Time Series Rule Discovery: Tough, not Meaningless

Zbigniew R. Struzik

Centrum voor Wiskunde en Informatica (CWI)
Postbus 94079, NL-1090 GB, Amsterdam, The Netherlands
email: Zbigniew.Struzik@cwi.nl

Abstract. 'Model free' rule discovery from data has recently been subject to considerable criticism, which has cast a shadow over the emerging discipline of time series data mining. However, other than in data mining, rule discovery has long been the subject of research in statistical physics of complex phenomena. Drawing from the expertise acquired therein, we suggest explanations for the two mechanisms of the apparent 'meaninglessness' of rule recovery in the reference data mining approach.
One reflects the universal property of self-affinity of signals from real life complex phenomena. It further expands on the issue of scaling invariance and fractal geometry, explaining that for ideal scale invariant (fractal) signals, rule discovery requires more than just comparing two parts of the signal. Authentic rule discovery is likely to look for the possible 'structure' pertinent to the failure mechanism of the (position and/or resolution-wise) invariance of the time series analysed.
The other reflects the redundancy of the 'trivial' matches, which effectively smoothes out the rule which potentially could be discovered. Orthogonal scale space representations and appropriate redundancy suppression measures over autocorrelation operations performed during the matches are suggested as the methods of choice for rule discovery.

1 Introduction

Recently, there has been considerable criticism of the mainstream rule discovery algorithm in data mining [1]. By performing scrutiny testing [1] suggests that the discussed algorithm [2] based rule discovery does not produce meaningful rules. In particular, the confidence of the rules recovered is not to be distinguished from the rules obtained from random noise. The overwhelming conclusions of the article would be disastrous for the domain of research in question if they lacked full explanation and understanding. In addition to the explanation provided in [1], the purpose of our paper is to propose a different look at the possible and plausible causes for the result reported in [1].

The primary investigated example in [1], coinciding with the example used by the primarily criticised paper by Das et al [2], is that of the S&P500 financial index. Indeed, the authors of [1] suggest that there is no more confidence in the particular rule advocated in [2] than in any other deterministic rule. Thus any

rule might do, which in actual fact means that such a rule is useless and irrelevant, holding at random, statistically meaningless instances. The mechanism of proving this conclusion has been devised by comparing the rule discovery algorithm from [2] on both the test time series (S&P500) and the surrogate time series (random walk). However, as the authors of [1] rightly indicate, the evidence for the lack of correlations in a financial time series like the S&P 500 index is so overwhelming that the 'meaninglessness' of any deterministic rule discovered may not seem surprising [3,4,5,6,7,8].

The article [1] suggests, however, that the same degree of meaninglessness is obtained no matter what input time series is used. The primary cause attributed to this failure is not in the clustering algorithm, which is the only rule extraction mechanism investigated, but in the pre-processing of the time series. In particular, the 'moving window' overlapping selection of candidate time series intervals leads to so-called spurious matches, destroying the resolution of the clustering algorithm.

The purpose of this writing is to look closer at the likely cause for the inability of the algorithms discussed blindly to extract rules from real life time series. In particular, the issue of scale invariance will be addressed, which characterises not only an overwhelming range of real-life and artificial time series but can also be attributed to isolated singularities - often the building blocks of the real-life and artificial time series.

Additionally, scaling invariance will be linked to the rate of auto-correlation decay, which determines the impact of 'redundant' spurious matches on the blind clustering algorithms. While auto-correlation decay is considered an important diagnostic tool in the study of long range dependence, for the purpose of blind clustering only the extrema (maxima or minima) of the autocorrelation (or the local match) may need to be considered to provide rule extraction with sufficient resolution and sensitivity.

2 Redundant information $\rightarrow$ spurious matches

The primary cause of the meaninglessness of the rule discovery has been attributed by the authors of the critical work [1] to the shortcoming of the time series pre-processing algorithm and in particular to the redundancy of the matching operation through the so-called 'trivial' matches. Indeed, matching two time-series intervals shifted with respect to one another by a time lag will indeed in many cases show a slow rate of decorrelation - which is reffered to as partial, trivial matches in [1].

Apart from the entire plethora of possible distance measures, the standard way of calculating the inner product of two time series is used for evaluating their 'correlation' level. For the time lag t shifted versions, the definition of the autocorrelation product/function $C(t)$ of a function $f(t)$ reads:

$$C(t) = \int_{-\infty}^{\infty} \bar{f}(\tau)\ f(t+\tau)\ d\tau \tag{1}$$

where $\bar{f}(\tau)$ is the complex conjugate of $f(t)$. Amazingly, the autocorrelation is simply given by the Fourier transform $\mathcal{F}$ of the absolute square of $f(t)$:

$$C(t) = \mathcal{F}(|f(t)|^2) \,, \tag{2}$$

and, of course, the Fourier transform of the second moment of the function is nothing else than its power spectrum $P(\omega) = \mathcal{F}(|f(t)|^2)$. This relationship is known by the name of the Wiener Khinchin theorem.

Thus, interestingly, the Fourier power spectrum is also related to the likely cause of the inability of the rule discovery to be selective enough in its pre-processing phase (feeding the rule extraction algorithms.) The importance of this in the context of our discussion lies in the fact that it links the scaling properties of the Fourier power spectrum with the decay rate of the auto-correlation function. Thus any property of the scaling invariance as discussed above will reveal itself in the invariance of the auto-correlation function. In particular, it will also determine the rate of decay of the auto-correlation function and will be inherited by the cross-correlation products of the time series with its sub-parts.

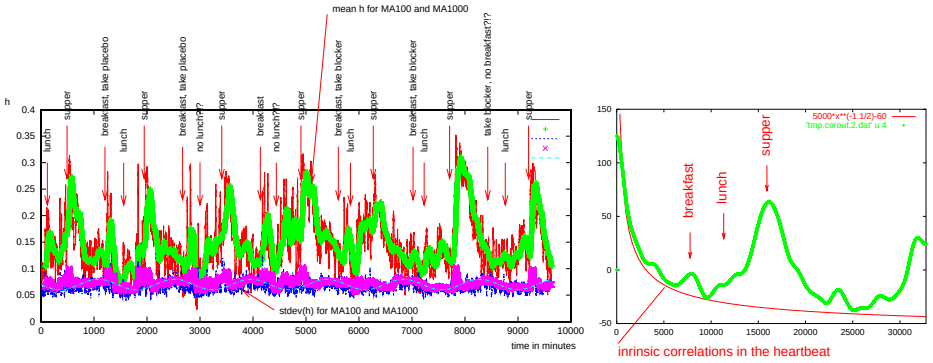

Fig. 1. Left: the plot of the variability of human heartbeat from a seven day long experiment where the test persons were given placebo or beta-blocker. For the variability estimate, local roughness (local correlation) exponent h is used, smoothed with a moving average (MA) filter with 100 and 1000 long window. An interesting pattern of response to food is evident [9]. Right: autocorrelation function confirms the presence of an invariant, intra-day periodic structure. The autocorrelation plot is in fact the autocorrelation of the local correlation exponent (as described with the Hölder h exponent.)

The sub-part matching operation is the key operation used in the rule discovery algorithms [2,1] and it clearly inherits the self-similarity properties of the time series. The explanation of the 'trivial match' redundancy which contributes to the inability of the algorithm to select sound rules thus comes from the spectral properties of the time series. The same spectral properties which, as we will show in the following, describe self-similarity properties of the time series.

It is worth noting that the autocorrelation decay is an important diagnostic tool widely used in investigating long range dependence (correlations), see e.g. [4] in the context of S&P500 analysis. However, due to the (Wiener Khinchin) equivalence referred to above, power spectrum decay has been extensively investigated in the same context. A modern method which allows local multiscale or multi-resolution decomposition of non-stationary signals as opposed to the global Fourier approach (useful for stationary processes), is the recently introduced wavelet transform. It allows local location-wise (temporal) and scale-wise (frequency) extraction of required information, including moments of the decomposition measure and regularity (scaling) exponents.

3 Estimating regularity properties of rough time series

The advent of multi-scale techniques (like WT), capable of locally assessing the singular behaviour, greatly contributed to the advance of analysis of 'strange' signals, including (multi)fractal functions and distributions. The wavelet transform [10,11,12,13] is a decomposition of the input time series into the discrete or continuous basis of localised (often compactly supported) functions - the wavelets. This decomposition is defined through an inner product of the time series with the appropriately rescaled and translated wavelet of a fixed shape. Wavelet decomposition schemes exist which allow decomposition optimisation through the choice from various wavelet bases [14,15] or adaptive decomposition (notably the lifting scheme [16]).

In the continuous formulation, the wavelet transform can be seen as a convolution product of the signal with the scaled and translated kernel, the wavelet $\psi(x)$:

$$(Wf)(s,b) = \frac{1}{s} \int dx\ f(x)\ \psi(\frac{x-b}{s}) \tag{3}$$

where $s, b \in \mathbf{R}$ and $s > 0$ for the continuous version.

For analysis purposes, one is not so much concerned with numerical or transmission efficiency or representation compactness, but rather with accuracy and adaptive properties of the analysing tool. Therefore, in analysis tasks, continuous wavelet decomposition is mostly used. The space of scale s and position b is then sampled semi-continuously, using the finest data resolution available. The numerical cost of evaluating the continuous wavelet decomposition is not as high as it may seem. Algorithms have been proposed which (per scale) have a complexity of the order n, the number of input samples, at a relatively low constant cost factor [17]. Additionally, computationally cheap, discretised, semi-continuous versions of the decomposition are possible [18,19].

In figure 2, we plot the input time series which is a part of the S&P index containing the crash of '87. In the same figure, we plot corresponding maxima derived from the WT decomposition with the Mexican hat wavelet. The maxima converging to the strongest singularity - the '97 crash have been highlighted in the top view.

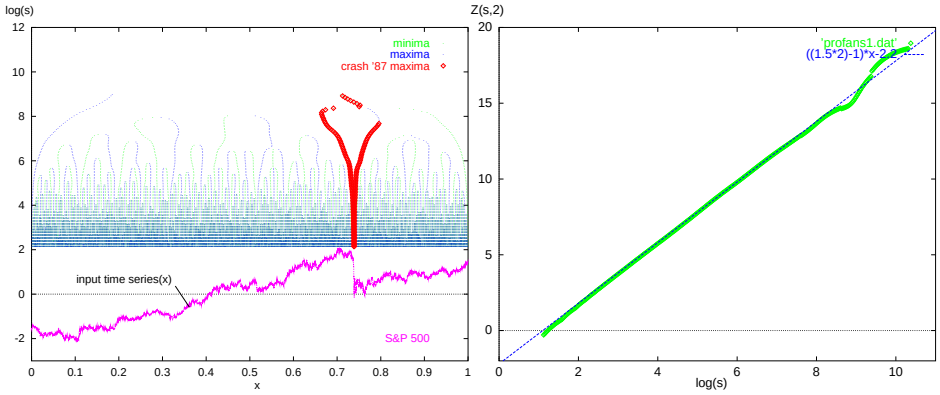

Fig. 2. Left: the L1 normalised S&P500 index time series with the derived WT maxima tree above it in the same figure. The strongest maxima correspond to the crash of '87. Right: the second moment of the partition function over the entire CWT (thus not only the maxima lines) see Eq. 4, shows consistent scaling invariance with the exponent $H + 1 = 1.5$. This corresponds with the Brownian walk scaling invariance exponent at $H = 0.5$.

There is an ultimate link between the global scaling exponent: the Hurst exponent H (compare figure 3), and the Fourier power spectral exponent. The power spectrum of the input signal and the corresponding scaling exponent γ can be directly evaluated from the second moment of the partition function $Z(s, q = 2)$:

$$\mathcal{Z}(s, q) = \sum_{\Omega(s)} (Wf\omega_i(s))^q \ , \tag{4}$$

where $\Omega(s) = \{\omega_i(s)\}$ is the set of maxima $\omega_i(s)$ at the scale s of the continuous wavelet transform $Wf(s, t)$ of the function $f(t)$.

Indeed the wavelet transform decomposes the signal into scale (and thus frequency) [1] dependent components (scale and position localised wavelets), comparably to frequency localised sines and cosines based Fourier decomposition, but with added position localisation. Scaling of the second moment of the decomposition coefficients provides γ, the power spectrum scaling, through $\gamma = 2H + 1$, the relation which links the spectral exponent γ with the Hurst exponent H.

4 Rules within rules, or the principle of self-similarity

Contrary to the long and widely accepted (Euclidean) view, real world time series/ signals are not smooth, but they are often non-differentiable and densely packed with singularities. Rough and wildly changing records are ever-present in

[1] The working scale of the wavelet s is inversely proportional to the (Fourier) frequency $f \sim 1/s$ and the continuous wavelet used is the second derivative of the Gaussian curve (*Mexican hat*).

nature [24,25,26,27,28]. The frequently adopted view that these signals consist of some smooth information carrying a component with superimposed noise is also very often inaccurate. Real life records are not necessarily contaminated by 'noise'. Instead, in the case of the lack of a better model, they often intrinsically consist exclusively of noise - indeed, they are 'noise' themselves.

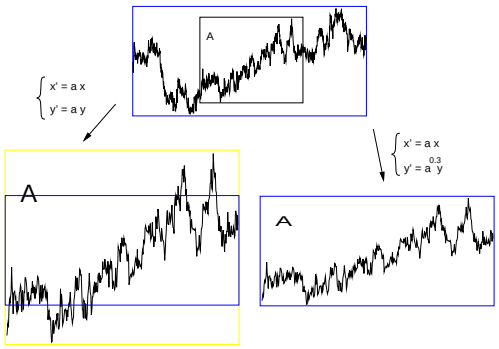

Fig. 3. The principle of self-affinity; *similar* rescaling in the bottom left figure versus *affine* rescaling, bottom right, of the fractional Brownian motion of $H = 0.3$. The rescaling factor used for the affine rescaling of the (x, y) axis is $(a, a^{0.3})$, while for a similar case both axes were rescaled using the a factor.

Their 'noisy' components are often distributed at various resolution and length scales - in other words each sub-part of the record is equally noisy and statistically similar (after affine rescaling of x,y coordinates with some factors β_x,β_y) to the entire record (or any other subpart). This kind of similarity can often be characterised by one single exponent $h = \log(\beta_y)/\log(\beta_x)$ for a range of β rescaling factors. This is the concept of local scaling which has been explored with the wavelet transform local scaling estimates in section 3. Additionally, the local scaling is often isotropic and the same one unique exponent can characterise both global and local ratio of the similarity rescaling. E.g. this is the case for 1-dim Brownian walk - the integral of white noise for which $h = 0.5$, and equals global $H = 0.5$, the so-called Hurst exponent. Such global scaling rules have been addressed through the partition function multifractal formalism [21].

Indeed the very essence of scaling, i.e. scale invariance, has the consequence that statistically similar patterns may occur at any resolution or scale length. This property characterising many real life signals may be behind the limited ability to extract meaningful deterministic 'rules' from such records [1], although it does permit statistical rule discovery [4,22]. Such can be used for distance evaluation for detecting rule violation for whole time series or streaming diagnosis etc., or streaming time series novelty assessment through departure from the model 'rule'.

In conclusion, the kind of redundancy inherent to a variety of signals analysed in [1] (for the discussion of the algorithm of [2]), as revealed in the 'trivial or

spurious matches' has been the subject of research into scaling, (multi-)fractal and long-range correlation properties in real-life phenomena and technology. Recently this has also been done using the advanced multiscale technique of wavelet transform permitting advanced 'inverse problem' type rule recovery.

The pressing question of course remains, what then is the meaningful methodology/strategy for dealing with signals inherently tough for rule discovery. The answer resides, in our opinion as outlined, in the spectral and auto-correlation properties of the time series. The rules which can be detected are instances of invariance violation, This can be manifested in the non-stationarity of spectral characteristics, be it a short-time power spectrum or multifractal spectrum. Or alternatively and simultaneously, in the breakdown of the scaling invariance - the structure which potentially emerges from the tree of the wavelet transform maxima [23,29,30,31,32], or possibly the structure emerging from the conduct of the self-adapting mechanism in multiscale/multiresolution decomposition or approximation bases [14,15,16,33].

References

1. J. Lin, E. Keogh, W. Truppel, When is Time Series Clustering Meaningful?, *preprint* Workshop on Clustering High Dimensional Data and its Applications, SDM 2003. will appear on the workshop site: `www.cs.utexas.edu/users/inderjit/sdm03.html`
2. G. Das, K. Lin, H. Mannila, G. Renganathan, P. Smyth, Rule Discovery from Time Series, in proceedings of the ..th Intl. Conference on Knowledge Discovery and Data Mining, New York, NY,Aug 27-31, 1998, pp 16-22, (1998).
3. R.N. Mantegna and H.E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* Cambridge, England: Cambridge University Press, (2000).
4. A. Arneodo, J.F. Muzy, D. Sornette, Eur. Phys J. B, **2**, 277 (1998). `\\http://xxx.lanl.gov/ps/cond-mat/9708012`
5. A. Johansen, D. Sornette, Stock Market Crashes are Outliers, Eur.Phys.J. B 1, pp. 141-143 (1998).
   A. Johansen, D. Sornette, Large Stock Market Price Drawdowns Are Outliers arXiv:cond-mat/0010050, 3 Oct 2000, rev. 25 Jul 2001.
6. B. Podobnik, P.Ch. Ivanov, Y. Lee, and H.E. Stanley. "Scale-invariant Truncated Lévy Process". Europhysics Letters, **52** pp 491-497, (2000).
7. Z. R. Struzik. Wavelet Methods in (Financial) Time-series Processing. Physica A: Statistical Mechanics and its Applications, 296(1-2):307-319, June 2001.
8. D. Sornette, Y. Malevergne, J.F. Muzy, Volatility Fingerprints of Large Shocks: Endogeneous Versus Exogeneous, arXiv:cond-mat/0204626, (2002).
9. Z. R. Struzik. Revealing Local Variablity Properties of Human Heartbeat Intervals with the Local Effective Hölder Exponent. Fractals **9**, No 1, 77-93 (2001).
10. S. Jaffard, Multifractal Formalism for Functions: I. Results Valid for all Functions, II. Self-Similar Functions, *SIAM J. Math. Anal.*, 28(4): 944-998, (1997).
11. I. Daubechies, *Ten Lectures on Wavelets*, (S.I.A.M., 1992).
12. M. Holschneider, *Wavelets - An Analysis Tool*, (Oxford Science Publications, 1995).
13. S.G. Mallat and W.L. Hwang, Singularity Detection and Processing with Wavelets. *IEEE Trans. on Information Theory* **38**, 617 (1992).
    S.G. Mallat and S. Zhong Complete Signal Representation with Multiscale Edges. *IEEE Trans. PAMI* **14**, 710 (1992).

14. S. Mallat, Z. Zhang, Matching Pursuit in a Time-frequency Dictionary, *IEEE Transactions on Signal Processing*, **41** pp. 3397-3415, (1993).
15. R.R. Coifmann M.V. Wickerhauser, Entropy-based Algorithm for Best-basis Selection. *IEEE Transactions on Information Theory*, **38**, pp. 713-718, (1992).
16. W. Sweldens, The Lifting Scheme: Construction of Second Generation Wavelets, *SIAM J. Math. Anal.* **29**, (2), pp 511-546, (1997).
17. A. Muñoz Barrutia, R. Ertlé, M. Unser, Continuous Wavelet Transform with Arbitrary Scales and O(N) Complexity, Signal Processing, vol 82, no. 5, pp. 749-757, May 2002
18. M. Unser, A. Aldroubi, S.J. Schiff, Fast Implementation of the Continuous Wavelet Transform with Integer Scales, IEEE Transactions on Signal Processing, vol. 42, no. 12, pp. 3519-3523, December 1994.
19. Z. R. Struzik, Oversampling the Haar Wavelet Transform. Technical Report INS-R0102, CWI, Amsterdam, The Netherlands, March 2001.
20. A. Arneodo, E. Bacry, J.F. Muzy, Oscillating Singularities in Locally Self-Similar Functions, *PRL*, **74**, No 24, 4823-4826, (1995).
21. A. Arneodo, E. Bacry and J.F. Muzy, The Thermodynamics of Fractals Revisited with Wavelets. *Physica A*, **213**, 232 (1995).
J.F. Muzy, E. Bacry and A. Arneodo, The Multifractal Formalism Revisited with Wavelets. *Int. J. of Bifurcation and Chaos* **4**, No 2, 245 (1994).
22. A.C.-C.Yang, S.-S. Hseu, H.-W. Yien, A.L. Goldberger, C.-K. Peng, Linguistic Analysis of the Human Heartbeat using Frequency and Rank Order Statistics, *PRL*, in press, (2003).
23. Z.R. Struzik, Taming Surprises, in proceedings of the *New Trends in Intelligent Information Processing and Web Mining* conference, Zakopane June 2-5, (2003).
24. K. Falconer, Fractal Geometry: Mathematical Foundations and Applications, John Wiley, 1990; paperback 1997.
25. A. Arneodo, E. Bacry, J.F. Muzy, Wavelets and Multifractal Formalism for Singular Signals: Application to Turbulence Data, *PRL*, **67**, No 25, 3515-3518, (1991).
26. H.E. Stanley, P. Meakin, Multifractal Phenomena in Physics and Chemistry, Nature, vol 335, 405-409, (1988)
27. P.Ch. Ivanov, M.G. Rosenblum, L.A. Nunes Amaral, Z.R. Struzik, S. Havlin, A.L. Goldberger and H.E. Stanley, Multifractality in Human Heartbeat Dynamics, *Nature* **399**, 461-465, (1999).
28. A. Bunde, J. Kropp, H.J. Schellnhuber, (Eds), The Science of Disasters, Climate Disruptions, Heart Attacks, and Market Crashes, Springer, (2002).
29. A. Arneodo, E. Bacry and J.F. Muzy, Solving the Inverse Fractal Problem from Wavelet Analysis, *Europhysics Letters*, **25**, No 7, 479–484, (1994).
30. A. Arneodo, A. Argoul, J.F. Muzy, M. Tabard and E. Bacry, Beyond Classical Multifractal Analysis using Wavelets: Uncovering a Multiplicative Process Hidden in the Geometrical Complexity of Diffusion Limited Aggregates. *Fractals* **1**, 629 (1995).
31. Z.R. Struzik The Wavelet Transform in the Solution to the Inverse Fractal Problem. *Fractals* **3** No.2, 329 (1995).
32. Z. R. Struzik, A. P. J. M. Siebes. Wavelet Transform in Similarity Paradigm. In Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, Volume 1394 of Lecture Notes in Artificial Intelligence, pp 295-309, Melbourne, Australia, April 1998.
33. A. Smola, B.Schl̈kopf, A Tutorial on Support Vector Regression, NeuroCOLT2 technical report NC-TR-1998-030, (1998).