

THE ALLAN VARIANCE AS AN ESTIMATOR OF THE LONG-MEMORY PARAMETER: TIME-DOMAIN AND WAVELET METHODS

Lara S. Schmidt, RAND Corporation

James G. Skinner, U.S. Naval Observatory

Abstract

The Allan variance is a well-known estimator of frequency stability and is often used to classify a time series into one of the standard clock noise types. By identifying the power-law model for clock noise with its long-memory equivalent, the Allan variance can also serve as an estimate for the long-memory parameter. Although the Allan variance is not a maximum likelihood estimator, it can be used with regression techniques that employ minimum variance estimates. This work describes the analytic basis for using the Allan variance to estimate the memory parameter, and performance of several Allan-variance-based estimators is illustrated via simulation study. Maximum likelihood estimation is also discussed, and the performance of maximum-likelihood estimators is contrasted with that of the Allan-variance-based estimators.

INTRODUCTION

This paper is intended to provide a synopsis of useful techniques for estimating the long-memory parameter with specific emphasis on Allan-variance-based techniques. The ideas described below are not new, although many of them may be new to the timing community. The intent is to provide some insight into these estimators, ample references, and an indication of the performance of these estimators via simulation studies – the hope being that the practitioner can then utilize these techniques as appropriate.

We begin with a brief introduction of long-memory processes and discuss why estimation of the long-memory parameter may be of interest to timekeepers. We then describe several techniques for estimation of this parameter that are found in the literature, and focus on two Allan-variance-based techniques. Finally, simulation studies are described that illustrate the performance of several of these estimators.

LONG-MEMORY PROCESSES

A class of time series processes known as long-memory or fractionally integrated processes (see, for example, [1]) have been used to describe the persistent correlation structures seen in fields such as hydrology, finance, astronomy, and timekeeping. These processes exhibit significantly large correlations between measurements separated by long time intervals, hence the name long-memory processes. The memory of these processes is characterized by a single parameter, d . $Y(t)$, a long-memory process with time index t , is denoted by $Y(t) \sim I(d)$ to indicate that it is the result of the fractional-integration of white noise of order d . The power spectral density of such a process is given by $S_Y(f) \sim c|f|^{-2d}$ for

small frequencies, f . That is, Y is a “power law” process. The slowly decaying autocorrelation function in terms of lag, h (for stationary processes where Stirling’s approximation holds), can be approximated by $\rho(h) \sim h^{2d-1} \Gamma(1-d)/\Gamma(d)$ as h tends to infinity, where $\Gamma(x)$ satisfies $\Gamma(x) = (x-1)!$. It is this form of the autocorrelation function that has provided much of the impetus for estimating d , as we will discuss in the next section.

In this work, we are interested in estimating the long-memory parameter. In timekeeping applications, we may wish to estimate d in order to describe the correlation structure of a time series for several reasons. For instance, it may be necessary to characterize clock behavior by identifying the autocorrelation structure as consistent with one of the standard noise types (white, flicker, random walk, etc.) or more generally by the value of d itself.¹ Or, we may want to separate autocorrelated noise from polynomial trend – perhaps by means of a prewhitening filter that employs an estimate of d . We may also wish to make predictions of future clock behavior via a statistical model (e.g., an *ARFIMA* (p,d,q) model; see [1]).

In the next section, we begin our discussion by outlining some available estimation techniques for d . We describe several techniques that make use of the Allan variance, a well-known estimator of frequency stability. We describe the analytical mechanism for these approaches and highlight the associated benefits and pitfalls. Monte Carlo simulation results are presented for several cases of interest. We conclude with recommendations regarding the use of the Allan variance as an estimator of the long-memory parameter.

ESTIMATORS OF THE LONG-MEMORY PARAMETER

There is a rich history² of estimation of d . Historically, these estimators were heuristic in nature with their graphic and analytic forms sometimes arising from the context of the applied problem. More recently, statistically optimal techniques for estimating d have emerged such as those based upon maximizing the likelihood function. We discuss several of those techniques below and put the Allan variance into context with other statistical approaches for estimating d .

GRAPHICAL TECHNIQUES

Graphical techniques have long served as an intuitive approach to estimation of d due to their visual nature. Hurst [2] developed the R/S statistic to aid hydrologists in making predictions of the flow of the Nile River. Based on dividing the range of adjusted cumulative sums by a measure of the variability of the process, the R/S statistic approximates a straight line with slope $d+1/2$ when plotted on a log-scale against lag. The correlogram [3] also allows estimation of d via a graphical technique. By plotting the

¹ Recall that the following identifications can be made:

Noise Type:	White PM	Flicker PM	White FM	Flicker FM	Random Walk FM
α	-2	-1	0	1	2
d	-1	-1/2	0	1/2	1

where d is not limited to these five values but can be any real number in the range $(-1, \infty), d \neq 0$. Thus, d can describe noise types “between” those typically considered and can be regarded as a way to classify the correlation properties of a process.

² The discussion here is by no means a complete listing of estimation techniques for the long-memory parameter. For more, see [4].

correlation function against lag on a log-log scale, d can be found by noticing that the slope is approximately equal to $2d-1$. A similar technique, the variogram, arose in geostatistics for analysis of spatial processes [4]. And the periodogram (an asymptotically unbiased estimate of the power spectral density) can be used with spectral-domain regression [5] to yield an estimate of d . In fact, there is an abundance of graphical techniques [4] that estimate d by approximating the slope in a linear relationship between some function of the variance and some description of lag – the Allan variance plot is one such technique.

THE ALLAN VARIANCE AS AN ESTIMATOR OF THE LONG-MEMORY PARAMETER

Timing professionals are well aware of the graphical nature of the Allan Variance plot (a.k.a. sigma-tau plot). Here, we discuss two approaches for using the Allan variance (*AVAR*) to estimate the memory parameter. First, we begin with time-domain regression on traditionally computed *AVAR* estimates; then we discuss the use of the wavelet representation of the *AVAR*.

Time-Domain Regression

The basis for using *AVAR* to estimate the memory parameter in the time domain hinges upon the following property of long-memory processes [6, 4]:

$$AVAR_Y(\tau) \sim \tau^{2d-1} \text{ for } -\frac{1}{2} < d < \frac{1}{2} \text{ and } \tau \rightarrow \infty,$$

where τ is averaging length and Y is, as usual, fractional frequency deviation. Therefore, an estimator for d may be found by computing the slope of the line between $\log \tau$ and $\log AVAR$ for sufficiently large τ . This slope may be estimated by regression. It is important to note, however, that this relationship holds only when τ is sufficiently large.³ But how large is large enough? This question has plagued many of the heuristic estimators (see, for example, the discussions of the “bandwidth parameter” [5, 7]), since choosing an inadequate lower bound for τ could result in estimates of d that are severely biased. Abry [8] suggests, when determining the lower cut-off for τ , to visually inspect a plot of confidence intervals (not simply points) for a linear relationship. The regression should exclude values of τ that are not large enough to support a linear relationship.

But determining the lower cutoff for τ is not the only difficulty when conducting time-domain regression to estimate d . Practitioners know well that *AVAR* estimates at high averaging lengths are substantially noisier than *AVAR* estimates for low- and mid-range τ . Therefore, we would expect *AVAR*-based estimates of d to be quite imprecise if we base these estimates on large values of τ . To alleviate this problem, we can assign more weight to the more precise *AVAR* estimates (at smaller τ) and less weight to the imprecise *AVAR* estimates (at larger τ).

Thus, we define an *AVAR* time-domain regression estimator for d as follows:

$$\hat{d}_{TR} = \frac{1}{2}(b_1 + 1)$$

where b_1 is the coefficient of $\log \tau$ (i.e., the slope term) in the weighted linear regression of $\log AVAR$ by $\log \tau$. This estimator is subscripted by *TR* to indicate that it is a Time-domain Regression estimator. The regression weights are defined by $w_i = 1/\text{var}(\log AVAR(\tau_i))$. Determining $\text{var}(\log AVAR(\tau_i))$ is the

³ Note that large τ is equivalent to low frequency.

most difficult aspect of the estimation process. Since our goal is to estimate the memory parameter, we cannot make use of the standard formulas [9] for estimating the variability of $AVAR$, since these formulas differ by noise-type (e.g., white, flicker, etc.) and hence require the noise-type to be known *a priori*, which is certainly not the case when trying to estimate d . Thus, one is forced to estimate the weights directly using maximum likelihood or iterative least squares techniques (see [10], Chapter 7), or is relegated to the use of unweighted regression – the resulting estimates will still be unbiased, but no longer have the minimum variance property. We will return to the discussion of estimation of weights and the lower cutoff for τ in the results section below.

Wavelet Regression

The second $AVAR$ -based estimator of d leverages the wavelet representation of $AVAR$. As shown in [11], the $AVAR$ is equivalent to the wavelet variance when the Haar mother wavelet is used. It is also known [12, 8], irrespective of the choice of mother wavelet, that the wavelet variance, $WVAR$, satisfies $WVAR \sim 2^{j(2d-1)}$, where j is the wavelet-level and $WVAR$ is the estimate of the variance of the wavelet coefficients. Thus, employing the Haar wavelet, we have

$$\log AVAR \sim j(2d - 1) + c.$$

Therefore, the following is also an estimator for d :

$$\hat{d}_{WR} = \frac{1}{2}(b_1 + 1)$$

where b_1 is the coefficient of j (i.e., the slope term) in the weighted linear regression of $\log AVAR$ by j . This estimator is subscripted by WR to indicate that it is a Wavelet Regression estimator. The weights may be estimated directly, may be ignored to pursue unweighted regression, or may be based upon the variance of $\log WVAR$. Percival [13] gives equations for confidence intervals for $WVAR$, the half-widths of which may serve as the basis for weights. For long-memory processes, the range of j for which the above relationship holds must be determined prior to estimation. Along the lines of the time-domain regression procedure, a lower cutoff must be established via visual inspection.

The use of weighted linear regression in both estimation techniques above overcomes the problem of unequal error variance at different levels of the independent variable. In this situation, the weighted linear regression technique is known to yield minimum-variance estimates of the regression coefficients (which is not the case with unweighted linear regression).

MAXIMUM LIKELIHOOD TECHNIQUES

Maximum likelihood estimation is an analytic technique that seeks to identify the value of the parameters for which the observed sample is the most likely. This is achieved by maximizing the likelihood function⁴ with respect to the parameters. In the timekeeping context, taking the fractional frequency deviations to be long-memory with parameter d , i.e., $Y(t) \sim I(d)$, it is reasonable to assume that Gaussian errors with mean 0 and variance σ^2 are appropriate. It can easily be shown that the resultant likelihood function is

⁴ The likelihood function, $L(\theta | x)$, for parameter θ and sample vector x , can be found by simply writing the probability distribution function, $f(x | \theta)$, and viewing the parameter as the free variable and the sample values as fixed. That is, $L(\theta | x) = f(x | \theta)$.

$$L(d, \sigma^2 | Y) = (2\pi)^{-n/2} \sigma^4 (\det(\Psi_d))^{-1/2} e^{-\frac{1}{2}(Y'(\sigma^2 \Psi_d)^{-1} Y)}$$

where n is the sample size and $\sigma^2 \Psi_d$ is the covariance matrix of Y which has the following form:

$$\sigma^2 \Psi_d = \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(n-1) \\ \gamma(1) & \gamma(0) & \dots & \gamma(n-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(n-1) & \gamma(n-2) & \dots & \gamma(0) \end{pmatrix} \text{ where } \gamma(h) = \frac{(-1)^h \Gamma(1-2d)\sigma^2}{\Gamma(h-d+1)\Gamma(1-h-d)}$$

The MLE is then found by maximization of the likelihood function (or, equivalently, the log of the likelihood function) with respect to the two parameters. In this case, maximization with respect to σ^2 can be achieved analytically, and the resultant estimator can be substituted into $L(d, \sigma^2 | Y)$. The reduced log-likelihood function then becomes

$$\log L(d | Y) = c - Y' \Psi_d^{-1} Y - \frac{1}{2} \log(\det \Psi_d).$$

However, due to the complexity of the matrix Ψ_d , maximization cannot be performed analytically. Thus, the MLE must be obtained via a numerical search over d . Although this technique appears to be computationally daunting, several approximate maximum likelihood techniques are available in both the time domain and spectral domain to speed estimation. In fact, software aids such as the R-language package [14] for computing approximate maximum likelihood estimates are readily available⁵ such that practitioners do not need to engage in the details.

Recent work with wavelets (see, for example, [15]) has produced yet another class of estimators of the long-memory parameter. Wavelet-based maximum likelihood techniques [12, 16] produce results that are similar to results from true MLE estimation. Instead of the fractional frequency process itself, these techniques use the wavelet-transform representation of the process that often has a covariance matrix of simpler form than the original process – often approximately diagonal. Thus, maximum likelihood estimation calculations (requiring inversion of the covariance matrix) are also simplified.

For the purposes of further discussion, we define the following notation: \hat{d}_{TM} for the Time-domain MLE and \hat{d}_{WM} for the Wavelet MLE. We now discuss the performance of the four estimators of d defined above by describing the results of a Monte Carlo simulation study.

SIMULATION RESULTS

A series of simulations were conducted to illustrate the behavior of *AVAR* and maximum-likelihood estimates of d . Five-hundred datasets, each of length 4096, were analyzed for each of four different values of d (0, 0.04, 0.14, 0.24). Thus, for each value of d , there are 500 estimates, \hat{d}_i , $i=1, 2, \dots, 500$ for each estimation technique. Since results were similar across levels of d , we will present only the results for $d=0.24$, a noise-type that is just shy of Flicker FM. Figure 1 displays the distribution of the 500

⁵ See <http://www.r-project.org>.

estimates when the true value of d was fixed at 0.24 (as denoted by the red horizontal line). For each estimation technique, a boxplot⁶ is shown which describes both the location (i.e., bias) and the spread (i.e., variance) of the estimates. Moving from left to right, the boxes represent the performance of each of the estimators: $\hat{d}_{TR}, \hat{d}_{WR}, \hat{d}_{TM}, \hat{d}_{WM}$.

It is clear from Figure 1 that the two MLE techniques on the right produce estimates that more faithfully reflect the true value of d – both in terms of bias and variance. However, all four techniques produce estimates that are reasonably symmetrically distributed and, to varying degrees, can be expected to deliver an estimate of d that is close to truth. The following describes the details for computation for each of the four techniques.

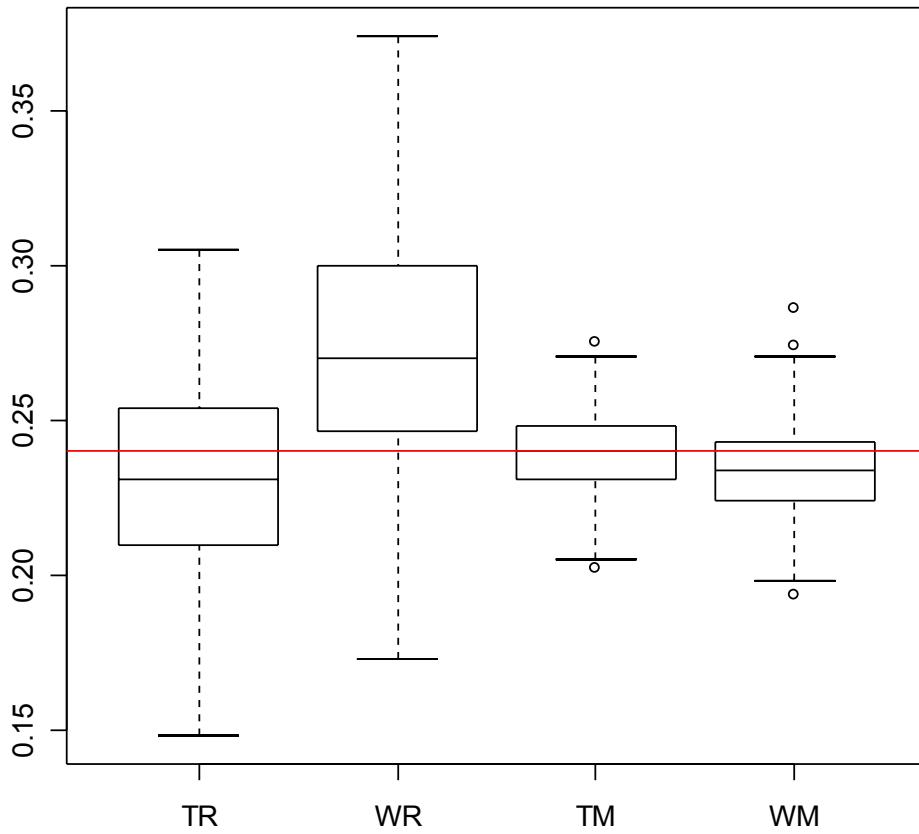


Figure 1

⁶ In a boxplot, the horizontal line on the interior of each box indicates the median, which is a robust estimator of central tendency. The top and bottom lines of the box represent the third and first quartiles respectively; thus, the central 50% of the observations lie within the box, and the height of the box represents the interquartile range. The interquartile range indicates the dispersion of the observations around the median and, thus, provides an indication of the spread of the distribution. The vertical “whiskers” protruding from the box extend to include all observations within 1.5 times the interquartile range. Any observations outside this range are denoted by open circles, and can be thought of as belonging to the tails of the distribution. The vertical axis represents the values of the estimates of the long-memory parameter.

The time-domain regression estimate, \hat{d}_{TR} , was formed as described in the preceding section, except that the weights were estimated empirically as part of the overall simulation study. Additionally, the first several weights were set to zero to avoid ill-advised weighting of lags (τ) that do not satisfy the asymptotic relationship between $AVAR$ and d , i.e., lags that are not “sufficiently large.” It should be noted that the identification of “sufficiently large” τ is not a well-defined process. Abry’s suggestion that the cutoff for τ be defined by including “all points whose confidence intervals support a line” failed in these simulations by admitting *all* values of τ , which yielded significantly biased results. Thus, we abandoned this technique and estimated the cutoff based on minimizing the mean squared error – an approach that can only be attempted during simulation studies. Thus, the results in Figure 1 should be regarded as a “best case scenario” for performance of the time-domain regression estimator. The fuzzy nature of the “sufficiently large” cutoff is a significant disadvantage to the time-domain regression approach. Figure 2 repeats the boxplots from Figure 1 (in black) and displays, in light blue, the boxplot that more accurately reflects the performance of the time-domain regression estimator in practice – formed without regard to cutoffs or weights. Although it yields more variable results with increased bias, we found this conservative approach to be less risky and much less time-intensive than using techniques for weights and cutoffs given in the literature.

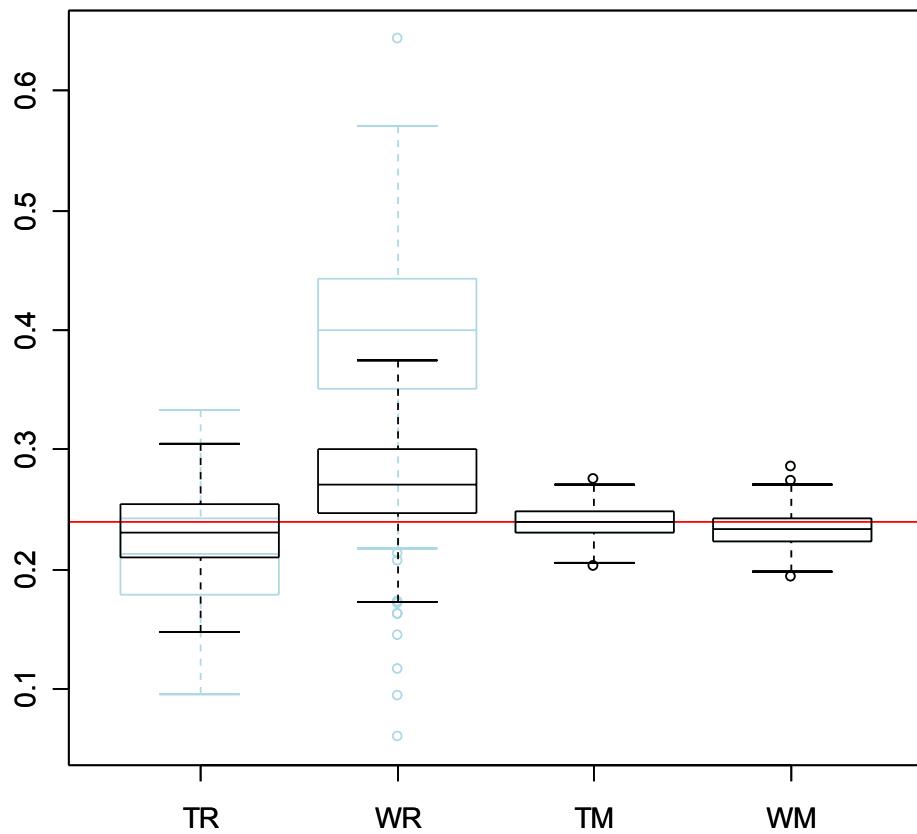


Figure 2

The wavelet regression estimate, \hat{d}_{WR} , was, again, formed as described in the preceding section, except that the weights were estimated empirically as part of the overall simulation study. After painful hours of experimenting with various weighting schemes based on confidence intervals for the *WVAR* [8], results were found to be misleading, often resulting in significantly biased results. Additionally, we abandoned ineffective and often-misleading graphical techniques (“visual inspection”) for identifying the lower cutoff for j . Instead, the cutoff was estimated by minimizing the mean squared error – a luxury only available to us due to the nature of simulation studies. Therefore, the results in Figure 1 should be regarded as “best case scenario” results for the wavelet regression technique. Figure 2 displays, in light blue, the boxplot that more accurately reflects the performance of the wavelet regression estimator in practice when estimates are formed without regard to cutoffs or weights. Although this conservative approach yields significantly noisy and biased results, we observed much more bias in results when weights were mis-specified. Since appropriate weights are unknown to the practitioner in general, mis-specification is a distinct possibility and brings with it the potential for significant errors in estimating d .

Table 1

Mean Squared Error			
TR	WR	TM	WM
.0027	.0290	.00016	.00024

The time-domain maximum likelihood estimator, \hat{d}_{TM} , was obtained using the “fracdiff” package in the R language which uses an approximate MLE approach. The wavelet MLE estimator, \hat{d}_{WM} , was obtained using the “waveslim” package in the R language. The Daubechies level 4 wavelet was employed. We can see that the performance of the two maximum likelihood techniques is similar. Both result in low mean-squared-error (as shown in Table 1) as a result of their small bias and low variability. Additionally, the MLE approaches, being available in readily accessible software, are easy to implement, run quickly, and are well documented. There is no need for visual inspection to determine lower cutoffs and no need for complex weighting schemes.

CONCLUSIONS

The *AVAR* estimates of d are both based on regression techniques. The time-domain version is based on a well-understood relationship between the *AVAR* and the memory-parameter (although historically stated in terms of α not d). The intuitive nature of these estimators (particularly the time-domain estimator) is attractive – graphical analysis, formalized via regression, yields the estimate of d . There are, however, two significant drawbacks to these *AVAR*-based approaches. First, estimation of the “sufficiently large” cutoff is troublesome. Despite much attention in the literature, an automated mechanism for identification of the appropriate cutoff is elusive. Techniques are *ad hoc*, time- and labor-intensive, and may, in the end, deliver misleading results. Secondly, due to the unequal error variances, weighted regression must be used to deliver minimum-variance estimates. But the appropriate weights are not known in general and, thus, must be estimated. Although this is not an insurmountable problem, it does add an additional layer of complexity to the regression techniques.

The maximum likelihood estimators are, on the other hand, easy to implement and produce outstanding results. In fact, the time-domain MLE for d is known [17] to converge almost surely to the true value of

d. In general, maximum-likelihood techniques are widely used and often produce minimum-variance estimates. Yet timekeeping practitioners may be reluctant to employ these techniques due to their slightly less intuitive (i.e., non-graphical) nature. We hypothesize, however, that once the small investment is made to explore the MLE and the available software, timekeepers will recognize the utility of this approach. Given the sound basis for MLE techniques and the superior performance of these estimators, we strongly recommend their use over *AVAR*-based approaches.

REFERENCES

- [1] P. Brockwell and R. Davis, 1987, **Time Series: Theory and Methods**, 2nd ed. (Springer Series in Statistics, Springer, Reading, Massachusetts).
- [2] H. E. Hurst, 1951, “*Long term storage capacity for reservoirs,*” **Transactions of the American Society of Civil Engineers**, **116**, 770-799.
- [3] G. Box, G. Jenkins, and G. Reinsel, 1994, **Time Series Analysis: Forecasting and Control**, 3rd ed. (Prentice Hall, Englewood Cliffs, New Jersey).
- [4] J. Beran, 1994, **Statistics for Long-Memory Processes**, Number 61 in Monographs on Statistics and Applied Probability (Chapman and Hall/CRC, Boca Raton, Florida).
- [5] J. Geweke and S. Porter-Hudak, 1983, “*The estimation and application of long memory time series models,*” **Journal of Time Series Analysis**, **4** (4), 221-238.
- [6] J. Barnes, J., *et al.*, 1971, “*Characterization of frequency stability,*” **IEEE Transactions on Instrumentation and Measurement**, **IM-20**, 105-120.
- [7] C. Hurvich, R. Deo, and J. Brodsky, 1998, “*The mean squared error of Geweke and Porter-Hudak’s estimator of the memory parameter of a long-memory time series,*” **Journal of Time Series Analysis**, **19** (1), 19-46.
- [8] P. Abry, P. Flandrin, M. Taqqu, and D. Veitch, 2000, “*Wavelets for the analysis, estimation and synthesis of scaling data,*” in K. Park and W. Willinger (eds.), **Self Similar Network Traffic Analysis and Performance Evaluation** (John Wiley & Sons, New York).
- [9] D. Howe, D. Allan, and J. Barnes, 1981, “*Properties of Signal Sources and Measurement Methods,*” in Proceedings of the 35th Annual Frequency Control Symposium, 27-29 May 1981, Philadelphia, Pennsylvania, USA (Electronic Industries Association, Washington, D.C.), pp. 669-715 (A1-A47).
- [10] G. Seber, 1977, **Linear Regression Analysis** (John Wiley & Sons, New York).
- [11] D. Howe and D. Percival, 1994, “*Wavelet Analysis for Synchronization and Timekeeping,*” in Proceedings of the 1994 IEEE International Frequency Control Symposium, 1-3 June 1994, Boston, Massachusetts, USA (IEEE 94CH3446-2), pp. 791-797.
- [12] M. Jensen, 2000, “*An alternative maximum likelihood estimator of long-memory processes using compactly supported wavelets,*” **Journal of Economic Dynamics and Control**, **24** (3), 361-387.
- [13] D. Percival, 1995, “*On estimation of the wavelet variance,*” **Biometrika**, **82** (3), 619-631.

- [14] R Development Team, 2003, **R: A language and environment for statistical computing** (R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3).
- [15] D. Percival and A. Walden, 2000, **Wavelet Methods for Time Series Analysis**, Cambridge Series in Statistical and Probabilistic Mathematics (Cambridge University Press, Cambridge).
- [16] P. Craigile, P. Guttorp, and D. Percival, 2004, “*Wavelet-based parameter estimation for trend contaminated fractionally differenced processes*,” NRCSE Technical Report NRCSE-TRS-77 (University of Washington, 4 February 2004).
- [17] Y. Yajima, 1985, “*On estimation of long-memory time series models*,” **Australian Journal of Statistics**, **27** (3), 303-320.