wavelet transform, we obtain the (logarithmic) slopes of the maxima values very closely following the correct values of these exponents, see figure 6 right.

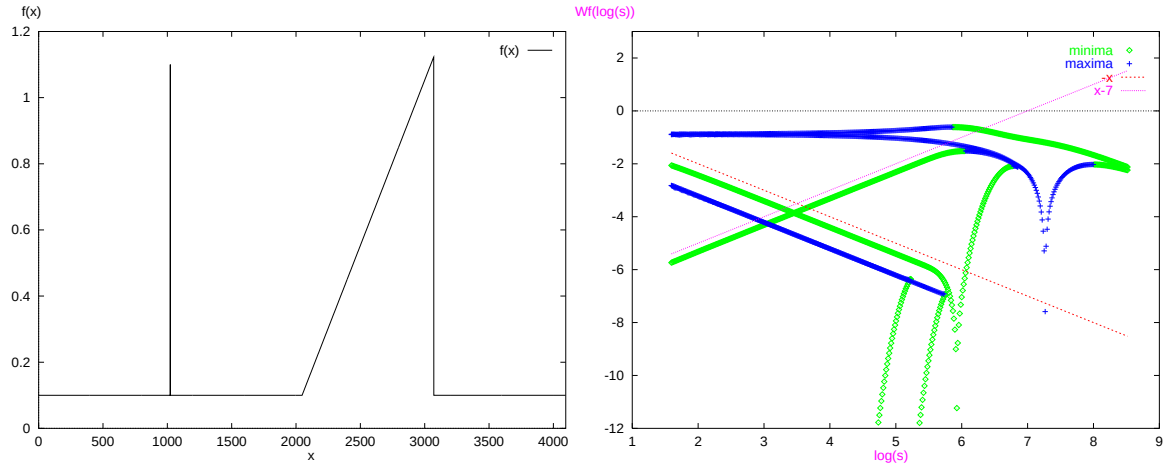

Figure 6: Left: Maxima representation of a single Dirac pulse. Right: The log-log plot of the central maximum, and of its logarithmic derivative. Indicated corresponding theoretical lines are $-x$ and -1. Mexican hat wavelet. Normalisation $1/s$.

Indeed, the slope of the maxima lines approaching the singularities reflects precisely the Hölder exponent of these singularities. This, of course, allows for the estimation of the Hölder exponent of these singularities. For Hölder singularities, the process of integration and differentiation adds and subtracts one from the exponent. This can be also verified in the results displayed.

3.2 Wavelet Transform Modulus Maxima Representation

The continuous wavelet transform described in Eq. 3.1 is an extremely redundant representation. Therefore, other, less redundant representations, are frequently used, including orthogonal representations and a variety of frames (almost orthogonal representations) [8].

For our purpose of comparison of the local features of time series, one critical requirement is the translation shift invariance of the representation; nothing other than the boundary coefficients of the representation should change, if the time series is translated by some Δt, see figure 8 for an illustration of this property.

A useful representation satisfying this requirement and of much less redundancy than the CWT is the Wavelet Transform Modulus Maxima (WTMM) representation, introduced by Mallat [9]. In the previous subsection we have also demonstrated the possibility of using the local maxima method to estimate the Hölder exponent of singularity.

Both above properties of the maxima lines representation make it particularly useful for our purpose. The WTMM is derived from the CWT representation by extracting lines of maxima of the modulus of the wavelet transform. The definition of the maxima (minima) along scale, the necessary requirement for the maximum is zero of the derivative of the WT with respect to the position coordinate x:

$$\begin{cases} & \frac{d(Wf)(s,x)}{dx} = 0 & \text{and} \\ \text{either} & \frac{d^2(Wf)(s,x)}{dx^2} < 0 & \text{for maximum} \\ \text{or} & \frac{d^2(Wf)(s,x)}{dx^2} > 0 & \text{for minimum}\,. \end{cases} \tag{3.3}$$

An additional condition for zero of the second derivative identifies the beginning of the maximum (minimum) line:

$$\begin{cases} \frac{d(Wf)(s,x)}{dx} = 0 \\ \frac{d^2(Wf)(s,x)}{dx^2} = 0\,. \end{cases} \tag{3.4}$$

An example WTMM tree is shown in figure 7, together with the highlighted bifurcations of the maxima lines [4].

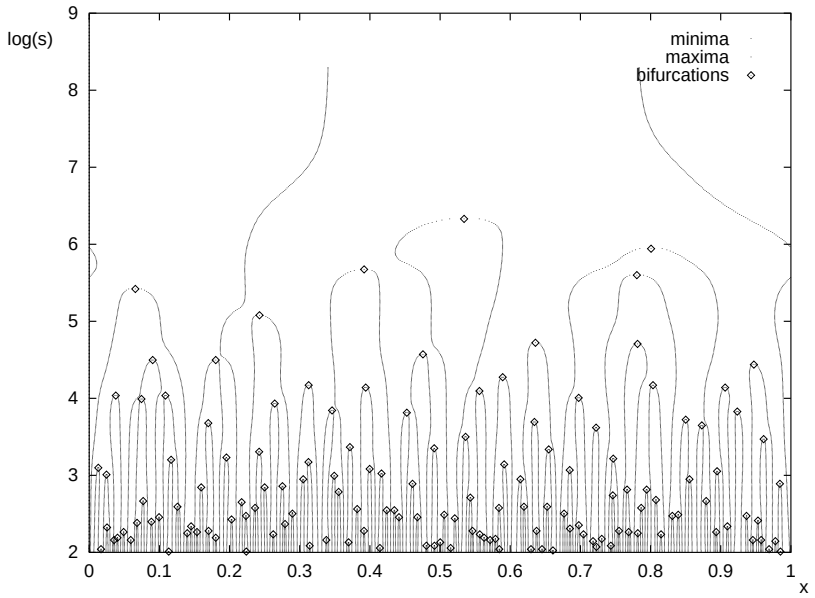

Figure 7: WTMM representation of the time series and the bifurcations of the WTMM tree. Mexican hat wavelet.

As demonstrated above, the wavelet has to be orthogonal to polynomials up to a certain degree n, in order to access the singularity exponent h by filtering out the polynomial bias. This operation of filtering the polynomial behaviour is nothing other than differentiating the time series to the degree n - the number of vanishing moments of the wavelet. This is evident from the fact that the wavelet transformation commutes with the operation of differentiation:

$$f(x)\ \psi(x,a) = f(x)\ s\ \frac{d}{dx}\theta(x,s) = s\ \frac{d}{dx}(f(x)\ \theta(x,s))\ . \tag{3.5}$$

Therefore, using wavelets with n vanishing moments one can perform stable derivation of n-th order; thus one can obtain a smoothed derivative $D^n_{(\theta(s))}$ of the time series at the given scale s. The degree of derivation can be controlled with n the number of vanishing moments. For $n = 1$, we obtain the representation corresponding to the first derivative of the function, the local slopes of the input time series.

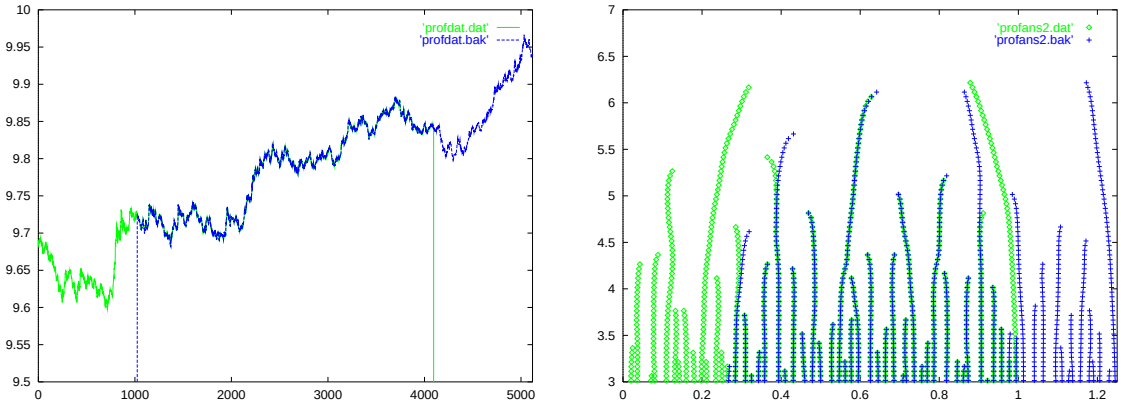

Figure 8: The translation invariance of the (WTMM): Left: - the original signal and its shifted version. Right: - the common part of the representation does not change (except for finite size effects). First derivative of the Gaussian wavelet.

$$D_{(\theta(s))}f(x) = \frac{1}{s^2}\int f(x)\ \psi(\frac{x-b}{s})\ dx \tag{3.6}$$

Note that the modulus maxima representation makes use of the maximum values of the same convolution product as the WT, compare Eq 3.1, but with the normalisation factor set to $1/s$. The maxima lines of the wavelet transform performed with the wavelet orthogonal to constants, the first derivative of a smoothing function, is therefore, up to the scaling factor s, proportional, locally in position and scale, to the strongest values of the first derivative of the analysed time series. These aspects of the decomposition will be further considered in the follow up to this work [12].

4. A Global (Statistical) Estimation of the Similarity of Time Series in the Presence of Scaling, Translation and Polynomial Bias

In the previous, section we have shown that the maxima lines of the WTMM representation of the time series are particularly useful for estimating the local scaling parameters of singularities. The link between this local characterisation and the global scaling properties of the time series has been developed by Arneodo et al [3].

In particular, one can show that the Hurst exponent is related to the q=2nd moment (correlation) of the scaling of the measure on the WTMM maxima tree:

$$s^{2(H+P_n)-1} \sim \sum_{all\ maxima\ at\ scale\ s} \mu^2(s) \tag{4.1}$$

where $\mu(s)$ is the amplitude of the maximum of the WT at the corresponding scale, and the sum - the partition function - is taken over all the maxima at the given scale s. P_n indicates the degree of the polynomial offset of the time series.

This relation (for $P_n = 1$) can easily be verified in figure 9 where, in log-log coordinates, the power law relation 4.1 should result in a straight line. We show the same, second moment for two examples of random noise and anti-correlated fractional noise. The Hurst exponent can easily be estimated from the slope of the linear fit to the scaling portion of the plot.

In figure 9 right, we show the same second moment estimated for the record of the financial index. It falls very well into the same category as the simulated Brownian path - indeed financial records are

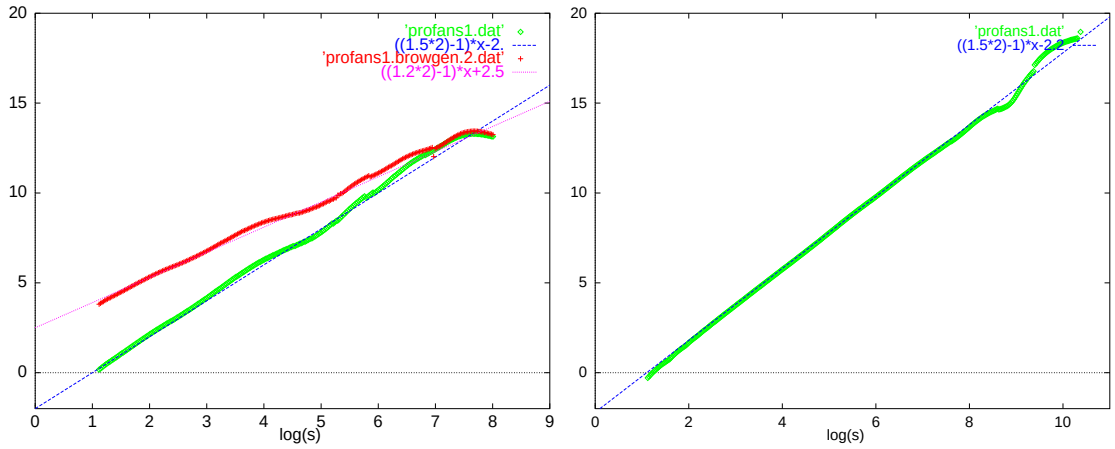

Figure 9: Left: the second moments of the modulus maxima representation of the noise samples from figure 2 leftmost and rightmost. Right: the same for the real life sample of a financial index shown in figure 17 left.

known to follow the H=0.5 law very closely, see e.g. [13]. The exponent H can thus be used to classify global similarity between time series or categorize them on the statistical grounds.

By design, H is limited to taking values from the interval $0 \leq H \leq 1$. With the WTMM based formalism, we are able to estimate not only the fractional scaling part but also the degree of the underlying polynomial Pn. For fBm trails $Pn = 1$, for the noise time series it would be $Pn = 0$. It is therefore more convenient to take the complete exponent $\beta = H + Pn$ as the (correlation) exponent representative to our time series.

With this exponent, we are able to distinguish between various categories of time series for $0 < \beta \leq 1$, $1 < \beta \leq 2$, $2 < \beta \leq 3$ or higher. As an illustration, see figure 10, where we show the scaling of the WTMM correlation dimension evaluated for random noise, its integral - Brownian trail - and, again, its integral. For each integration step, the increase of the slope of the second moment is two, and the corresponding increase of the correlation dimension is one.

Note, that WTMM based formalism will correctly estimate the correlation exponent β only if the wavelet used has enough vanishing moments. In most practical situations, this condition is satisfied with $n = 2$ or $n = 3$. For example, $n = 1$ is enough for noise time series $0 < \beta \leq 1$, like the leftmost example in figure 10 [2]. But we need a wavelet with at least $n = 2$ for the $1 < \beta \leq 2$ class (central example in figure 10) and with at least $n = 3$ for the $2 < \beta \leq 3$ class (rightmost example in figure 10). Of course, it should be noted, that the exponent β is independent of any scaling, translation or polynomial bias up to $n-1$ degree, which may be affecting the investigated time series. Still, considerable care should be taken to prevent finite size effects from distorting the estimation of the exponent.

Estimation of the exponent β from the scaling of the second moment of the partition function 4.1 picks out the most pronounced feature of the global characteristics of the time series, but is not complete. Generically, it is possible to calculate the entire range of moments of the partition function, say from minus ten to plus ten. This range will depend on the length of the available time series, limiting the number of moments to those meaningful.

Statistically we will observe two distinct behaviours. One possibility is the linear dependence of the moments, indicating that the Holder exponent in time series takes just one value. In this case, we call the time series mono-fractal. Another possibility will show the non-linear dependence of the moments. For such a time series, the Holder exponent varies from point to point, indicating the presence of multi-fractal scaling.

[2] In fact, for this example we could even use $n = 0$ since, as it happens, this noise example does not possess any constant bias. Still, generically it is a good idea to use higher rather than lower n.

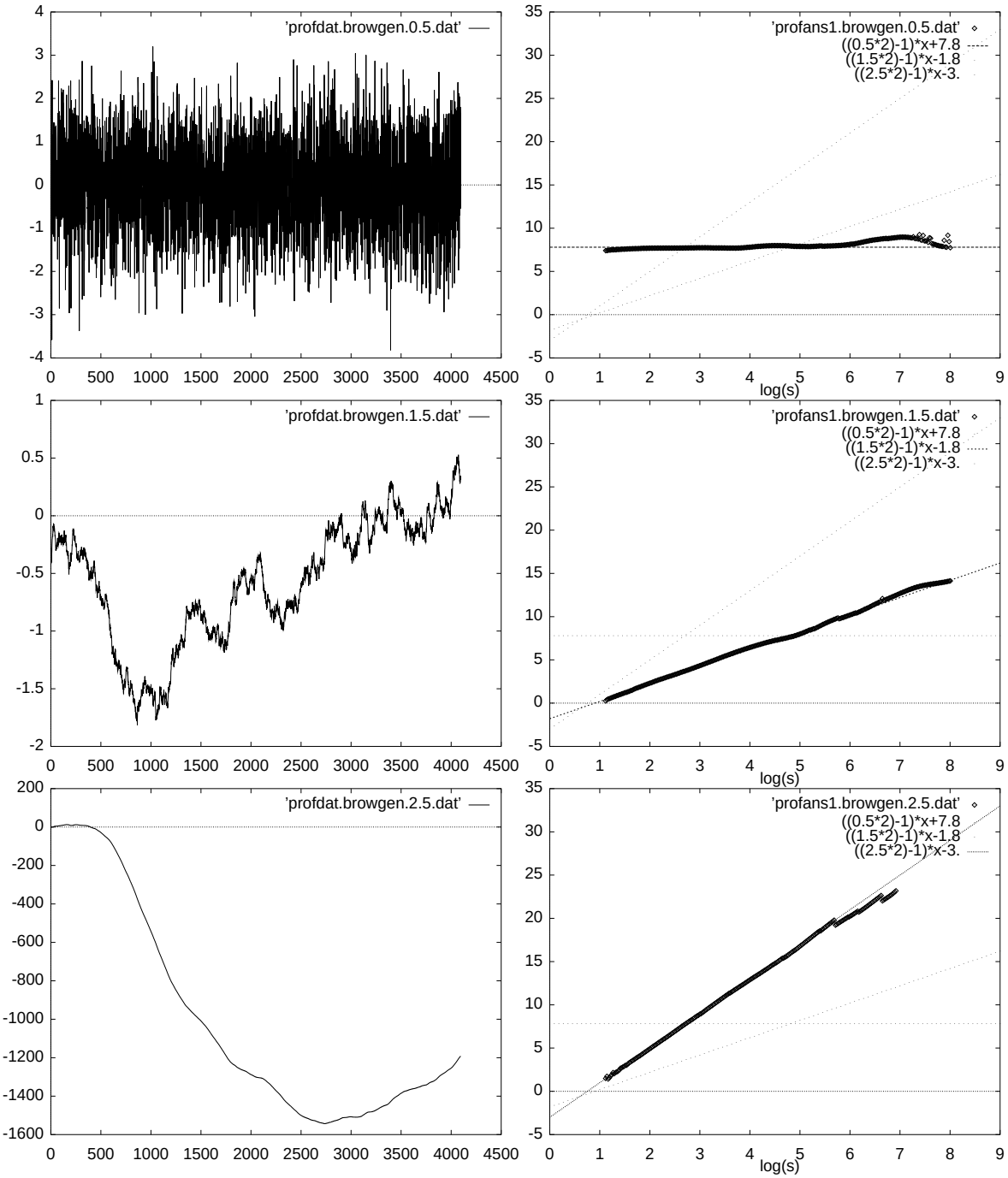

Figure 10: Left, from the top to bottom: the random noise sample, its integral - Brownian trail and again its integral. Right, from the top to bottom: corresponding scaling of the second moments of WTMM maxima.

A generally accepted approach is calculating the 'histogram' of these Hölder exponents, known as the *spectrum of singularities*, which will show either a single peak for the mono-fractal time series or a broad spectrum for a multi-fractal. The spectrum of singularities can be calculated either by taking the Legendre transformation of the exponents extracted from the partition function moments, or the so-called direct approach, see Muzy [11] for both methods.

Below we show two example spectra for mono-fractal and multi-fractal time series. In the left upper part of fig. 11, we show the record of a Brownian noise with the (theoretically) single Hurst exponent $H = 0.5$. The corresponding spectrum reveals a narrow band near the $h = -0.5$. In the right upper part of fig. 11, we show the record of a real-time time series (human heart beat) showing rare, strong events - peaks far exceeding what one would sense as consistent with the average behaviour. This phenomenon is reflected in the spectrum below which shows considerable width. It is also centered at a very small value of $h \sim 0.1$, indicating the presence of strong (long-range) correlations in the time series.

The key difference between the two classes of signals is that one can be represented statistically with one single exponent $h = H$ while the other requires a wide spectrum $f(\alpha), \alpha = h$. Nevertheless, the bulk of the singularities falls near the central value of h, and this exponent provides the often most relevant characterisation of the time series.

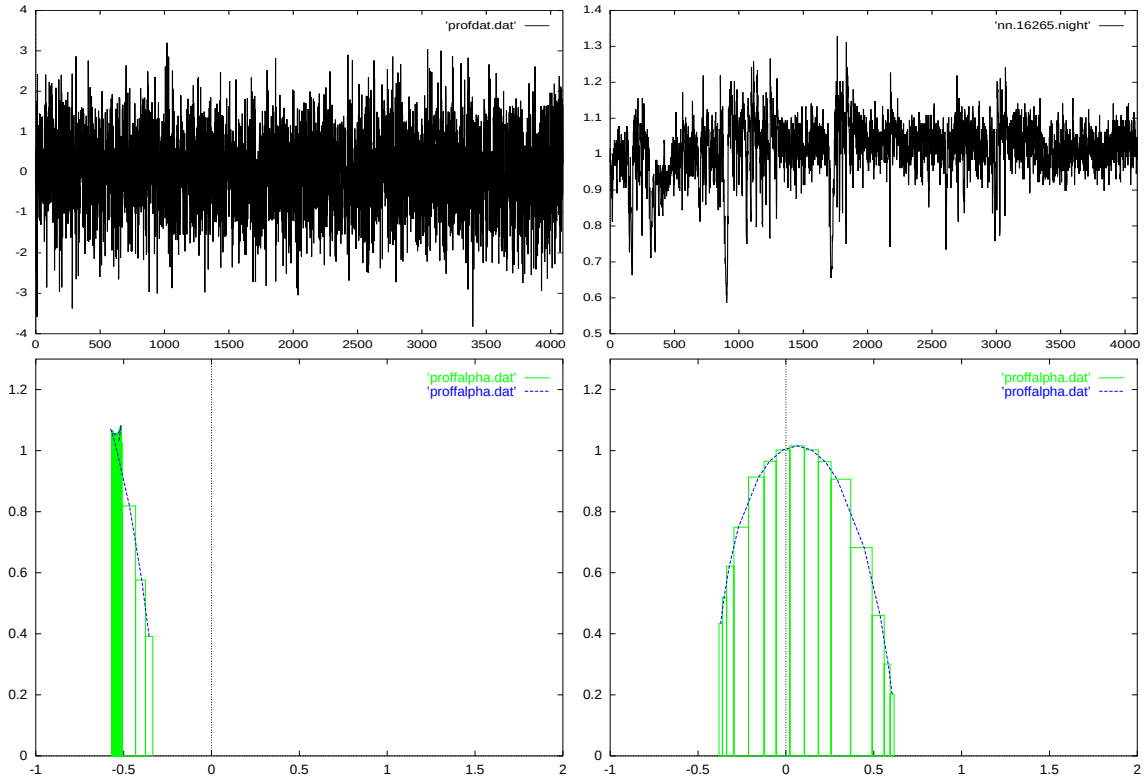

Figure 11: In the left upper plot, the record of a Brownian noise with the (theoretically) single Hurst exponent $H = 0.5$. The corresponding spectrum at the left below reveals a narrow band near the $h = -0.5$. In the right upper plot, the record of a multi-fractal time series. The corresponding wide spectrum of singularities at the right below.

5. Local Similarity Estimation in the Presence of Scaling, Translation and Polynomial Bias

In order to evaluate the local similarity, we will turn to local features of the time series, the bifurcations of the WTMM tree. Bifurcations [4] form a set of highly sensitive 'landmarks' in the WT landscape of the decomposed function. By reflecting the scale-position development of the maxima tree, they capture the singular structure of the time series. Each bifurcation can be represented with its position and scale coordinates, plus the corresponding value of the WT in the bifurcation point.

Just like the wavelet transform itself, the bifurcations can be evaluated for the time series up to a certain resolution, meaning that only the coarse features are taken into account. Alternatively a range of scales can be determined in the application to be covered by WTMM tree and its bifurcations. The numbers used for our experiments ranged from 20-100 bifurcations covering the span of maximum of two decades of scales from the highest resolution available.

We will (mainly and unless otherwise indicated) use bifurcations obtained from the WTMM tree of the wavelet transform of the time series with the Mexican hat wavelet, the second derivative of the Gaussian kernel. This means that the maxima lines follow singular features in the second derivative of the function and the bifurcation representation reflects the structure of these features in the second derivative of the function. This will allow for looking for similarities with respect to linear bias - such bias is filtered out from the time series by the wavelet with two vanishing moments. One can of course use wavelets with fewer or more vanishing moments to suit one's particular needs.

5.1 Local Distance (Similarity) Measure between Two (or More) Bifurcations

Essentially, the method uses the bifurcation representations of two time series to be compared and estimates the degree of similarity of these representations. Additionally one can shift both representations with respect to one another in order to find whether there is a better match if shift and scaling are involved. Note that a shift in the logarithmic scale corresponds with the scaling operation in the original time series.

The simplest but quite reliable measure of the similarity of two sets of bifurcations is given by the occurrence of a bifurcation in one representation, within a distance ϵ of some 'reference' bifurcation in the 'reference' set of bifurcations. Counting the fraction of reference bifurcations which have such a matching counterpart in the bifurcation set compared gives the estimate of the similarity between the two sets - a number from the range 0..1. A useful improvement of this scheme is easily made by counting only bifurcations in which WT has the same sign. This procedure, subject to one ϵ parameter only, gives good results.

A straightforward extension to the box of ϵ size is a two-dimensional correlation function which has a smooth decay of the distance between the bifurcations. As the measure of correlation between two bifurcation points: $biff_a(a, \sigma_b)$ and $biff_b(b, \sigma_b)$, we took for our experiments the auto-correlation function of two Gaussian kernels $C(biff_a, biff_b)$. This correlation can be parametrized by the additional shift in (logarithmic) scale and position Δs and Δx respectively:

$$\begin{aligned} C(biff_a, biff_b) &= C(a, b, \sigma_a, \sigma_b) = \\ &= \frac{1}{\sqrt{\sqrt{\pi}\sigma_a \sqrt{\pi}\sigma_b}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(\frac{x-a}{\sigma_a})^2} e^{-\frac{1}{2}(\frac{x-b}{\sigma_b})^2} dx = \\ &= \sqrt{2}\, \frac{\sqrt{\sigma_a \sigma_b}\, e^{-\frac{1}{2}\frac{(b-a)^2}{{\sigma_a}^2+{\sigma_b}^2}}}{\sqrt{{\sigma_a}^2+{\sigma_b}^2}} \,. \end{aligned} \tag{5.1}$$

This measure does not, however, decay quickly enough in our experience and therefore we equipped it with two adjustable parameters F_{scale}, F_{posit}, independently affecting the decay along position and scale:

$$C(biff_a, biff_b)_{(s,p)} = \sqrt{2}\ \frac{\sqrt{\sigma_a \sigma_b}\ e^{-\frac{1}{2}\frac{F_{posit}(b-a)^2}{{\sigma_a}^2+{\sigma_b}^2}}}{\sqrt{{\sigma_a}^2+{\sigma_b}^2}}\ . \tag{5.2}$$

The resulting half-width along scale, in logarithmic scale coordinates $S = \log(\sigma)$, is then:

$$1/F_{scale}\ \ln(\frac{4+\sqrt{15}}{4-\sqrt{15}})\ ,$$

and along position it is:

$$1/F_{posit}\ 2\sqrt{2}\ \sqrt{ln(2)}\ \sqrt{{\sigma_a}^2+{\sigma_b}^2}\ .$$

Note that without the factor F_{scale}, the half-width of the measure along the scale is over 4.13 in logarithmic coordinates. The effective neighbourhood within which we would like to locate another bifurcation is smaller by one order of magnitude (therefore we use $F_{scale} > 10$). Also contrary to the half-width along position, half-width along scale is independent of the location of bifurcations in position and scale.

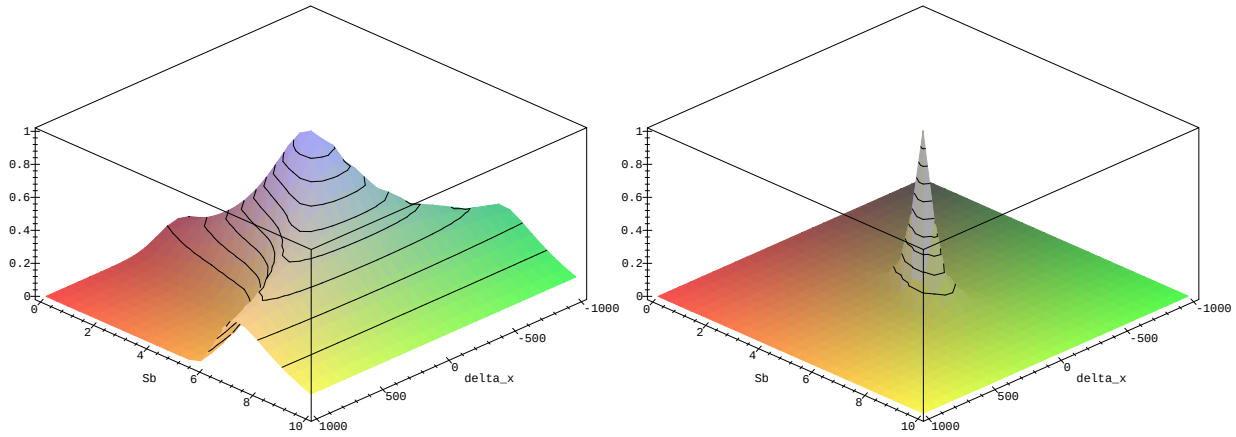

Figure 12: The measure used for matching bifurcations for two values of the parameters F_{scale} and F_{posit}. Left default values $F_{scale} = 1$ and $F_{posit} = 1$. Right $F_{scale} = 5$ and $f_{posit} = 5$.

With this measure for the correlation of two (or more) bifurcations, we can now estimate the total correlation of two bifurcation representations of the time series we want to compare. The most straightforward measure is simply

$$M(\Delta x, \Delta s) = \frac{1}{norm(N_1, N_2, \Delta x, \Delta s)} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} C(biff_i, biff_j)(\Delta x, \Delta s)\ V(biff_i, biff_j)\ , \tag{5.3}$$

where N_1 and N_2 are the respective numbers of bifurcations in both time series representations compared. $V(biff_i, biff_j)$ is an optional factor (but used in all our experiments) which weights the

correlation in the values of the wavelet transform in $WT(biff_i)$ and $WT(biff_j)$ respectively. This can be just a two valued function giving zero for respective values having different signs and one for consistent sign:

$$\begin{array}{llll} \text{if} & WT(biff_i)\,.\,WT(biff_j) > 0 & \text{then} & V(biff_i, biff_j) = 1 \\ & & \text{else} & V(biff_i, biff_j) = 0. \end{array} \tag{5.4}$$

Alternatively, as will be used in our last example, it can be a function giving -1 for WT values with opposite signs and 1 in the case of the same signs. One can of course design continuously valued correlation functions.

$$\begin{array}{llll} \text{if} & WT(biff_i)\,.\,WT(biff_j) > 0 & \text{then} & V(biff_i, biff_j) = 1 \\ \text{if} & WT(biff_i)\,.\,WT(biff_j) < 0 & \text{then} & V(biff_i, biff_j) = -1 \\ & & \text{else} & V(biff_i, biff_j) = 0. \end{array} \tag{5.5}$$

The *norm* factor is designed to normalise the measure so that it would reach 1 for two identical samples and 0 for samples with no resemblance of one another according to the criteria employed. There is also possibility of extending the measure from 0 to -1 in the case of negative or 'anti-match'. This can for example be done with the weighting factor described above, Eq. 5.5, which will be used in our last example in the following.

There are several nontrivial issues that need to be taken into account in order to design the appropriate *norm*. The first is the length of the time series, which is generally not equal. Additionally, since we work with the decomposition of the time series in scale the range of scales may differ. With complete wavelet representation these parameters are known, however, in the case of the bifurcation representation information is only approximately known (through lower bounds only).

The approach we took is to relate the *norm* factor to the number of bifurcations taking their geometric mean:

$$\frac{1}{norm_1} \sim \frac{1}{\sqrt{N_1\ N_2}} \tag{5.6}$$

This simple normalising approximates the overlap of the two sets of bifurcations. Even though it is sufficient for directly comparing two bifurcation representations, the scale and position shift changes the effective overlap of the two samples. In order to compensate for this we introduce two component normalisation factors. The first approximates the change of the effective overlap due to the scale shift, and the second does the same to compensate for the position shift. In the $norm_2$ formula these two factors look respectively like:

$$\frac{1}{norm_2} \sim \frac{1}{\exp(\Delta s)}\ \frac{1}{1 - \Delta x\ \exp(\Delta s)} \tag{5.7}$$

Note that the scale shift Δs in the formula above is in the logarithmic scale. The resulting normalisation to be used, is therefore a combination of the above factors Eq. 5.6 and Eq. 5.7.

Another aspect which generally needs normalisation is the overlap of the correlation functions Eq. 5.2 associated with respective bifurcations. As designed, they have infinite support and therefore introduce some degree of additional measure due to mutual overlapping. This overlap can be a priori evaluated and removed, however, we took the liberty of neglecting this bias in the measure. This was possible due to very quick decay which we imposed on the correlation function Eq. 5.2, and the the

application of an additional window, cutting off the tails of this correlation function. (We chose for the window the scale range spanning $|\Delta s| < 0.5$ and $|\Delta x| < 0.5\ \exp(\Delta s)$.)

Last component of the normalisation is actually more related to removing the irrelevant/unlikely part of the measure for the case of extreme scale position shift. It consists of a band of a certain width which is too short to contain any trustworthy matches. We set this arbitrarily at one quarter of the reference sample but this is of course subject to free choice. The effective width of the band changes with the scale shift to give the finite size factor $norm_3$:

$$\begin{aligned} \text{if} \quad & \frac{1-\Delta x}{\exp(\Delta s)} > 0.25 \qquad \text{then} \quad \frac{1}{norm_3} = 1 \\ & \qquad\qquad\quad \text{else} \quad \frac{1}{norm_3} = 0\ . \end{aligned} \tag{5.8}$$

Again, this factor is used to modify the total measure 5.3; $\frac{1}{norm} = \frac{1}{norm_1}\,\frac{1}{norm_2}\,\frac{1}{norm_3}$. It results in the zero measure band visible on the right side, $x \to 1$, in the measure plots in the experimental subsections to follow.

In the rest of this section we will demonstrate the ability of the method to localise similarity in time series using the measure just designed, Eq. 5.3, to compare their intricate structure - the scale-position behaviour of the second derivative of the time series. This structure is captured by the bifurcation representation obtained with the wavelet orthogonal to linear bias ($n = 2$) which, unless otherwise indicated, we will use in the examples in the rest of this paper. It is, of course, possible to use a structure of different derivative of the time series or the time series itself by taking a wavelet with an appropriate n - number of vanishing moments.

5.2 The Effect of Translation and Scaling

In this example, we demonstrate how the similarities can be found for the time series which is scaled and translated. Let us take the example time series record, see figure 13.

Using this example, we will demonstrate that the algorithm using the measure 5.3 to compare two bifurcation representations is capable of finding similarities between the time series (or their parts), with respect to the operations of translation and scaling.

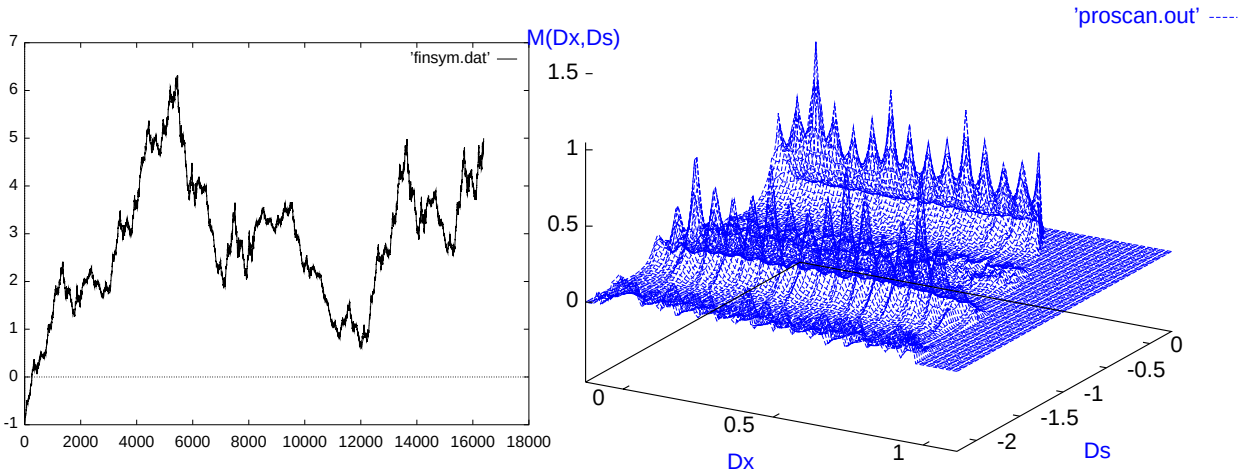

Figure 13: Example fractal time series consisting of four self-affine sub-parts (left). The measure shows high regularity and reveals the scaling and translation - elements of the invariance of the time series.

Instead of comparing one time series with another, in this test we will actually compare the time series with itself. Covering an adequate scope of values $\Delta x, \Delta s$ reveals in the measure $M(\Delta x, \Delta s)$ the presence of a very structured self-affinity relation within this time series - there are four main peaks located at $\Delta x = 0$, $\Delta x = 1/4$, $\Delta x = 1/2$ and $\Delta x = 3/4$. In between these, there are lower peaks, starting from the big peak at $\Delta x = 0$. The next smaller peak is at $\Delta x = 1/16$, followed by another at $\Delta x = 2/16$, and the next at $\Delta x = 3/16$. This sequence is repeated at $\Delta s = -1.39 = \log(1/4)$.

This, in fact, goes to show that we discovered that within the time series there are four similar parts, which in sequel contain four similar parts etc. This similarity is evaluated with respect to the second derivative of the time series - the analysing wavelet is the Mexican hat - all the masking linear trend at different scales has been removed. Indeed the test time series is an IFS fractal [14] with four non-overlapping self-affine transformations as the construction rule.

5.3 The Effect of Random Noise

Speaking of the influence of random noise has, of course, a very special meaning in the context of our analysis - most of the example test time series we considered were records of pure or correlated noise. Still such noise is a perfectly valid and valuable time series (in fact a record which in absence of an appropriate model seems to be just an uncorrelated noise is likely to contain perfectly coded information). Therefore, in addition to comparing two different noise records, it makes perfect sense to evaluate how one noise record is corrupted by other noise.

We took three examples.

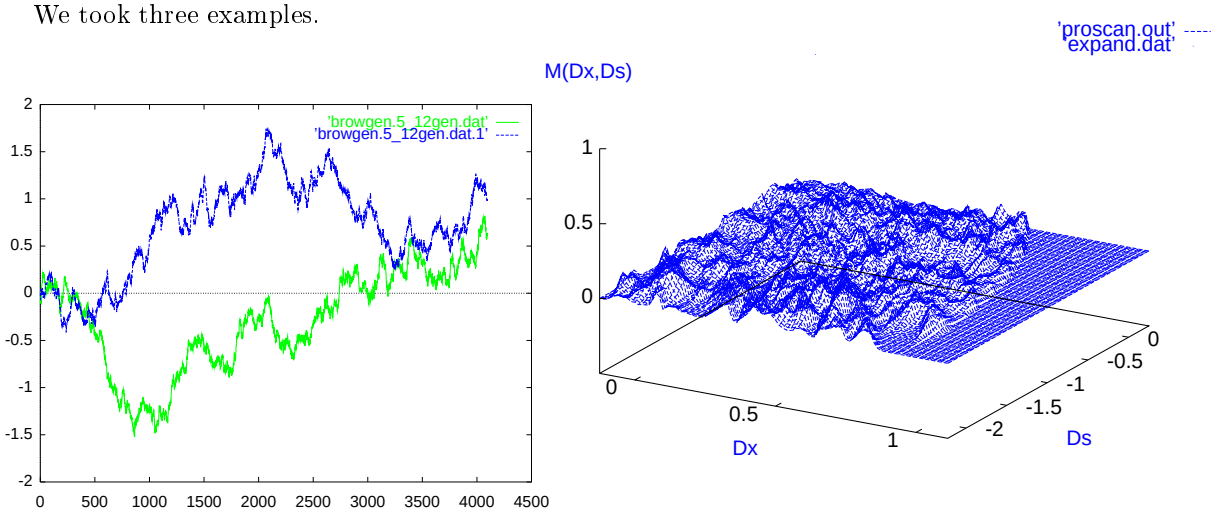

Figure 14: Two independent random walks (left) investigated for the presence of similar parts with respect to scaling translation and linear bias show only slight local similarities at a residual level of measure fluctuations (right).

In the first, figure 14, two independent random walks are scanned for similarities. The level of the measure remains low but significant within the searched range of scale and position shift. It can be considered as the fluctuations of the similarity measure reaching a significant level due to random occurrences of parts remotely looking like one another. These similarities are more likely to be assumed when going to higher Δx,Δs, due to the simple fact of considering the overlap of just a few bifurcations.

In the second example, figure 15, we consider two noises created with the same random sequence. One is uncorrelated with $H = 0.5$ and the other anti-correlated with $H = 0.2$. Both are created with the 'random midpoint displacement method', see e.g. [2], using the same random sequence, meaning the sign of displacement remains the same. Indeed, the similarity measure gives quite good a response, reaching about 0.5 at no shift.

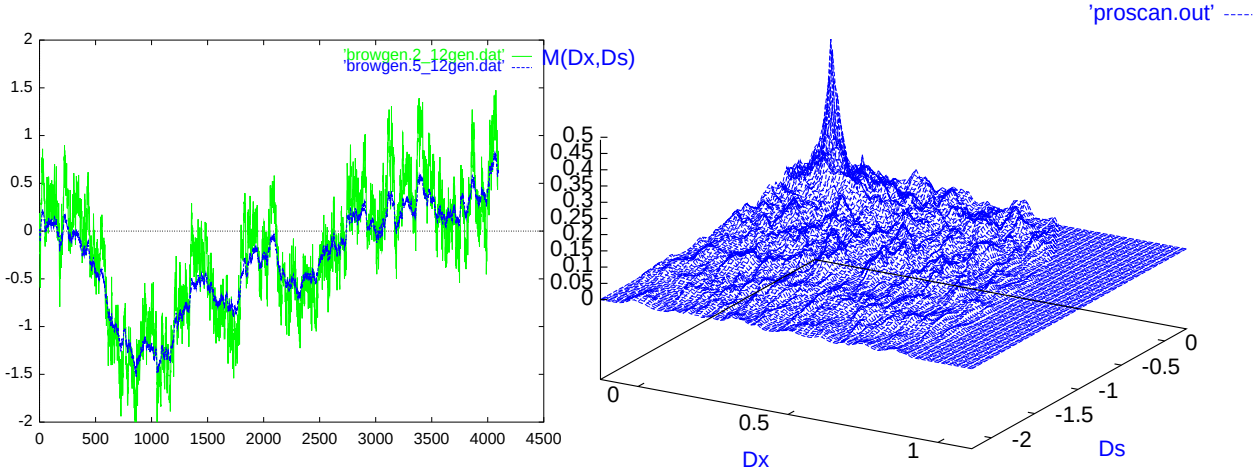

Figure 15: Two noises with $H = 0.5$ and with $H = 0.2$ created with the same random sequence. Right, the similarity measure response of about 0.5 at no shift.

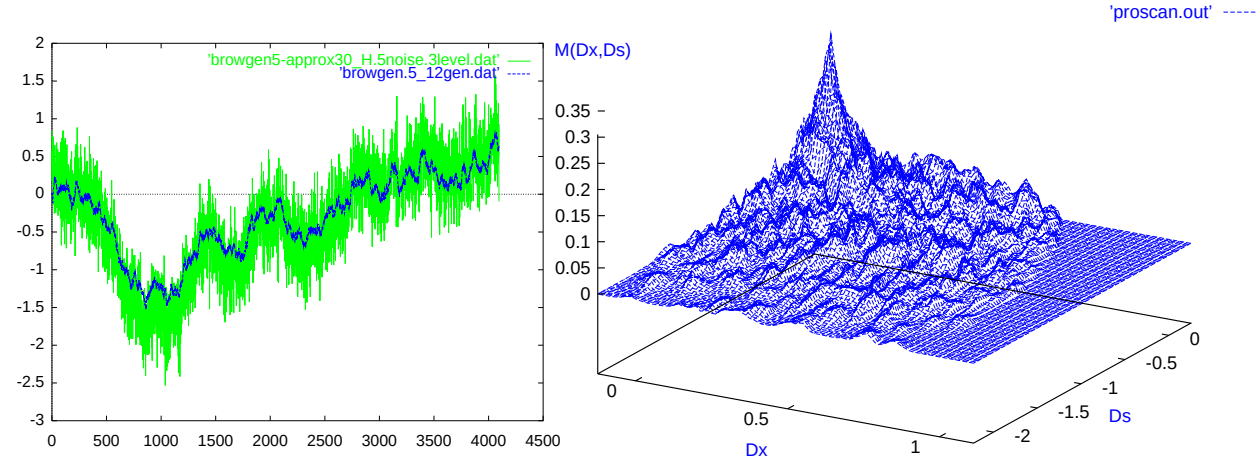

Figure 16: Example random walk with the addition of random noise at $-10dB$ level. Right the resulting measure.

The third example, figure 16, demonstrates the influence of the additive random noise at a level of about $-10dB$ (amplitude factor 0.3). The result is a drop in the measure associated with a noticeable widening of the maximum cone. The resulting measure is at a level of one third of the measure norm for two identical time series.

5.4 A Real Life Sample within a Sample Example

So far we were matching samples with equal length, this time we took two unequal samples of a real life time series. Two records of a financial index were scanned for the presence of similar parts with respect to scaling, translation and linear bias.

While from a visual inspection, it is rather difficult to establish the degree of similarity between the two, the similarity measure reveals a high degree of similarity at $\Delta x = 0.3$ and $\Delta s = -1.4$ (this corresponds to a shift by $0.3 \,.\, 4096 = 1229$ samples and rescaling by a factor $e^{-1.4} = 0.25$). At almost maximum level $C = 0.87$, it is far stronger than the rest of the measure. Applying the shift parameters to the second plot confirms a close fit, but only after the linear trend is restored! (Indeed the second

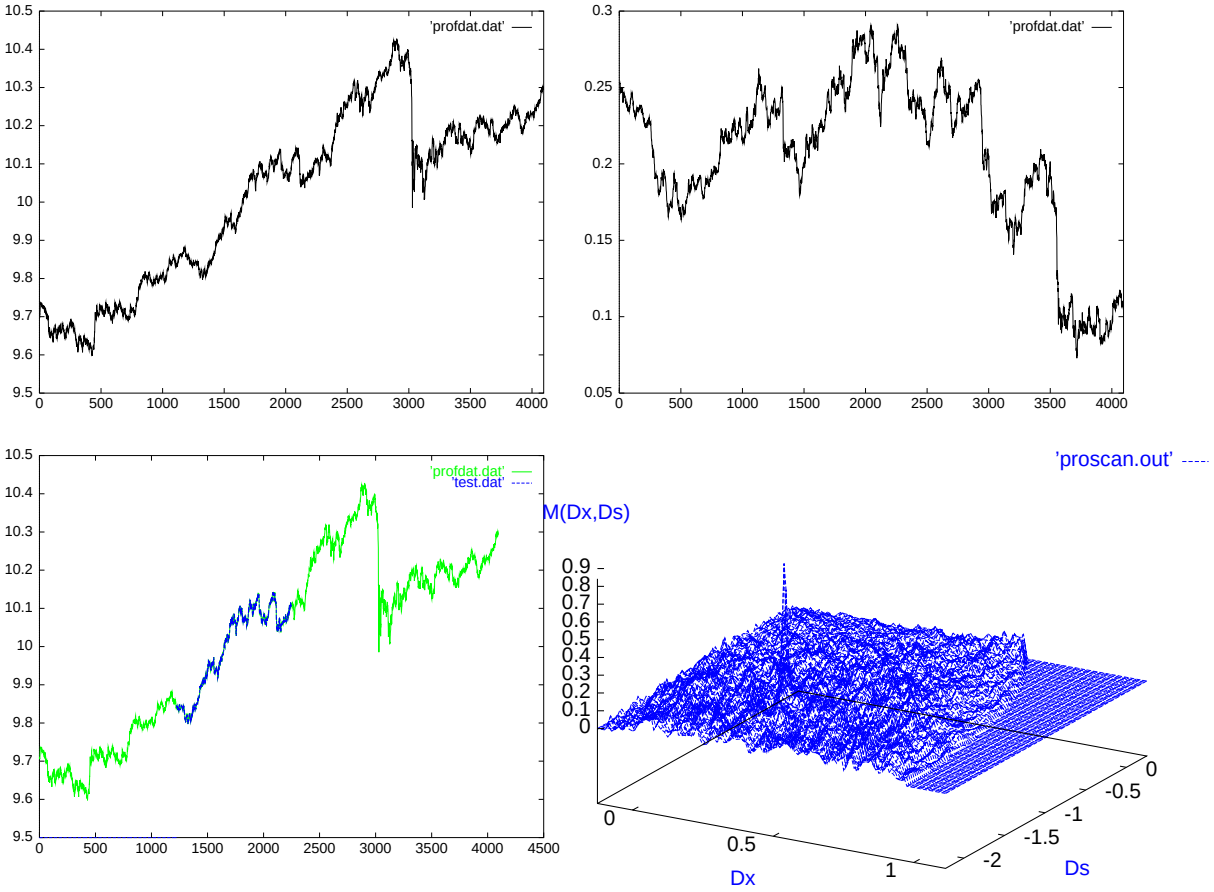

Figure 17: Above, two example real life time series subject to investigation for the presence of similar parts with respect to scaling, translation and linear bias. Left below the match revealed in the maximum of the measure shown at right below.

time series is just a part of the first with added linear bias.) We added a relatively strong bias in order to illustrate our purpose. The method will, of course, also work for smaller levels of bias for which visual check of the similarity of both time series will be possible. Still, even in such cases, techniques without the ability to filter polynomial trends are likely to fail.

5.5 Checking for slope similarity between two independent samples

This last example consists of two correlated fBm walks of unequal length and scaling exponent; $H = 0.9$ and $H = 0.7$ respectively, independent from one another. We scanned these two time series for the similarity in slope which is perhaps the most generic idea of similarity in data mining on time series. The first derivative of the smoothing kernel served for the wavelet - it is sensitive to local slope and filters out constant bias. Note that in the examples this far we used the bifurcation set obtained with the second derivative of the smoothing kernel insensitive to constants and slopes.

For another modification, in the examples so far, only the sign of the value of the wavelet transform at the bifurcation points was considered in the similarity measure. Here, we slightly modified the measure function to allow for correlation of sign of the value to reach either 1 or -1 for the case of correlated (both positive or both negative) and anti-correlated (one positive, one negative) values, Eq. 5.5. Note that such a measure is still not sensitive to exact values of local slope, although such

sensitivity can easily be achieved by introducing the appropriate continuous correlation function.

The example time series are shown in figure 18 left and right above. The scale axis is set in the direction so that the first time series is searched for the presence of parts resembling the second time series, but this can, of course, be reverted or the scale axis can be extended to cover both directions. Compared with the previous experiments we also slightly extended the width of the correlation kernel 5.2 to allow larger deviations for the bifurcation coordinates.

The result of the modified measure is shown in figure 18 right below, additionally using contours. The measure revealed one strong maximum at $(0.3125, -1.23)$ but also a strong minimum located at (0.6125,-1.39). The minimum is an indication of a good 'anti-match', which in the slope domain simply means a good match with the negated slope. In figure 18 left below, we marked the parts of the first 'reference' signal which show best slope similarity with the second 'matched' signal, with its slope unaffected and negated for the two shown matches respectively.

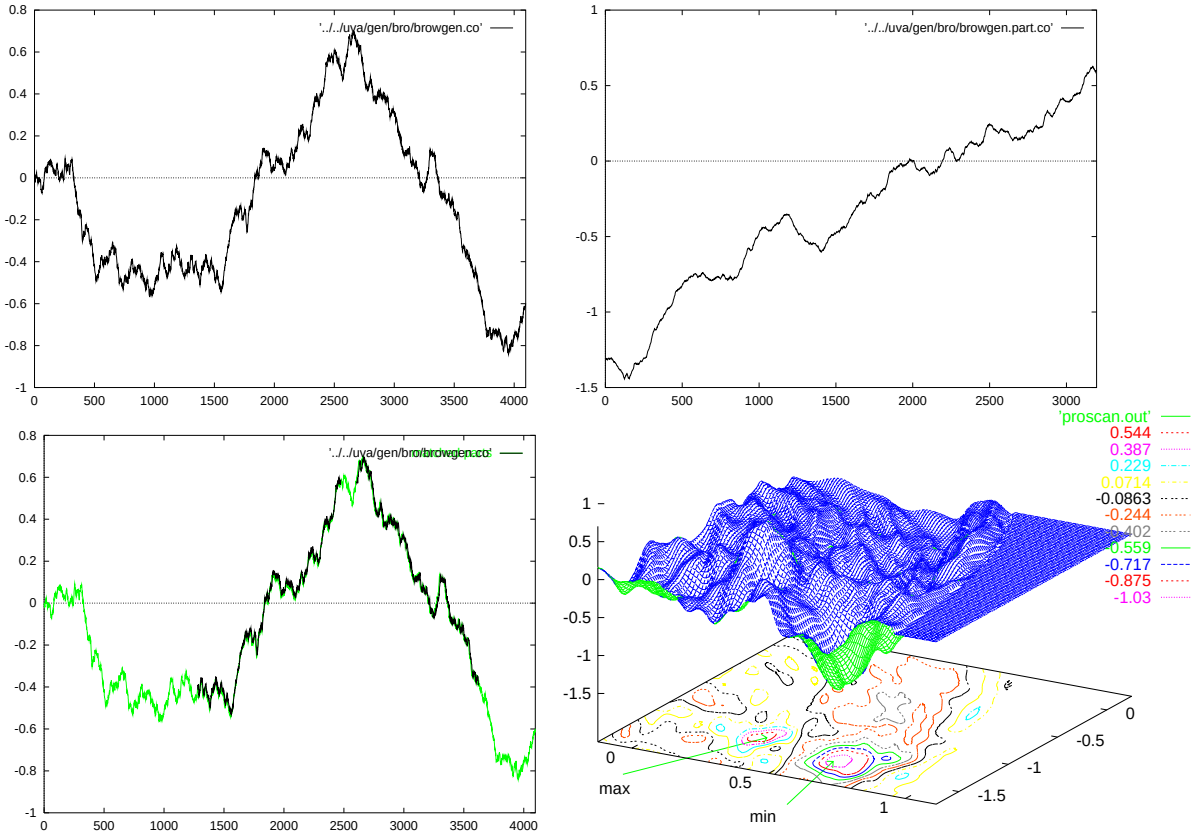

Figure 18: Above, two example fBm time series with different Hurst exponent scanned for similarity. Left has the correlation exponent $H = 0.7$ while the right one has $H = 0.9$. Left below the parts of the time series $H = 0.7$ above showing greatest similarity and anti-similarity in slope with the entire $H = 0.9$ time series. Right below the measure plus contour plot showing the location of the maximum and the minimum of the similarity measure.

6. Final Remarks and Conclusions

We presented a powerful technique allowing the evaluation of similarity between time series in the presence of scaling, translation and polynomial bias. Two main classes of similarity evaluation measures

were distinguished and the appropriate measures were proposed.
The global, statistical similarity was estimated with the Wavelet Transform derived Hurst exponent. It classifies time series according to their global scaling properties. The local, detail oriented measure, used the scale-position bifurcation representation of the wavelet modulus maxima transform of the time series. It makes it possible to obtain good matches of (the parts of) the time series with respect to scaling translation and polynomial bias. The degree of the polynomial bias filtered can be affected as well as the range of the translation and scaling parameters. The measure used for matching the two bifurcation representations of the time series can also be adapted to the specific user requirements. In addition to this, since they represent two extremes, both the global and the local measure can be used together with appropriate weighting factors.

Future work on this methodology is expected to include investigating features of the WTMM representation other than bifurcations, namely instantaneous frequencies - time-frequency 'atoms'.

Along with identifying other then presented generic matching criteria, (e.g. piece-wise linear or IFS approximation) tailoring the representations to fit specific matching criteria will be pursued.

An important aspect of the WTMM representation which has not been fully used in this work is its hierarchical structure. This can be built into the searching algorithms for increasing speed and match optimization - selecting best matching parts of the time series. The (bifurcation) tree matching algorithm presented in [4] is a good candidate for developing the time-series similarity methodology in this direction.

References

1. U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, R. Uthurusamy, Eds., *Advances in Knowledge Discovery and Data Mining*, AAAI Press/MIT Press, 1996.
2. K. Falconer, *Fractal Geometry - Mathematical Foundations and Applications*, John Wiley (1990).
3. A. Arneodo, E. Bacry, J.F. Muzy, The Thermodynamics of Fractals Revisited with Wavelets, *Physica A*, **213**, 232-275, (1995).
4. Z.R. Struzik, The Wavelet Transform in The Solution to the Inverse Fractal Problem, *Fractals* **3** No.2, 329-350 (1995).
5. R. Agrawal, K-I. Lin, h.S. Sawhney, K, Shim, Fast Similarity Search in the Presence of Noise, Scaling and Translation in Time Series Databases, In *Proceedings of the 21 VLDB Conference*, Zürich, 1995.
6. G. Das, D. Gunopulos, H. Mannila, Finding Similar Time Series, In *Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial intelligence 1263, Springer, 1997.
7. R. Agrawal, C. Faloutsos, A. Swami. Efficient Similarity Search in Sequence Databases, In *Proc. of the Fourth International Conference on Foundations of Data Organization and Algorithms*, Chicago, 1993.
8. I. Daubechies, *Ten Lectures on Wavelets*, S.I.A.M. (1992).
9. S.G. Mallat, S. Zhong, Complete Signal Representation with Multiscale Edges, *IEEE Trans. PAMI* **14**, 710-732 (1992).
10. S.G. Mallat, W.L. Hwang, Singularity Detection and Processing with Wavelets, *IEEE Trans. on Information Theory* **38**, 617-643 (1992).
11. J.F. Muzy, E. Bacry, A.Arneodo, The Multifractal Formalism revisited with Wavelets, *International Journal of Bifurcation and Chaos* **4**, No 2, 245-302 (1994).
12. Z.R. Struzik, A. Siebes, Wavelet Transform in Similarity Paradigm II, *CWI Reports*, in preparation, (1998).
13. C.J.G. Evertsz, Fractal Geometry of Financial Time Series, *Fractals* **3** No.3, 609-616 (1995).
14. M.F. Barnsley, *Fractals Everywhere*, Academic Press, NY, (1988).