Stochastic models for characterisation and prediction of time series with long-range dependence and intermittency

Vo Anh, School of Mathematics, Queensland University of Technology, GPO Box 2434, Brisbane, Q. 4001.
Karen Lunney, Katestone Scientific Pty Ltd, PO Box 2012, Bardon, Q. 4065.

Abstract Time series from various fields such as geophysics, meteorology, hydrology, air pollution are often intermittent and display long-range dependence. This paper develops a new class of stochastic models to represent such properties. An efficient estimation procedure is outlined and tested on two concentration time series collected in an environmental wind tunnel. These time series simulate two different types of odour sources and possess quite different statistical properties that are well described by the new model.

1. INTRODUCTION

Long-range dependence (LRD) is characterised by slowly decaying serial correlations so that the autocorrelation function of the time series is not absolutely summable. The presence of long-range dependence invalidates many of the traditional approaches of data description using ARMA models (Beran (1992)).

Another dominant feature of turbulent time series is their intermittency, which is traditionally described by the rate of dissipation of kinetic energy. Arneodo *et al.* (1992), Farge (1992), and Davis *et al.* (1994) advocated the use of wavelet analysis in studying intermittency. But, as apparent in the numerical results of Arneodo *et al.* (1992), difficulties are inherent in this wavelet approach.

Apart from some attempts based on stochastic differential equations, most recent studies have been geared towards analysis of turbulent systems rather than building explicit models for prediction of future behaviour of these systems. A notable example is the investigation of zero-crossings of the wavelet transform of turbulence time series, resulting in a scale-space representation of the time series (see Arneodo *et al.* (1992), Farge (1992)). The representation gives information on the singularities of the time series across scales, which helps understanding the underlying generating process but does not yield a model for its prediction. In this paper, we shall pay attention to developing explicit models for describing simultaneously LRD and intermittency in turbulent time series. These models can be used directly for prediction purposes.

In Section 2, we define the concepts of LRD and intermittency in terms of the spectral density of the stochastic process. An appropriate form of the spectral density is then used to model LRD and intermittency simultaneously. In Section 3, a discrete approximation for the models is obtained. In Section 4, we give a fast algorithm to estimate all the parameters of the resulting discrete model. As an example of application of the method, we analyse two odour records simulated in the Monash University experimental wind tunnel.

2. LONG-RANGE DEPENDENCE AND INTERMITTENCY

A stochastic process $X(t)$ is said to exhibit LRD if its spectral density has the form

(2.1) $$f(\omega) = f_*(\omega)\omega^{-2\beta}\ , \ \beta > 0, \omega \in \Re,$$

where $f_*(\omega)$ is slowly varying as $\omega \to 0$ (i.e. $\dfrac{f_*(a\omega)}{f_*(\omega)} \to 1$ as $\omega \to 0$ for any $a > 0$). The spectral density has an integrable singularity at the origin if $0 < \beta < \frac{1}{2}$, with the characteristic effect that the covariance function of $X(t)$ decays to zero at a very slow rate.

LRD is a contribution to the process from the low-frequency components of the spectrum. On the other hand, intermittency is a contribution to the process from the high-frequency components of the spectrum. These latter components are very irregular. Their scattering effect yields a positive dimension to their inverse image. Berman (1972) showed that if $X(t)$ has continuous sample paths and if

(2.2) $$|\omega|^{2\alpha-1}\left(1+\omega^2\right)f(\omega) \geq C \text{ as } |\omega| \to \infty$$

for some α and C, $0 < \alpha < 1$, $C > 0$, then the set

(2.3) $$\left\{x; \dim\left\{t; 0 \leq t \leq T, X(t) = x\right\} < 1-\alpha\right\}$$

is nowhere dense almost surely. If $X(t)$ also satisfies

(2.4) $$|\omega|^{2\beta-1}\left(1+\omega^2\right)f(\omega) \leq B \text{ as } |\omega| \to \infty$$

for some β and B, $0 < \beta < 1$, $B > 0$, then

(2.5) $$\dim\left\{t; 0 \leq t \leq T,\ X(t) = x\right\} \leq 1-\beta.$$

For $X(t)$ stationary and ergodic, the conditions (2.2) and (2.4) imply, with $\alpha = \beta$,

(2.6) $$\dim\left\{t; t \geq 0,\ X(t) = x\right\} = 1-\alpha,\ \forall x, a.s.$$

Since intermittency can be viewed as the crossings at a certain level $X(t) = x$, the results (2.5) and (2.6) suggest that the parameter α can be used as an indicator of the degree of intermittency of the process $X(t)$. A special spectral density featured in the work of Berman (1972) is

$$(2.7) \qquad f(\omega) = \frac{|\omega|^{1-2\alpha}}{1+\omega^2}, \; 0 < \alpha < 1, \; \omega \in \Re.$$

It can be shown that $f(\omega)$ of (2.7) satisfies conditions (2.2) and (2.4). In view of (2.1) and (2.7), we shall consider spectral densities of the form

(2.8)

$$f(\omega) = \frac{\sigma^2 |\omega|^{1-2\alpha}}{(b^2+\omega^2)^\beta}, \; 0 < \alpha < \tfrac{1}{2}, \; \beta > 0, \; \sigma, b, \omega \in \Re$$

to model LRD and intermittency simultaneously. The term b^2 is included in the denominator of (2.8) to enhance stationarity and to indicate the local self-similarity of the process. When $b = 0$, the process is self-similar.

3. STOCHASTIC MODELS

In this section, we obtain a discrete approximation to the spectral density (2.8) with $b \to 0$. We first consider an approximation when $\omega \to 0$, which is relevant for LRD. The intermittency components have little contribution to the spectrum at very low frequencies (near 0). Hence we may put $\alpha = 0$ in (2.8) as $\omega \to 0$, and the spectral density becomes

$$(3.1) \qquad f_X(\omega) = \frac{\sigma^2}{\omega^{2\beta-1}}, \; \beta > 0, \; \omega \in \Re.$$

It follows that $X(t)$ may not be stationary; but the related process

$$(3.2) \qquad Y(t) = \int_{t-\frac{1}{2}}^{t+\frac{1}{2}} X(u)\,du$$

is stationary and its spectral density

$$(3.3) \qquad f_Y(\omega) = \sigma^2 \left(4\sin^2 \frac{\omega}{2}\right) \omega^{-1-2\beta}, \; \omega \in \Re$$

approximates (3.1) closely as $\omega \to 0$. The discrete process which arises from $Y(t)$ by periodic sampling will now have a spectral density which is the Poisson sum of (3.3), that is,

$$f_D(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} f_Y(\omega + 2k\pi)$$

$$= \frac{2\sigma^2}{\pi} \sin^2 \frac{\omega}{2} \sum_{k=-\infty}^{\infty} |\omega + 2k\pi|^{-1-2\beta}, \; \omega \in [-\pi, \pi]$$

$$(3.4) \qquad \sim \frac{\sigma^2}{2\pi} \omega^{1-2\beta} \text{ as } \omega \to 0.$$

Noting that $|1-e^{i\omega}|^2 = 4\sin^2 \frac{\omega}{2} \to \omega^2$ as $\omega \to 0$, the result (3.4) suggests the approximation

$$(3.5) \qquad f_D(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|1-e^{i\omega}|^{2d}}, \; \omega \in [-\pi, \pi],$$

$$(3.6) \qquad d = \beta - \tfrac{1}{2}$$

to (3.1) as $\omega \to 0$.

On the other hand, the intermittency in (2.8) signifies a high-frequency behaviour of the process. Therefore, we shall consider a discrete approximation of (2.8) in a neighbourhood of $\omega = \pi$ and put $\beta = 1$ in (2.8) as LRD has little contribution at these frequencies. The spectral density of the discrete process which arises from (2.8) by periodic sampling will now have the form

(3.7)

$$f_D(\omega) = \frac{\sigma^2}{2\pi} \sum_{k=-\infty}^{\infty} \frac{|\omega + 2k\pi|^{1-2\alpha}}{b^2 + (\omega + 2k\pi)^2}, \; \omega \in [-\pi, \pi]$$

$$\to \frac{\sigma^2}{2\pi} \sum_{k=-\infty}^{\infty} \frac{1}{(2\pi)^{1+2\alpha}} \frac{1}{|k+\frac{1}{2}|^{1+2\alpha}} \quad as\; b \to 0, \; \omega \to \pi$$

$$\approx \frac{\sigma^2}{2\pi} \frac{4\alpha^2 + 8\alpha + 3}{6\alpha\, \pi^{1+2\alpha}}$$

As a result, we shall consider the following discrete approximation to (2.8), with $b \to 0$, in a neighbourhood of $\omega = \pi$:

$$(3.8) \; f_D(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|1-\theta e^{i\omega}|^{2\gamma}}, \quad 0 < |\theta| < 1, \; 0 < \gamma < 1.$$

The approximations (3.5) and (3.8) suggest the following model for LRD and intermittency:

$$(3.9) \qquad f_D(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|1-e^{i\omega}|^{2d}} \frac{1}{|1-\theta e^{i\omega}|^{2\gamma}},$$

$$d = \beta - \tfrac{1}{2}, \; 0 < |\theta| < 1, \; 0 < \gamma < 1, \; \omega \in [-\pi, \pi].$$

The spectral density (3.9) can be extended to include an $AR(p)$ component to model short-memory features of the time-series:

(3.10)

$$f_D(\omega) = \frac{\sigma^2}{2\pi} \frac{1}{|1-\theta_1 e^{i\omega} - \cdots - \theta_p e^{ip\omega}|^2} \frac{1}{|1-e^{i\omega}|^{2d}} \frac{1}{|1-\theta e^{i\omega}|^{2\gamma}},$$

which is the spectral density of the discrete process generated by the fractional difference equation:

$$(3.11) \; (1-B)^d (1-\theta B)^\gamma (1-\theta_1 B - \cdots - \theta_p B^p) X_t = \varepsilon_t,$$

where B is the backshift operator $BX_t = X_{t-1}$, and ε_t is white noise with variance σ^2.

4. PARAMETER ESTIMATION

This section gives a method to estimate the parameter vector $\Psi = (d, \theta, \gamma, p, \theta_1, \ldots, \theta_p, \sigma^2)$ of model (3.10). This is carried out in the following stages:

(i) Preliminary estimation of the LRD parameter d

Denote the spectral density of the series $Y_t = (1-B)^d X_t$ by $f_Y(\omega)$. Then

(4.1) $f_D(\omega) = \left|1-e^{i\omega}\right|^{-2d} f_Y(\omega)$, $\omega \in [-\pi, \pi]$,

which yields

(4.2)

$$\ln \hat{f}_D(\omega) = \ln f_Y(0) - d\ln\left|1-e^{i\omega}\right|^2 + \ln\frac{\hat{f}_D(\omega)}{f_D(\omega)} + \ln\frac{f_Y(\omega)}{f_Y(0)}$$

where $\hat{f}_D(\omega)$ is an estimator of $f_D(\omega)$. For a time series of length T, denote the Fourier frequencies by $\omega_j = \frac{2\pi j}{T} \in (0,\pi)$. Since the AR part in (3.10) can be assumed to have no roots inside the unit circle, i.e. $1-\theta_1 z - \cdots - \theta_p z^p \neq 0$ for $|z|<1$, we have $f_Y(\omega)$ smooth and bounded near $\omega = 0$. As a result, we can choose $m = m(T)$ such that $m/T \to 0$ as $T \to \infty$ and consequently $\ln\frac{f_Y(\omega_m)}{f_Y(0)} \to \ln 1 = 0$ as $T \to \infty$. With such a choice of m, we have from (4.2) and for sufficiently large T that

(4.3) $v_j \approx a - du_j + w_j$, $j = 1,2,\ldots,m$,

$$v_j = \ln \hat{f}_D(\omega_j), \qquad u_j = \ln\left|1-e^{i\omega_j}\right|^2 = \ln\left(4\sin^2\frac{\omega_j}{2}\right),$$

$$w_j = \ln\left(\hat{f}_D(\omega_j)/f_D(\omega_j)\right), \ a = \ln f_Y(0).$$

(ii) Preliminary estimation of the intermittency parameters θ and γ

In view of the form (3.11), we denote the spectral density of the series $U_t = (1-\theta B)^\gamma X_t$ by $f_U(\omega)$. Then

$$f_D(\omega) = \left|1-\theta e^{i\omega}\right|^{-2\gamma} f_U(\omega), \ \omega \in [-\pi,\pi],$$

which yields

(4.4)

$$\ln \hat{f}_D(\omega) = \ln f_U(\pi) - \gamma \ln\left|1-\theta e^{i\omega}\right|^2 + \ln\frac{\hat{f}_D(\omega)}{f_D(\omega)} + \ln\frac{f_U(\omega)}{f_U(\pi)}$$

Since $f_U(\omega)$ may be assumed smooth and bounded near $\omega = \pi$, we can choose $n = \left[\frac{T}{2+1/T}\right]$ so that $\ln\frac{f_U(\omega_n)}{f_U(\pi)} \to 0$ as $T \to \infty$. We then have from (4.4) that, for sufficiently large T,

(4.5) $v_j \approx b - \gamma\ln\left|1-\theta e^{i\omega_j}\right|^2 + w_j$, $j = n,\ldots,\left[\frac{T}{2}\right]$,

where

$$v_j = \ln\hat{f}_D(\omega_j), \ b = \ln f_U(\pi), \ w_j = \ln\frac{\hat{f}_D(\omega_j)}{f_D(\omega_j)}, \ \omega_j = \frac{2\pi j}{T}.$$

(iii) Estimation of the AR parameters

Initially, suppose that the model order p is known. We are led to minimise

(4.6)

$$\hat{\sigma}^2_{p,d,\theta,\gamma} = \frac{2\pi}{T}\sum_j I_X(\omega_j)\left|\Theta\left(e^{i\omega_j}\right)\right|^2\left|1-e^{i\omega_j}\right|^{2d}\left|1-\theta e^{i\omega_j}\right|^{2\gamma}$$

for $d, \theta, \gamma, \theta_1, \ldots, \theta_p$, where $\Theta(B) = 1-\theta_1 B - \ldots - \theta_p B^p$. For any d, θ and γ, an AR(p) model can be fitted to the inverse Fourier transform $I_X(\omega_j)\left|1-e^{i\omega_j}\right|^{2d}\left|1-\theta e^{i\omega_j}\right|^{2\gamma}$ using the Durbin-Levinson recursion. Since this recursion is very fast for each d, θ and γ, we can estimate d, θ and γ by a gradient-based optimisation procedure. Initial estimates of d, θ and γ are given by stages (i) and (ii) above. As the model order p is not known, it must be estimated from data. Following Anh and Kavalieris (1994), we base model selection on the concept of minimum description length due to Rissanen (1989). This leads to an approximation of the log likelihood by

(4.7) $$T\log\hat{\sigma}_{p,d,\theta,\gamma} + \left(p + \Delta_{d,\theta,\gamma}\right)\log T$$

where $\hat{\sigma}^2_{p,d,\theta,\gamma}$ is the prediction variance estimated in (4.6), $p + \Delta_{d,\theta,\gamma}$ is the number of model parameters.

Algorithm

Step 1. Taper the data after correcting for the mean. Append at least $p_{\max}+1$ zeroes to the data to obtain the series

$$X'_t = \begin{cases} X_t, & t = 1,\ldots,T \\ 0, & t = T+1,\ldots,T' \end{cases}$$

Step 2. Compute the periodogram

$$I_X(\omega_j) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} X'_t e^{i\omega_j}\right|^2, \ \omega_j = \frac{2\pi j}{T'}.$$

For efficient computation of the fast Fourier transform, T' should be a power of 2.

Step 3. Compute $\hat{d}_{LS}$ of (4.3) using least squares. Then assuming that $\gamma = 1$, use the Durbin-Levinson recursion to estimate the *AR(1)* parameter as a first estimate of $\hat{\theta}$ by computing

$$r_{\hat{d}}(k) = \frac{1}{T'}\sum_j I_X(\omega_j)\left|1-e^{i\omega_j}\right|^{2d} e^{ik\omega_j}, \ k = 0,1,$$

the covariances for the process $Y_t = (1-B)^{\hat{d}_{LS}} X_t$. Given the initial value of $\hat{\theta}$, an initial value for $\hat{\gamma}_{LS}$ can be obtained using the least squares estimate of (4.5).

Step 4. For any values of $\hat{d}, \hat{\theta}$ *and* $\hat{\gamma}$, compute

$$r_{d,\theta,\gamma}(k) = \frac{1}{T'} \sum_j I_X(\omega_j) \left|1 - e^{i\omega_j}\right|^{2d} \left|1 - \theta e^{i\omega_j}\right|^{2\gamma} e^{ik\omega_j}, \; k = 0, \ldots, p_{\max}$$

using the fast Fourier transform. Use the Durbin-Levinson recursion to estimate the *AR(p)* parameters $\theta_1, \ldots, \theta_p$ and prediction variance $\hat{\sigma}^2_{p,d,\gamma}$ from the covariances $r_{d,\theta,\gamma}(k)$, $k = 0, \ldots, p$. Select $\hat{p}_{d,\theta,\gamma}$ to minimise (4.7) for $p = 1, \ldots, p_{\max}$.

Step 5. Repeat *Step 4* for a series of values of (d, θ, γ) near $\left(\hat{d}_{LS}, \hat{\theta}, \hat{\gamma}_{LS}\right)$ and select $\left(\hat{d}, \hat{\theta}, \hat{\gamma}\right)$ to minimise

$$T \ln \hat{\sigma}^2_{\hat{p}_{\hat{d},\hat{\theta},\hat{\gamma}}, \hat{d}, \hat{\theta}, \hat{\gamma}} + \left(\hat{p}_{\hat{d},\hat{\theta},\hat{\gamma}} + \Delta_{\hat{d},\hat{\theta},\hat{\gamma}}\right) \log T.$$

This step is done using a downhill simplex method, a computationally easy and failsafe way of non-linear minimisation.

Step 6. After calculating the estimated values of (d, θ, γ), the value of α can be estimated from the non-linear equation $\frac{4\alpha^2 + 8\alpha + 3}{6\alpha\, \pi^{1+2\alpha}} = \frac{1}{\left(1 - 2\theta \cos \omega_j + \theta^2\right)^{\gamma}}$ using an algorithm that is a combination of the bisection method and the Newton-Raphson method. This hybrid technique takes a bisection step whenever the Newton-Raphson would take the solution out of the given bounds, or whenever the Newton-Raphson is not converging quickly enough.

5. NUMERICAL RESULTS

The wind tunnel at Monash University was commissioned to carry out simulations of concentrations to represent odours emanating from a piggery installation for a number of source characteristics. Two of the chosen source configurations were an area source (typically 100 x 100 m) representing around slurry spreading, waste ponds or large sheds, located in flat terrain, and a tall point source unaffected by turbulence from nearby buildings. The time series for these two source configurations for a downwind distance of 1000m are shown in Figure 1. The dominant characteristics obvious from these plots are the presence or absence of meandering and the intermittent nature of the concentrations. From Figure 2, which shows the autocorrelation function up to a lag of 150, it appears that a model including long-range dependence would be appropriate. In order to quantify long range dependence and intermittency, we use the method outlined in the previous sections.

The results of running the algorithm on the two simulated data series is shown in Table 1. Figure 3 shows the log of the spectral density of the raw point source data and compares this to the log of the spectral density after the removal of the estimated model. Figure 4 shows the log of the spectral density of the raw area source data and compares this to the log of the spectral density after the removal of the estimated model.

6. CONCLUSIONS

This paper introduced a new class of stochastic models based on established theories to represent time series which exhibit long-range dependence and intermittency simultaneously. An efficient iterative method was developed to estimate all the model parameters from observed data. These models can be used directly for prediction purposes. The numerical results on two odour time series with quite different degrees of intermittency indicated that the models are fully implementable and useful in describing turbulent time series. Detailed statistical results and an extensive simulation study confirming the power of the method are reported in Anh *et al.* (1995) and Lunney and Anh (1995).

7. REFERENCES

Anh, V.V. and L. Kavalieris, Long-range dependence in models for air quality, in *Statistics in Ecology and Environmental Monitoring*, D.J. Fletcher and B.F.J. Manly (Eds.). Dunedin: University of Otago Press, pp 100-209, 1994.

Anh, V.V., K.E. Lunney and S. Peiris, Statistical analysis of time series with long-range dependence and intermittency (submitted to *J. Forecasting*).

Arneodo, A., F. Argoul, E. Bacry, J. Elezgaray, E. Freysz, G. Grassea, J.F. Muzy and B. Pouligny, Wavelet transform of fractals, in *Wavelets and Applications*, Meyer, Y. (Ed.), Masson, pp. 286-352, 1992.

Beran, J., Statistical models for data with long-range dependence. *Statistical Science* 4, 404-27, 1992.

Berman, S.M., Gaussian sample functions: Uniform dimension and Hölder conditions nowhere. *Nagoya Math. J.* 46, 63-86, 1972.

Davis, A., A. Marshak and W. Wiscombe, Wavelet-based multifractal analysis of non-stationary and/or intermittent geophysical signals, in *Wavelets in Geophysics*, Foufoula-Georgiou, E. and P. Kumar (Eds.), Academic Press, 1994.

Farge, M., Wavelet transforms and their applications to turbulence. *Annu. Rev. Fluid Mech.* 24, 395-457, 1992.

Lunney, K.E. and V.V. Anh, Time series with long-range dependence and intermittency: A simulation study (submitted to *Communications in Statistics*).

Rissanen, J., *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.

8. ACKNOWLEDGEMENTS

The authors wish to thank Chris Watson and Peter Best of Katestone Scientific for computing assistance and general comments respectively. The wind tunnel simulations were carried out by Monash University as part of a project sponsored by the Environment Protection Authority of New South Wales. The work is partially supported by ARC grant A69531724.

Table 1

Data	$\hat{d}$	$\hat{\theta}$	$\hat{\gamma}$	$\hat{\alpha}$	$\hat{p}$	$\hat{\theta}_j, j = 1,\ldots,\hat{p}$	$\hat{\sigma}^2$
Area Source	0.7829	0.2042	0.1021	0.1672	5	1.1436 -0.2911 -0.1605 0.1185 -0.0292	2.97×10^{-6}
Point Source	0.2089	0.7222	0.3810	0.2388	11	0.3260 0.2339 0.1210 -0.0062 0.01882 0.0816 -0.0015 0.0351 -0.0032 -0.0012 0.0626	8.65×10^{-9}

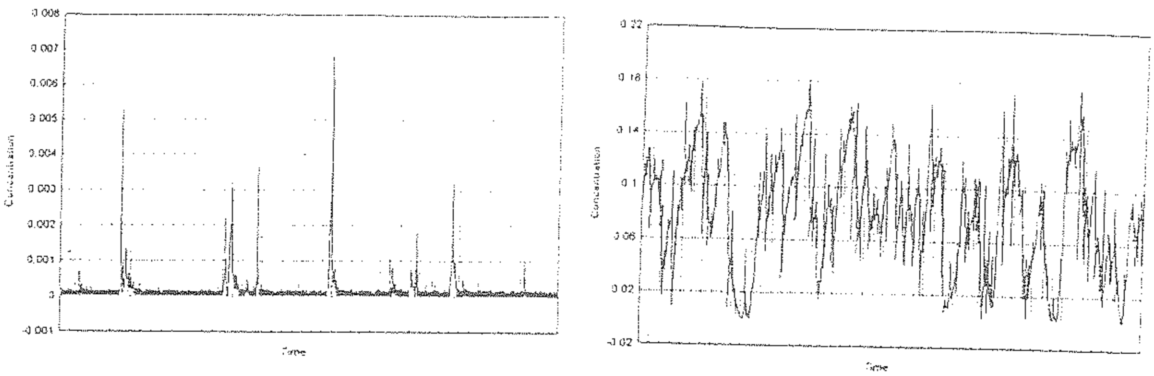

Figure 1: Time series plots of simulated concentration data from a wind tunnel. The two series are meant to represent odour from a piggery installation at 1000 m downwind on the centreline and having source configurations of (a) a point source stack and (b) an area source.

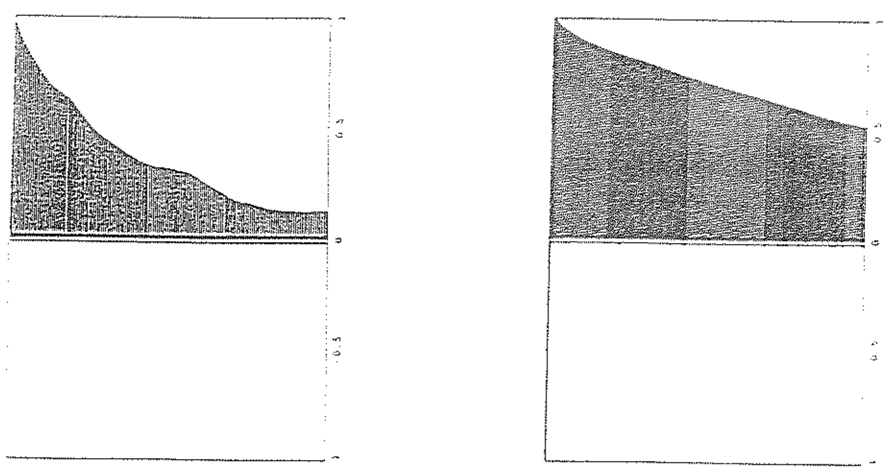

Figure 2: Auto correlation function for the data series (a) a point source 1000 m downwind of a point source and (b) 1000 m downwind of an area source.

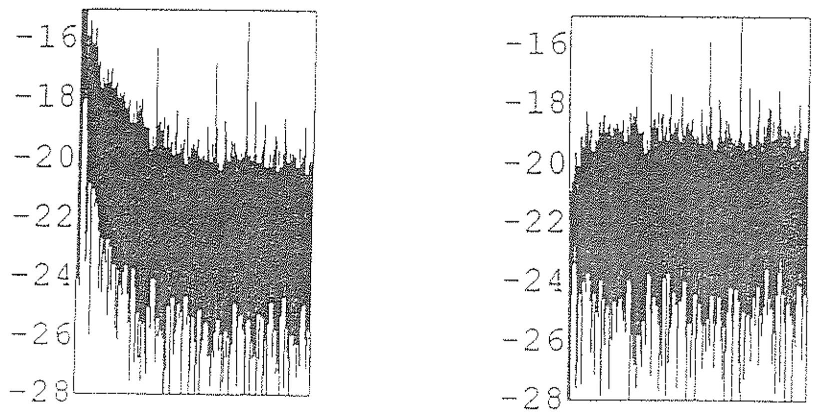

Figure 3: Spectral density plots for the series representing concentrations measured 1000 m downwind of a point source with (a) the log of the spectral density of the raw data, (b) the log of the spectral density of the raw series after removal of the estimated model.

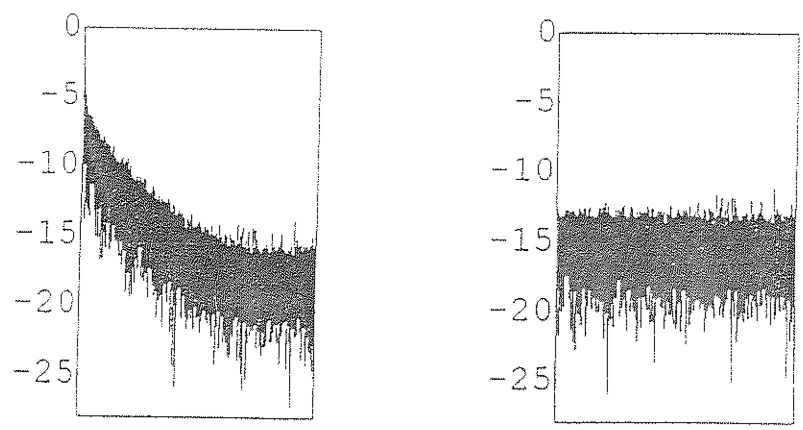

Figure 4: Spectral density plots for the series representing concentrations measured 1000 m downwind of an area source with (a) the log of the spectral density of the raw data, (b) the log of the spectral density of the raw series after removal of the estimated model.