Long-memory analysis of time series with missing values

P. S. Wilson,* A. C. Tomsett, and R. Toumi
Space and Atmospheric Physics, Blackett Laboratory, Imperial College, London, SW7 2BW, United Kingdom
(Received 19 November 2002; published 21 July 2003)

The estimation of long memory is often restricted by missing data. We examine the effects on the estimation of long memory of three simple gap-filling techniques: interpolation, random, and mean filling. Numerical simulations show that the gap-filling techniques introduce significant deviations from the expected scaling behavior for both persistent and antipersistent time series. For persistent time series the interpolation method provides a reliable estimation of long memory for scales longer than the largest likely gap.

DOI: 10.1103/PhysRevE.68.017103 PACS number(s): 02.50.−r, 05.40.−a

I. INTRODUCTION

The presence of long memory (the terms long-range dependence, long-range persistence, and long-range antipersistence are also used) has been identified in diverse fields ranging from geophysics and atmospheric phenomena [1,2], to biological systems [3–5], and financial volatility [6]. Long memory is a term used to describe a time series where correlations obey the power-law scaling relationship

$$\rho(\tau) \approx c_\rho |\tau|^{-\alpha}. \tag{1}$$

Here $\rho(\tau)$ is the autocorrelation function at lag τ, c_ρ is a positive constant, and α is the long-memory parameter, where $0<\alpha<1$. The decay of correlations with τ is hyperbolic, leading to divergence of the sum,

$$\sum_{\tau=-\infty}^{\infty} \rho(\tau) = \infty. \tag{2}$$

This is in contrast to random processes where no correlations exist and to short-memory processes (e.g., ARMA and Markov models) where correlations decay exponentially with τ. Beran [7] defines a stationary process with long memory as a process for which

$$\lim_{\tau\to\infty} \rho(\tau)/[c_\rho |\tau|^{-\alpha}] = 1. \tag{3}$$

Equation (3) is an asymptotic definition requiring a long homogeneous dataset for accurate estimation of the long-memory parameter α. Equations (1)–(3) apply to persistent time series. For antipersistent time series the square of the autocorrelation function, $\rho(\tau)^2$, must be used such that α is replaced by $\alpha+2$.

A variety of techniques are available to estimate long memory in time series (see Beran [7], Malamud and Turcotte [8], and Taqqu and Teverovsky [9] for reviews). The application of all methods to ''real world'' data suffer from the competing factors of data length and homogeneity. The restriction of long-memory analysis to only homogeneous sections of time series [10] is a severe limitation particularly in nonrepeatable experiments typical in geophysics and related fields. It is the aim of this paper to examine the effect of several standard techniques for replacing missing data on the estimation of long memory, with the aim of establishing the applicability of the techniques to incomplete datasets.

Several previous studies have examined the effects of missing data on long-memory estimation [11,12]. These have, however, focused on estimation in state space using the Kalman filter to account for the missing data. Researchers in the physical sciences consistently prefer a more direct, heuristic approach to estimation of long memory.

II. METHODOLOGY

Analysis techniques to identify long memory in a time series can be categorized according to the underlying theory, random walk or spectral, and also by the detrending or nondetrending properties. Accurate estimation of long memory with nondetrending methods requires the preprocessing of data to remove trends. We will use the Hurst exponent H to define long memory, where $H>0.5$ indicates long-range persistent data, $H=0.5$ implies the absence of long memory, and $H<0.5$ indicates long-range antipersistent data. As such, $0<H<1$ defines the limits of our study.

We examine two popular techniques, detrended fluctuation analysis (DFA) developed by Peng *et al.* [3] (a random walk detrending method), and the modified periodogram technique [13] (a spectral nondetrending method). These are representative of the range of techniques, and reflects the recommended approach to long-memory analysis [14], where multiple methods are integrated to avoid potential misdiagnosis.

First-order DFA gives the fluctuation function $F(\tau)$, the square root of the variance of the linearly detrended profile of the time series averaged over segments of size τ, such that $F(\tau) \propto \tau^H$ [3]. Higher-order DFA with polynominal detrending [15] is not widely used and will not be examined here. The exponent H is estimated directly from a linear fit of $F(\tau)$ versus τ on a logarithmic plot and is related to α [Eq. (1)] by $H=1-\alpha/2$. Modifications to the DFA method, which correct for deviations from scaling at small time scales intrinsic to the standard DFA method, have been developed [4,15]. This study will use the modified DFA method developed by Kantelhardt *et al.* [15].

The standard periodogram method estimates the spectral

*Corresponding author.
Email address: paul.wilson@imperial.ac.uk

density $S(f)$ of the times series. The spectral density is simply the Fourier transform of the autocorrelation function $\rho(\tau)$ [Eq. (1)]. Thus $S(f)$ is related to f by $S(f) \approx f^{-\beta}$, where $\beta = 1 - \alpha$. The modified technique accounts for the predominance of high-frequency points on a logarithmic plot by averaging over logarithmically equispaced bins. The exponent β is related to H by $H = (\beta + 1)/2$.

The analysis techniques are applied to 30 independent realizations of fractal Gaussian noises (FGNs), synthetic time series lying between classic white noise and brownian motion [16,17]. FGNs are generated with specific H exponents from a Gaussian distribution with zero mean and unit variance via the method of Davies and Harte [18]. A specific percentage P_g of the time series is simulated as missing by selecting the appropriate percentage of randomly distributed data points to remove, resulting in a geometric distribution of gap lengths [19]. Missing data are replaced via appropriate gap-filling techniques. This study is restricted to between 0% and 20% missing data, representing the realistic range of missing data observed in real world time series (e.g., UK Meteorological Office European synoptic station precipitation records available at www.badc.rl.ac.uk).

The gap-filling techniques applied are linear interpolation, random, and mean fills as described below.

Interpolation fill. Data are replaced by linearly interpolating across the data gaps.

Random fill. Data are replaced with random data drawn from the empirical distribution of the time series, in this case, a Gaussian distribution with unit mean and variance.

Mean fill. Data are replaced using the mean of the empirical distribution of the time series, in this case zero.

For examples of the application of these techniques, see Refs. [20,21]. For a discussion on more sophisticated gap-filling techniques, see the work of Little and Ruben [22].

III. RESULTS

The results are presented as the ensemble average, normalized to the mean, of the 30 independent realizations. In all cases the results of the modified periodogram method (not shown) are equivalent to the DFA results shown. This has been analytically justified by Heneghan and McDarby [23] who examined the relationship between DFA and the power spectral density, showing that DFA and spectral measures provide equivalent characterizations of long-memory properties for stationary signals.

Results for the specific case $H = 0.5$, the absence of long memory (not shown), show no deviation for the random or mean fill techniques. The interpolation technique adds short range correlations limited to short lags of $\tau \lesssim 30$ for 10% missing data. Gap filling cannot lead to the spurious diagnosis of long memory if it is absent from the time series.

A. Long-ranged persistent $H > 0.5$

Figure 1 shows results of DFA for $H = 0.7$, representative of the effects of gap filling over the range $0.5 > H > 1$. The results clearly indicate deviation from the expected scaling behavior with no data gaps at short lags (high frequency). The random fill and mean fill methods show a curvature towards the absence of long memory while the interpolation method shows a curvature towards highly correlated behavior. The variance for each gap-filling technique and percentage of missing data is comparable at equivalent lags. The standard deviation is less than 3% of the ensemble mean for short lags and less than 17% at long lags. Figure 2 shows the

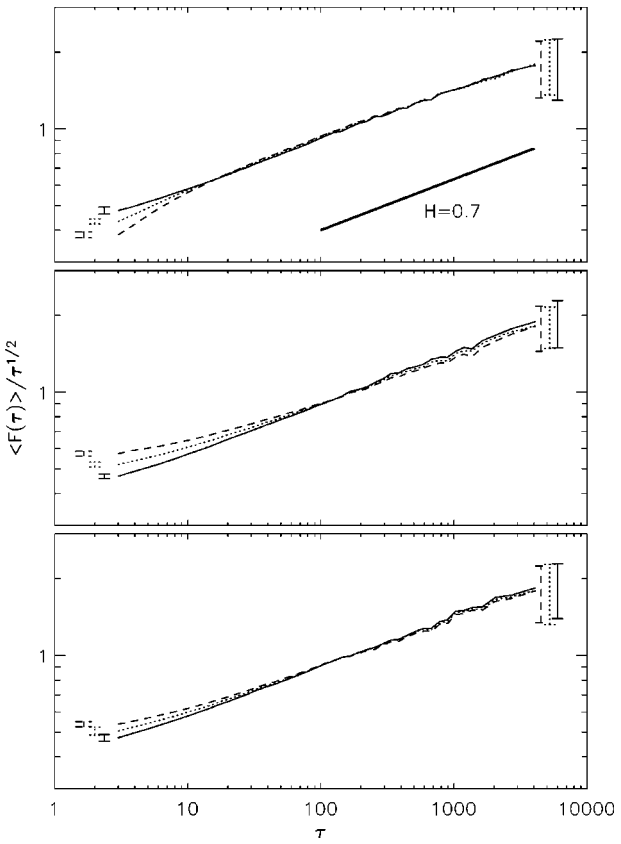

FIG. 1. Ensemble averages, normalized to the mean, of 30 independent FGN simulations with 8192 data points and $H = 0.7$. Top panel: interpolation of missing data; middle panel: random fill of missing data; and bottom panel: mean fill of missing data. Solid lines—0% missing data, dotted lines—10% missing data, and dashed lines—20% missing data. ± 1 standard deviation error bars for the shortest and longest lags are included.

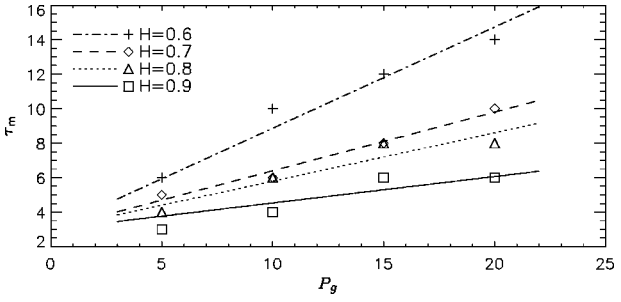

FIG. 2. Variation in the minimum lag τ_m for which the ensemble mean of $F(\tau)$, of 30 FGN realizations with P_g% gaps and interpolation filling, deviates from the expected scaling with no data gaps by less than 1 standard deviation.

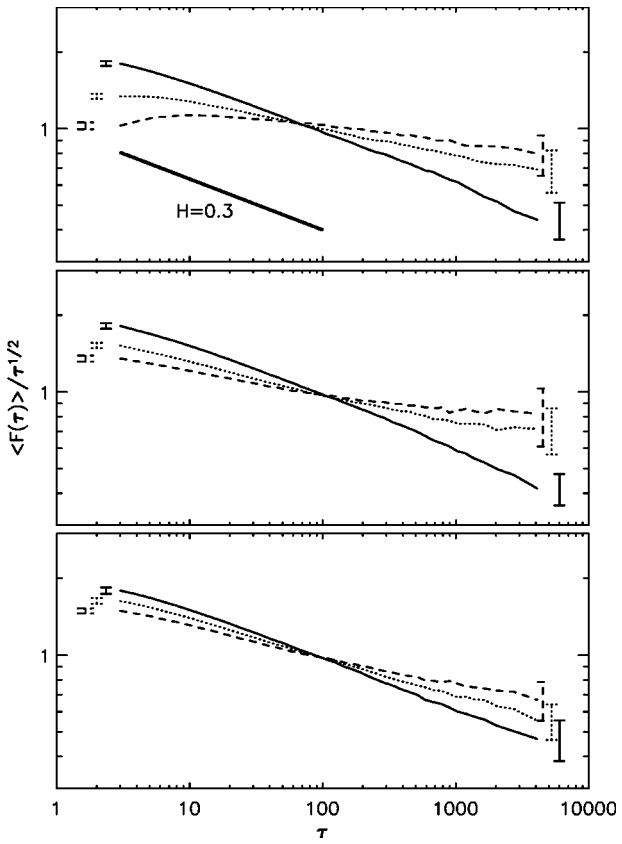

FIG. 3. Ensemble averages, normalized to the mean, of 30 independent FGN simulations with 8192 data points and $H=0.3$. Top panel: interpolation of missing data; middle panel: random fill of missing data; bottom panel: mean fill of missing data. Solid lines: 0% missing data; dotted lines: 10% missing data; dashed lines: 20% missing data. ± 1 standard deviation errors bars for the shortest and longest lags are included.

minimum lag for which there is no deviation from true scaling with the interpolation fill.

B. Long-ranged antipersistent $H<0.5$

The results for the long-range antipersistent case $H<0.5$, Fig. 3, show deviation from the expected scaling behavior with no data gaps at all lags. The magnitude of the deviation is greater at long lags (low frequency) than at short lags (high frequency). The deviations towards the absence of long memory are significantly greater than those for long-ranged persistent time series, consistent with the conclusion of Chen *et al.* [24] that antipersistent signals are less robust to the effects of nonstationarities compared to persistent signals. The potential for underestimating the strength of long memory for antipersistent time series with missing data is significantly greater than for persistent time series.

IV. DISCUSSION

The results are consistent with intuitive expectations. Random and mean fills destroy correlations between data points such that the deviations from the true scaling tend asymptotically to $H=0.5$. Chen *et al.* [24], in studying the general application of DFA to time series with nonstationarities, showed that the DFA signal $F(\tau)$ comprises the sum of the squares of $F(\tau)$ of the individual components of the time series, with the condition that the individual components were uncorrelated. It was argued that the randomly located jumps between zero (mean fill method) and nonzero segments of a time series introduces a random component to the analysis. The results of the DFA analysis for random and mean fill techniques are completely consistent with Chen *et al.* where the gap-filling technique is simply considered a form of nonstationarity. Chen *et al.* were, however, unable to accurately distinguish the short lag features, using the standard DFA method.

Considering the autocorrelation function, it is possible to present an argument for the effect of random or mean gap-filling techniques on the autocorrelation function $\rho(\tau)$ such that $\rho(\tau)$ is modified by a multiplicative factor of P_g^2 for $\tau \geqslant 1$ [25]. The resulting modification of the step $\rho(0)$ to $\rho(1)$ can be considered a simple form of short-range dependence with $H=0.5$. Qualitatively this can be considered to act in competition with the true long memory.

No similar simple statement regarding the fluctuation function or autocorrelation function can be made for the interpolation gap-filling technique where the gap-filling is not independent of the actual data. Hu *et al.* [26] analytically demonstrated that DFA on a linear trend gives a Hurst exponent of 2. $H=2.0$ should thus represent the asymptotic limits of the observed deviations. For correlated data the interpolation fill shows deviations towards this asymptotic limit but only at scales shorter than or similar to the expected maximum gap length: 6 and 4 for 20% and 10% gaps, respectively (calculated from the expected maximum of a geometric distribution with known series length using the theory of large numbers [27]). At longer scales the dependence of the interpolation fill on the actual time series maintains the true long memory properties. For antipersistent signals there is evidence that at short lags the deviation is tending to the expected asymptotic limit; however, at larger scales the interpolation destroys the correlation structure in a similar manner to the random or mean fills.

The significant feature of the results for the estimation of long memory in ''real world'' data is the scale of influence of the gap filling techniques. For persistent data the low frequency components of the time series $\tau \gtrsim 100$, which are principally manifest as local trends, are robust to the gap filling techniques. For antipersistent data the mean reversion property intrinsic to the time series inhibits drift from the mean resulting in sensitivity at all scales to the gap filing techniques.

V. CONCLUSION

We have examined the effects of three simple gap-filling techniques on the estimation of long memory via two popular methods, DFA and Modified Periodogram. While supporting the intuitive expectations of the filling techniques on the correlations, the accurate estimation of the long-memory parameter is shown to be a viable prospect only if the effects of the gap-filling technique are considered. The results indicate

that long-ranged antipersistent time series are significantly less robust to gap-filling techniques than long-ranged persistent time series. The results also indicate that for persistent time series maintaining some dependence between the data fill and the true time series limits the scale of influence of the gap-filling technique to the scale of the gaps. This study has focused on generalities: no attempt is made to address specific situations, e.g., where missing data is not randomly distributed, such as whole years of missing data in a daily time series. While it is obviously possible to construct more sophisticated methods for specific situations, the analysis provides a succinct outline of the general expectations.

ACKNOWLEDGMENT

The work was supported by a NERC CASE award with the Environment Agency of England and Wales.

[1] J.S. Syroka and R. Toumi, Geophys. Res. Lett. **28**, 3255 (2001).
[2] E. Koscielny-Bunde, A. Bunde, S. Havlin, H.E. Roman, Y. Goldreich, and H.J. Schellnhuber, Phys. Rev. Lett. **81**, 729 (1998).
[3] C.K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H. Stanley, and A. Goldberger, Phys. Rev. E **49**, 1685 (1994).
[4] S. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **51**, 5084 (1995).
[5] Y. Ashkenazy, P.C. Ivanov, S. Havlin, C.K. Peng, A.L. Goldberger, and H.E. Stanley, Phys. Rev. Lett. **86**, 1900 (2001).
[6] R.T. Baillie, Econometrica **73**, 5 (1996).
[7] J. Beran, *Statistics for Long-Memory Processes* (Chapman and Hall, New York, 1994).
[8] B.D. Malamud and D.L. Turcotte, *Long-Range Persistence in Geophysical Time-Series*, Advances in Geophysics Vol. 40 (Academic Press, New York, 1999).
[9] M.S. Taqqu and V. Teverovsky, *A Practical Guide to Heavy Tails* (Birkhäuser, Boston, 1998), Chap. 8, pp. 177–218.
[10] A. Tsonis, P. Roebber, and J. Elsner, J. Clim. **12**, 1534 (1999).
[11] W. Palma and N.H. Chan, J. Forecasting **16**, 395 (1997).
[12] W. Palma and G.D. Pino, Biometrika **86**, 965 (1999).
[13] M.S. Taqqu, V. Teverosky, and W. Willinger, Fractals **3**, 785 (1995).
[14] G. Rangarajan and M. Ding, Phys. Rev. E **61**, 4991 (2000).
[15] J.W. Kantelhardt, E. Koscielny-Bunde, H.H.A. Rego, S. Havlin, and A. Bunde, Physica A **295**, 441 (2001).
[16] B.B. Mandelbrot and J.W.V. Ness, SIAM Rev. **10**, 422 (1968).
[17] B.B. Mandelbrot and J.R. Wallis, Water Resour. Res. **5**, 228 (1969).
[18] R.B. Davies and D.S. Harte, Biometrika **74**, 95 (1987).
[19] M. Evans, N. Hastings, and B. Peacock, *Statistical Distributions* (Wiley, New York, 2000).
[20] P. Talkner and R.O. Weber, Phys. Rev. E **62**, 150 (2000).
[21] A.R. Rao and D. Bhattacharya, J. Hydrol. **216**, 183 (1999).
[22] R.J.A. Little and D.B. Ruben, *Statistical Anylsis with Missing Data* (Wiley, New York, 1987).
[23] C. Heneghan and G. McDarby, Phys. Rev. E **62**, 6103 (2000).
[24] Z. Chen, P.C. Ivanov, K. Hu, and H.E. Stanley, Phys. Rev. E **65**, 041107 (2002).
[25] R. Chandler (private communication).
[26] K. Hu, P.C. Ivanov, Z. Chen, P. Carpena, and H.E. Stanley, Phys. Rev. E **64**, 011114 (2001).
[27] D. Sornette, *Critical Phenomena in the Natural Sciences* (Springer, Berlin, 2000).