Accuracy analysis of measurements on a stable power-law distributed series of events

2006 J. Phys. A: Math. Gen. 39 13967

(http://iopscience.iop.org/0305-4470/39/45/007)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **39** (2006) 13967–13982 doi:10.1088/0305-4470/39/45/007

Accuracy analysis of measurements on a stable power-law distributed series of events

J O Matthews[1], **K I Hopcraft**[1,3], **E Jakeman**[1,2] **and G B Siviour**[1]

[1] School of Mathematical Sciences, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

[2] School of Electrical and Electronic Engineering, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

Received 8 May 2006, in final form 9 May 2006
Published 24 October 2006
Online at stacks.iop.org/JPhysA/39/13967

Abstract

We investigate how finite measurement time limits the accuracy with which the parameters of a stably distributed random series of events can be determined. The model process is generated by timing the emigration of individuals from a population that is subject to deaths and a particular choice of multiple immigration events. This leads to a scale-free discrete random process where customary measures, such as mean value and variance, do not exist. However, converting the number of events occurring in fixed time intervals to a 1-bit 'clipped' process allows the construction of well-behaved statistics that still retain vestiges of the original power-law and fluctuation properties. These statistics include the clipped mean and correlation function, from measurements of which both the power-law index of the distribution of events and the time constant of its fluctuations can be deduced. We report here a theoretical analysis of the accuracy of measurements of the mean of the clipped process. This indicates that, for a fixed experiment time, the error on measurements of the sample mean is minimized by an optimum choice of the number of samples. It is shown furthermore that this choice is sensitive to the power-law index and that the approach to Poisson statistics is dominated by rare events or 'outliers'. Our results are supported by numerical simulation.

PACS numbers: 02.50.−r, 05.40.−a, 05.45.Tp, 06.20.Dk

1. Introduction

Complex systems exhibiting scale-free behaviour are ubiquitous in nature. Fractal and self-affine systems have been studied extensively for many decades following recognition of the existence of objects that manifest hierarchical structure. In the context of continuous random

[3] Author to whom any correspondence should be addressed.

media, power-law spectra are commonly used to characterize multi-scale phenomena such as turbulence and rough surfaces. More recently interest has focused on probability densities with power-law tails, such as the Lévy stable distributions that can be used to characterize the connectivity of natural and fabricated scale-free networks. These range from genome, proteome and other symbiotic interactions to the topology of the www and economic and social interactions [1].

The characterization of random processes exhibiting power-law behaviour raises interesting questions regarding the accuracy of parameter estimation from finite data sets. For example, it has been shown that the accuracy of measurements of the spectral power-law of a Gaussian random fractal does not necessarily improve inversely as the number of samples but at a slower rate depending on the power-law index [2]. This means that the characterization of objects such as rough surfaces, which are known to exhibit multi-scale behaviour, is subject to larger errors than might have been anticipated.

A related problem that will be investigated in the present paper is encountered specifically in measurements on systems characterized by heavy tailed single-interval distributions such as the Lévy stable class [3]. Properties of these distributions such as variance, autocorrelation function and higher order moments are ill defined. Although in practice these quantities will always be finite for a finite data set, they increase with time without limit due to the slow decay of the distribution function tail. Classification has therefore to date largely relied on direct distribution measurements [4, 5]. However, the validity of such measurements will also be constrained by the length of the data set and more particularly by the dynamic range imposed in the measuring process, which may well be unknown.

Many systems characterized by power-law single interval distributions can legitimately be modelled using a mean-field approximation [6] where the size of the system is assumed to be sufficiently large so that it can be approximated by a continuum. However, some authors have noted that the discrete nature of certain processes is important, both in terms of formulation [7] and effect [8], with quantization incorporating intrinsic noise into the models in a natural way. With this in mind we have previously developed a stochastic model that generates a population of individuals that can have fluctuations with a Lorentzian spectrum but governed by a single interval distribution that belongs to the discrete-stable class [9]. Allowing individuals to escape from this population in a manner similar to the emission of photons from a cavity [10] forms a time series of events that can be monitored through counting the number of emigration events n in the time intervals T [11]. The probability distribution of the n counts of this external series of events, $p_n(T)$, possesses a similar power-law behaviour to the population that produces them and the associated moments and correlation functions are undefined [9]. This provides a model that can be used to explore the measurement problems outlined above, being at the same time analytically tractable and capable of numerical simulation.

In order to limit the excursions of the data generated by this model we adopt a nonlinear signal-processing technique known as 'clipping' [12–15] which regularizes the power-law fluctuations. An extreme form of clipping that has proved useful in other contexts [16] transforms the process into a 1-bit data stream. In practice, this procedure might be adopted to overcome the problem of an unknown detector response to the large fluctuations characterizing discrete-stable fluctuations. The resulting data stream is easy to analyse theoretically and previous work has shown that it retains sufficient information to characterize the original process at a relatively small cost in statistical accuracy. Moreover, since the moments and correlation functions of the 1-bit data stream are finite, they provide a *new* means for characterizing discrete-stable processes. The principal novel content of this paper is a statistical accuracy analysis of power-law data that have been processed in this way.

Section 2 provides a brief overview of the population model, formulation of the monitoring scheme, implementation of the clipping technique and subsequent construction of the clipped mean. Section 3 revisits calculation of the variance of a sampled mean for the well-known thermal model of photon counting statistics and also the equivalent clipped process, revealing the existence of an optimum number of samples for a given experiment duration. Section 4 extends this analysis to the case of a series of events generated by a discrete-stable process. In particular, we present the results of a calculation of the variance of an estimate of the clipped mean. Results and conclusions are contained in section 5.

2. The population model

We consider a population model whose dynamics are governed by deaths that occur at rate μ in proportion to the instantaneous population size N and immigrants that enter the population in multiples with rates α_m, particular to their order m. The rate equation for this process is

$$\frac{\mathrm{d}P_N(t)}{\mathrm{d}t} = \mu(N+1)P_{N+1} - \mu N P_N - P_N \sum_{m=1}^{\infty} \alpha_m + \sum_{m=1}^{N} \alpha_m P_{N-m}. \tag{1}$$

This describes the evolution of $P_N(t)$, the probability that the population comprises N members at time t. The multiple immigration rates allow the population model to be 'tuned' to adopt specific behaviours. In particular, the following choice leads to a discrete-stable process:

$$\alpha_m = \frac{a\Gamma(m-\nu)}{\Gamma(-\nu)m!} \qquad 0 < \nu \leqslant 1. \tag{2}$$

Solving the rate equation (1) with the help of the generating function

$$Q(s;t) = \sum_{n=0}^{\infty} (1-s)^n P_N(t) \tag{3}$$

leads to the solution [9]

$$Q(s;t) = (1 - f(s;t))^M \exp\left(-\frac{as^\nu}{\nu\mu}(1 - \exp(-\nu\mu t)\right) \tag{4}$$

where

$$f(s;t) = s\exp(-\mu t).$$

Here a is a positive real constant and the process is initiated with M individuals present. In the large time limit (4) reduces to the generating function for the class of discrete-stable distributions:

$$Q(s) = \exp\left(-\frac{as^\nu}{\nu\mu}\right) \equiv \exp(-As^\nu). \tag{5}$$

For the case $\nu = 1$, (5) is the generating function for a Poisson distribution: the equilibrium distribution for a process comprising deaths and only single immigrants. Furthermore, the scaling factor A becomes the *mean* in this case. According to (3), the probability distribution and factorial moments are both derived from the generating function by differentiation,

$$P_N = P_N(t\to\infty) = \frac{(-1)^N}{N!}\left.\frac{\mathrm{d}^N Q(s)}{\mathrm{d}s^N}\right|_{s=1} \tag{6}$$

$$\frac{\langle N(N-1)(N-2)\ldots(N-r+1)\rangle}{\langle N\rangle^r} = \left(\frac{-1}{\langle N\rangle}\frac{d}{ds}\right)^r Q(s)\Big|_{s=0} \tag{7}$$

and so the solution (5) implies that if $N \gg 1$, $P_N \sim 1/N^{1+\nu}$, giving an inverse power-law asymptote with index lying between one and two. Thus the moments of these distributions are theoretically infinite for values of ν in the allowed range except for the Poisson case $\nu = 1$. It is readily shown from the conditional solution (4) that the corresponding correlation functions are also ill defined.

A series of events is now generated by allowing individuals to leave the population at a rate η that is proportional to the number present, leading to an increased death rate within the population of $\bar{\mu} = \mu + \eta$. This process has an analogue in quantum optics [18] where inferences about the population inside a laser cavity, for example, can be made from samples of the series of photon events registered during a finite time due to light leaving the cavity. It will be assumed here that the emigration events are sampled by counting the number $n(t, T)$ in contiguous time intervals of duration T (the 'integration' time). This utilizes the maximum amount of data that is available although there will in general be correlation between samples.

The probability of there being N individuals within the population and recording n emigrations from it within an integration time, $p_{N,n}(T)$, can be determined using the joint generating function for the monitored population

$$Q_c(s, z; T) = \sum_{N,n=0}^{\infty} (1 - s)^N (1 - z)^n p_{N,n}(T).$$

This satisfies the partial differential equation [9]

$$\frac{\partial Q_c}{\partial t} = (\eta z - \bar{\mu} s)\frac{\partial Q_c}{\partial s} - a\nu s^\nu Q_c. \tag{8}$$

The full solution of this equation was derived in [9] and is reproduced in the appendix. For present purposes, interest resides in the marginal distribution $p_n(T)$ of counting n external events within an integration time T. This is constructed from the generating function by setting $s = 0$ in the solution of (8):

$$Q_c(0, z; T) = \exp\{-Az^\nu[\eta(1 - \mathrm{e}^{-\bar{\mu}T})/\bar{\mu}]^\nu F(1, \nu; \nu + 1; 1 - \mathrm{e}^{-\bar{\mu}T})\} \tag{9}$$

where $F = {}_2F_1$ is the hypergeometric function [19]. The z-dependence of this quantity is similar to the s-dependence in (5) and so the distribution of the monitored population also has a power-law tail for all integration times T,

$$p(n; T) \approx Cn^{-(\nu+1)}, \tag{10}$$

and all its moments are infinite. Furthermore, it is possible to calculate the joint distribution of finding n counts in an integration time T centred at time t and n counts in an identical interval at a later time $t + \tau$(see appendix of [9]). These joint probabilities again have power-law tails implying that all joint statistics such as autocorrelation functions are infinite. Consequently, there is no moment-based method for classifying raw power-law data generated by this model.

Numerical simulation shows that individual realizations of the process described above exhibit intermittent bursts of large numbers of emigrants from the population and this is the source of the heavy tailed distributions and infinite moments. However, any detection system will have a finite dynamic range (due, for example, to its finite response time or to saturation of the detector) and it may not correctly register large excursions in $n(t; T)$. Indeed, the consequent distortion of the data may be unknown. One way to ameliorate this problem and at the same time regularize the measured statistics is to apply a *known* operation on the data to eliminate regimes where the response of the detection system is suspect. The simplest such processing technique is 'clipping' or 'limiting' [12], in which the number of counts in an integration time is replaced by one or zero according to whether it lies above or below a pre-determined level. Hard limiting [13] is the simplest but most severe form of clipping where the clipping level is set equal to zero:

$$c_j(T) = \begin{cases} 0 & n(t_j, T) = 0 \\ 1 & n(t_j, T) \geqslant 1. \end{cases} \tag{11}$$

Here $n(t_j, T)$ is the sampled time series, $c_j(T)$ is the new clipped time series and $t_j = jT$. The generating function for the clipped distribution is particularly simple,

$$Q_{\mathrm{cl}}(z; T) = \sum_{n=0}^{1} (1-z)^n p_n(T) = p_0(T) + (1-z)(1-p_0(T)). \tag{12}$$

Noting that $p_0(T) = Q_c(0, 1; T)$, the clipped mean can be written as a function of the integration time from (11) and (12) as follows:

$$\bar{c}(T) = 1 - p_0(T) = 1 - Q_c(0, 1; T) \tag{13a}$$

$$= 1 - \exp\{-A[\eta(1 - \mathrm{e}^{-\bar{\mu}T})/\bar{\mu}]^{\nu} F(1, \nu; \nu + 1; 1 - \mathrm{e}^{-\bar{\mu}T})\} \tag{13b}$$

It is shown in [9] that both the clipped mean and autocorrelation function exhibit power-law behaviour with T when the integration time is much less than the correlation time of the population fluctuations. In particular, the clipped mean *does not scale linearly with the integration time*, but rather has the asymptotic behaviour

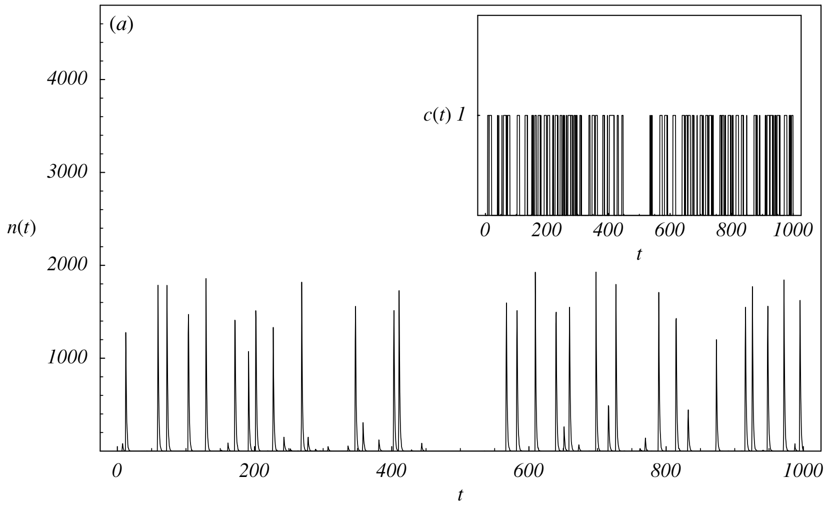
$$\bar{c}(T) \approx A(\eta T)^{\nu}; \qquad \bar{\mu}T \ll 1. \tag{14}$$

The power-law scaling with T manifest in (14) despite the loss of amplitude information is a consequence of the fact that the power-law fluctuations in the *size* of the original population are transferred to the *dispersion* of emigration events in time. It is possible to construct realizations of this population process using techniques detailed in [20]. Figure 1(*a*) shows a realization of the sampled emanations from the discrete-stable process with $\nu = 0.1$. Note the large infrequent fluctuations. The inset figure shows the equivalent clipped time series for this realization and serves to highlight the intermittent bursts of events produced by a process with small index ν. Figure 1(*b*) shows a realization for the same process but with $\nu = 0.8$. For this realization, the fluctuations are far more frequent but their magnitude is smaller, which is highlighted by the inset figure showing the very 'busy' clipped signal.

It follows from (14) that ν is the slope of the line obtained by plotting the logarithm of the clipped mean against the logarithm of the sample time. In practice, this would be deduced from estimates of the clipped mean constructed over a finite experiment time and is subject to error due to the random nature of the process. However, it does provide a simple alternative method for estimating the power law characterizing the process and the remainder of this paper will be concerned with the accuracy of such measurements.

3. Accuracy of estimates of the mean of a birth–death–immigration process

Before considering the problem of measurement accuracy in the case of the stably distributed discrete series of events of interest here, we shall briefly review the standard calculation for a birth–death–immigration process [21], including the effect caused by clipping. This will revise the approach in terms of a well-understood train of events, establish the notation and enable the special features encountered in the case of stable processes to be highlighted. As indicated above, we are concerned here with one of the simpler statistical properties of the data, namely, its mean value as a function of integration time T. There is, of course, a difference



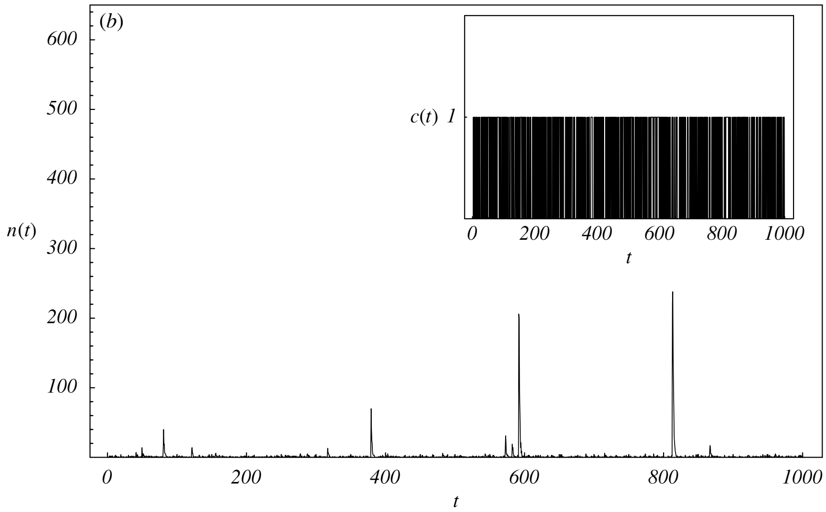


Figure 1. (*a*) A realization of sampled emigrations for the discrete-stable population model, with $a = 1, \eta = 1, \bar{\mu} = 1, \nu = 0.1$ and $T = 0.5$. Note the large fluctuations in population size and the intermittent events. An outer scale of 2000 was imposed on the simulation. (*b*) A realization of the same process but with $\nu = 0.8$.

between the true mean and the sample mean acquired in an experiment time of finite duration. The sample mean is an estimate that may be biased and the values of a series of estimates will be scattered around the true mean with a variance that depends on the experiment time and other factors. It is an experimenter's objective to minimize this error by optimizing the measurement parameters under his control.

Consider a process with mean $\bar{n}$, an estimator for which is the sample mean

$$\hat{n}(T) = \frac{1}{N}\sum_{j=1}^{N} n(t_j, T) \equiv \frac{1}{N}\sum_{j=1}^{N} n_j. \tag{15}$$

In (15) the time at which each sample is taken is $t_j = jT$ and $NT = t_{\text{exp}}$ is the experiment time. The estimate $\hat{n}(T)$ is unbiased because $\langle \hat{n} \rangle = \bar{n}$. We can now construct the variance of the sample mean

$$Var\hat{n}(T) \equiv \delta^2(\hat{n}) = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}[\langle n_j n_k\rangle - \bar{n}^2]. \tag{16}$$

This can be simplified further by exploiting the symmetry properties of the correlations and noting that the diagonal terms are identical, whereupon equation (16) becomes

$$\frac{\delta^2(\hat{n})}{\bar{n}^2} = \frac{\delta_n^2}{N\bar{n}^2} + \frac{2}{N\bar{n}^2}\sum_{k=1}^{N-1}\left(1-\frac{k}{N}\right)[R(kT,T)-\bar{n}^2] \tag{17}$$

where $R(kT, T)$ is the autocorrelation function of n and δ_n^2 is the variance of the estimate (15). Note that this is a function of the number of samples N, the integration time T and the correlation time of the process.

For the case of a birth–death–immigration population model we can use the following well-known results for the mean, variance and autocorrelation function (see, for example, [17]):

$$\bar{n} = rT \tag{18a}$$

$$\delta_n^2 = \bar{n} + \frac{\bar{n}^2}{\beta}\left(\frac{1}{\gamma} + \frac{\exp(-2\gamma)}{2\gamma^2} - \frac{1}{2\gamma^2}\right) \tag{18b}$$

$$R(t,T) = \bar{n}^2\left(1 + \frac{1}{\beta}\left(\frac{\sinh(\gamma)}{\gamma}\right)^2 \exp[-(\bar{\mu}-\lambda)t]\right) \tag{18c}$$

$$\gamma = \frac{(\bar{\mu}-\lambda)T}{2}, \qquad t \geqslant T. \tag{18d}$$

Here r is the average rate at which events occur and β is the cluster parameter defined as the ratio of the immigration rate to the birth rate, $\bar{\mu}$ is the population death rate enhanced by emigration as discussed in the last section and λ is the birth rate. We have shown in a previous paper that these results are structurally identical to those for the death-*multiple immigration* process with geometrically distributed immigrant groups [16]. Substituting the above into equation (17), performing the sum and expressing the total measurement time in terms of the integration time T, $t_{\text{exp}} = NT$ gives the result

$$\frac{\delta^2}{\bar{n}^2} = \frac{1}{t_{\text{exp}}}\left(\frac{1}{r} + \frac{2}{\beta(\bar{\mu}-\lambda)}\right). \tag{19}$$

Note that this formula depends only on the *product* of the number of samples and the integration time, i.e. the total experiment time. Otherwise it is a function of parameters of the population process that are beyond the control of the experimenter. Thus the error on a measurement of the mean can only be reduced by increasing the length of the data run and not by changing the way that the data are sub-divided into its constituent integration periods. This is a consequence of *all* events being recorded for contiguous sampling and the fact that in the case of a simple birth–death–immigration process the mean (18*a*) increases *linearly* with integration time.

It is not difficult to generalize the above calculation to the case where the data samples are not contiguous but separated by fixed periods. The number of samples within the experiment time is reduced in this case and the error increases because information has been discarded.

The same approach can be used to evaluate the variance of an unbiased estimate $\hat{c}(T)$ of the mean $\bar{c}(T)$ of the equivalent clipped process. The diagonal terms of the correlation matrix

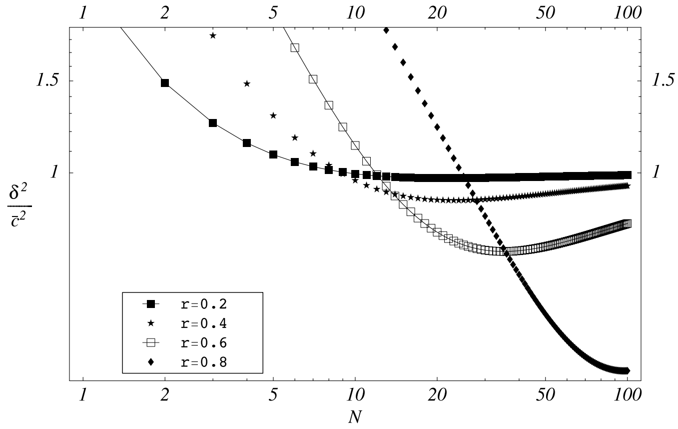


Figure 2. Relative variance of estimates of the clipped mean of a birth–death–immigration population plotted against sample number for the event rates shown and fixed experiment time. The graphs are normalized against the (constant) relative variance of an estimate of the true mean.

are equal to the variance of the clipped mean. However, $c^2(t, T) = c(t, T)$ since the clipped counts are either 0 or 1. Hence the error is given by

$$\frac{\delta^2(\hat{c})}{\bar{c}^2} = \frac{1}{N\bar{c}} + \frac{2}{N\bar{c}^2}\sum_{k=1}^{N-1}\left(1 - \frac{k}{N}\right)[\zeta(kT, T) - \bar{c}^2]. \tag{20}$$

This result replaces (17) for any series of events clipped at zero as in equation (11). As indicated in equation (13), the clipped mean is simply related to the generating function for the original series of events. Expressions for both this and $\zeta(t, T)$, the clipped autocorrelation function, are given in the appendix for the birth–death–immigration process. Figure 2 shows a plot of the relative variance of measurements of the clipped mean *normalized* by that of measurements of the true mean, equation (19), for contiguous sampling of a birth–death–immigration population model. The graphs are plotted as a function of the number N of samples taken in a fixed experiment time equal to 10 correlation times (i.e. $10/(\bar{\mu} - \lambda)$; see equations (18*c*), (18*d*) for various values of the count rate r. The variance of the clipped mean exhibits a minimum, converging towards the unclipped result when the N is large because the number of events in each integration time is small in this limit and clipping has little effect. On the other hand, as $N \to 1$, $(T \to t_{\text{exp}})$ the clipped mean eventually saturates at unity so more information is lost leading to an increasing relative error. The minimum of the curve is deeper and occurs at larger values of N when the count rate is large. This is because the clipped mean increases more rapidly to unity with increasing integration time in this case with an associated decrease in the first of the two contributions appearing on the right-hand side of equation (20). The location of the minimum value of the variance is determined by the balance between these contributions. Note that in the special case of a Poisson process where the birth rate is zero $(\beta \to \infty)$ the second contributions in (19) and (20) vanish and the relative variance curve for the clipped process exhibits no minimum as the number of samples is reduced but only a monotonic increase above the value taken by the original process at large values of N. It is not difficult to deduce that in this case the ratio of results (20) to (19) is $\bar{n}/[1 - \exp(-\bar{n})]$ with $\bar{n} = rT = rt_{\text{exp}}/N$.

It must be emphasized that although under certain conditions the relative variance of the clipped mean is less than that of the true mean this does not imply that measurement accuracy

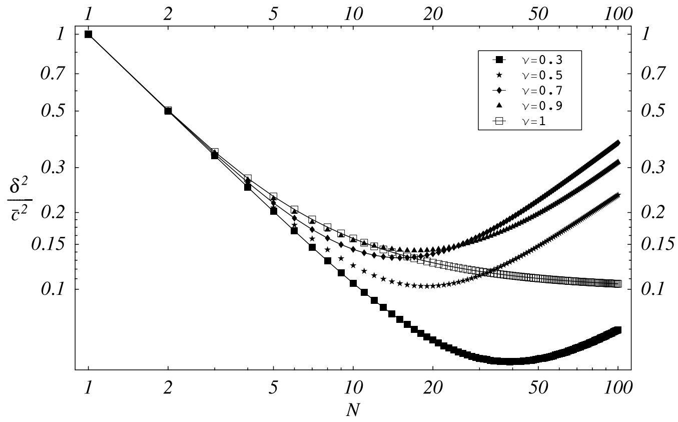


Figure 3. The variance of estimates of the clipped mean of a discrete-stable process as a function of the number samples, for $a = 1$, $\bar{\mu} = 1$ and the values of ν shown. The experiment time $t_{\text{exp}} = 10$.

can be improved by clipping. This is because, after clipping, an estimate of the true mean can only be deduced through formula (13*a*) that relates the clipped mean to the true mean, namely the probability of finding no events in the time interval T. It is readily shown from this formula that the relative variance of estimates of the true mean is always increased by clipping.

4. Accuracy of estimates of the clipped mean of a discrete-stable process

The clipped result (14) provides a method for determining ν based on a measurement of the clipped mean as a function of the integration time T. For example, a simple two-point fit gives $\hat{\nu} = \ln(\hat{c}_1/\hat{c}_2)/\ln(T_1/T_2)$. According to the calculation of small variations, the variance of the index value obtained in this way is proportional to the *relative* variance of the clipped mean:

$$(\delta\hat{\nu})^2 \propto \frac{(\delta\hat{c})^2}{\bar{c}^2}. \tag{21}$$

Figure 3 shows the relative variance of the clipped mean computed from equation (20) for a fixed experiment time as a function of the number of samples N and for a selection of values for the index of the power-law ν. Result (13*b*) has been used in this calculation together with the correlation function for the clipped process derived in [9] and quoted in the appendix. Just as in the case of the clipped birth–death–immigration process (figure 2) the curves exhibit minima for specific values of N or, equivalently, integration time T. If the number of contiguous samples is increased beyond the minimum in figure 3, however, the relative error in measurements of the clipped mean increases without limit for fixed experiment time. This is because improved resolution reveals more events and clipping continues to discard information in this limit unlike the case of the clipped birth–death–immigration process discussed in section 3. The relative variance of an estimate of the clipped mean in that case converged to the *finite* unclipped result when the samples contain so few events that clipping had little effect (figure 2).

Another feature of figure 3 is the crossover of curves relating to different values of ν, implying that processes with different indices may display the same spread of estimates in an experiment of finite duration. This effect can be attributed to the form of the normalized

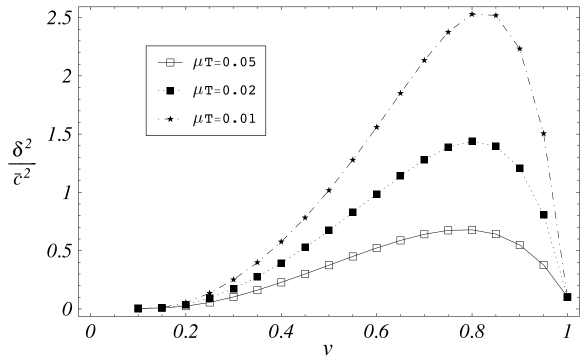


Figure 4. First channel of the normalized clipped autocorrelation function ($k = 1$) as a function of the index of power-law index for the values of $\bar{\mu}T$ shown ($a = 1$, $\bar{\mu} = 1$).

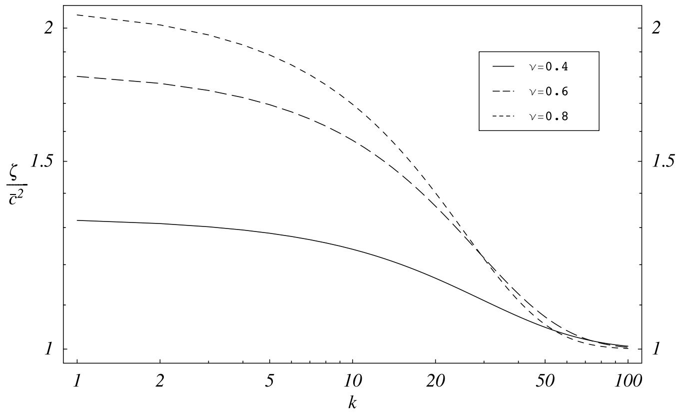


Figure 5. Normalized clipped autocorrelation function as a function of delay time for the values of the power-law index shown.

clipped autocorrelation function ζ that appears in equation (20). For all integration times, this is given at $k = 1$ by

$$\frac{\zeta(T, T)}{\bar{c}^2} \approx 1 + \frac{\bar{\mu}\nu}{a}(2 - 2^\nu)(\eta T)^{-\nu}. \tag{22}$$

Figure 4 shows that, for all integration times, the degree of correlation at $k = 1$ increases from its value for small power-law index to a maximum near $\nu = 0.8$ and then decreases again as the index approaches unity. Figure 5 shows $\zeta/\bar{c}^2$ as a function of k. As ν increases towards the value of 0.8, the degree of correlation $\zeta(kT, T)$ increases at small delay times but decays more rapidly for large values. Thus at small k the curve for $\nu = 0.6$ is below than that for $\nu = 0.8$; however, the correlation between samples for the smaller index persists to higher values of k. In fact for $\nu \sim 0.1$ (not shown) there is no discernable change in the clipped autocorrelation until $k \sim 2000$. Conversely, for larger indices ($\nu \sim 0.8$) the correlation function falls off quite quickly beyond $k \sim 10$. This behaviour is manifest in the realizations shown in figure 1, where part (a) suggests large but infrequent events that result in weak and long lasting correlations whereas part (b) indicates small but frequent fluctuations, which cause strong and short-lived correlations. It is the competition between these properties

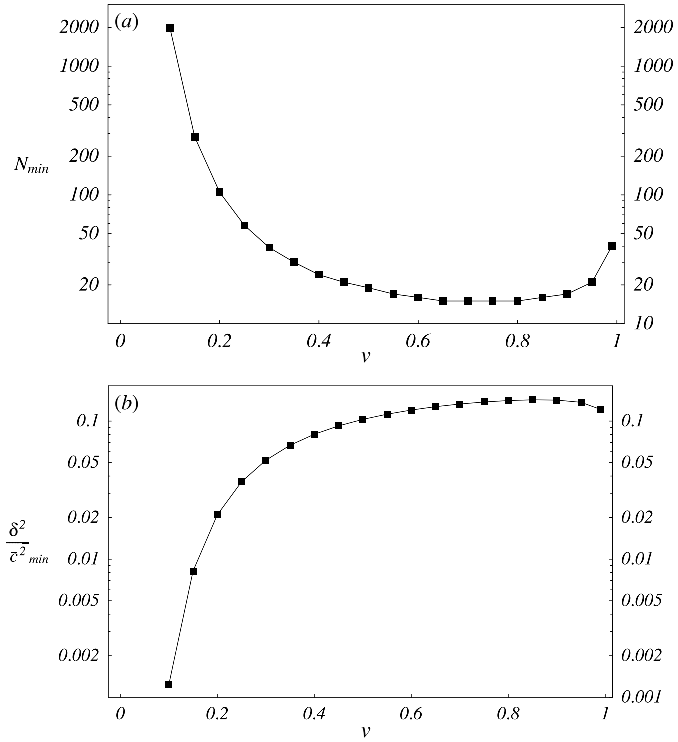


Figure 6. (*a*) Number of samples required to minimize the variance of measurements of the clipped mean as a function of the power-law index. (*b*) Value of the minimum error produced by optimum sampling for different values of the power-law index.

in the second term of equation (20) that determines the positions of the minima of the relative error curves displayed in figure 4 and the crossover between curves with different values of the power-law index.

Figure 6(*a*) shows the number of samples required to minimize the variance of measurements of the clipped mean as a function of the index of the distribution. It indicates that a greater number of samples are required to produce the minimum error for small indices compared with those close to 0.8, which require relatively few. The weak but persistent correlations for small values of ν, as opposed to the stronger shorter-lived correlations of larger indices ($\nu \sim 0.8$), account for this difference. Figure 6(*b*) displays the value of the minimum error produced by sampling at that optimum number as a function of the index. Clearly a longer experiment time is required to achieve the same accuracy in a measurement of the clipped mean for the larger values than for the smaller values of ν. Figure 7 shows plots of the un-normalized variance of estimates of the clipped mean measured with optimum data resolution (integration time/number of samples) as a function of experiment time. These decrease more slowly than linear and are reminiscent of the results obtained for measurements on fractal surfaces [2].

Figure 8 shows the normalized variance of the clipped mean for values of ν close to 1 and for $\nu = 1$. It is clear that the curves evolve smoothly to the Poisson case in accord with the behaviour of $N_{\rm min}$ in figure 6(*a*). As noted previously (section 3), in that case there is no minimum in the relative error versus N curve because clipping has no effect when N is

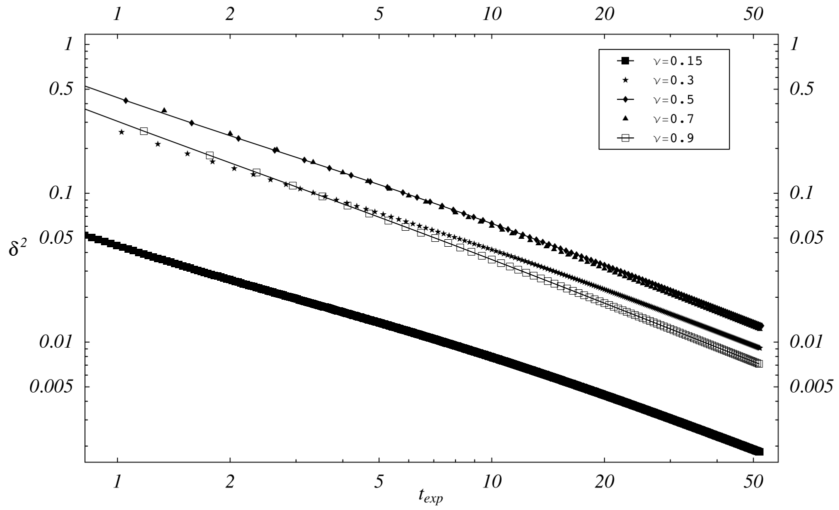


Figure 7. Relative variance of estimates of the clipped mean in the case of optimal sampling as a function of experiment time for the values of the power-law index shown.

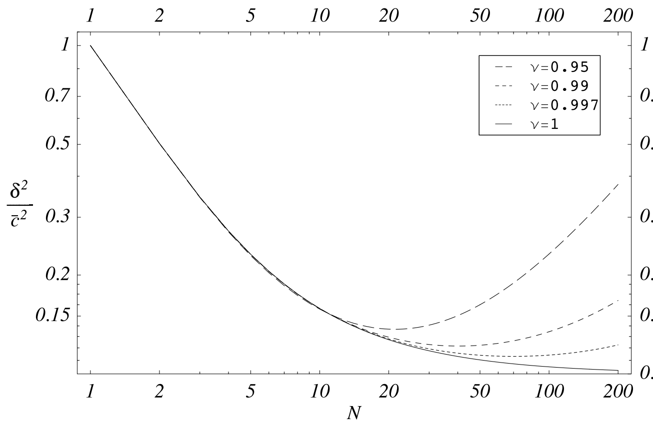


Figure 8. Relative variance of estimates of the clipped mean for values of ν close to 1 and for $\nu = 1$.

sufficiently large and the error in the clipped mean saturates at that of the mean of the original series of events.

From an experimental perspective, monitoring stable processes with indices close to unity could prove problematic, for if insufficient samples were taken, it would not be possible to distinguish a discrete-stable process from a Poisson process. One might be tempted however to argue that a process with $\nu \approx 1$ *is* effectively a Poisson process. Figure 9(*a*) shows the Poisson probability distribution [22] together with two examples of discrete-stable probability distributions that have indices very close to 1. It is apparent that these discrete-stable distributions are indistinguishable from a Poisson distribution up to the point that their power-law tails become established. Such a process would consist largely of small fluctuations

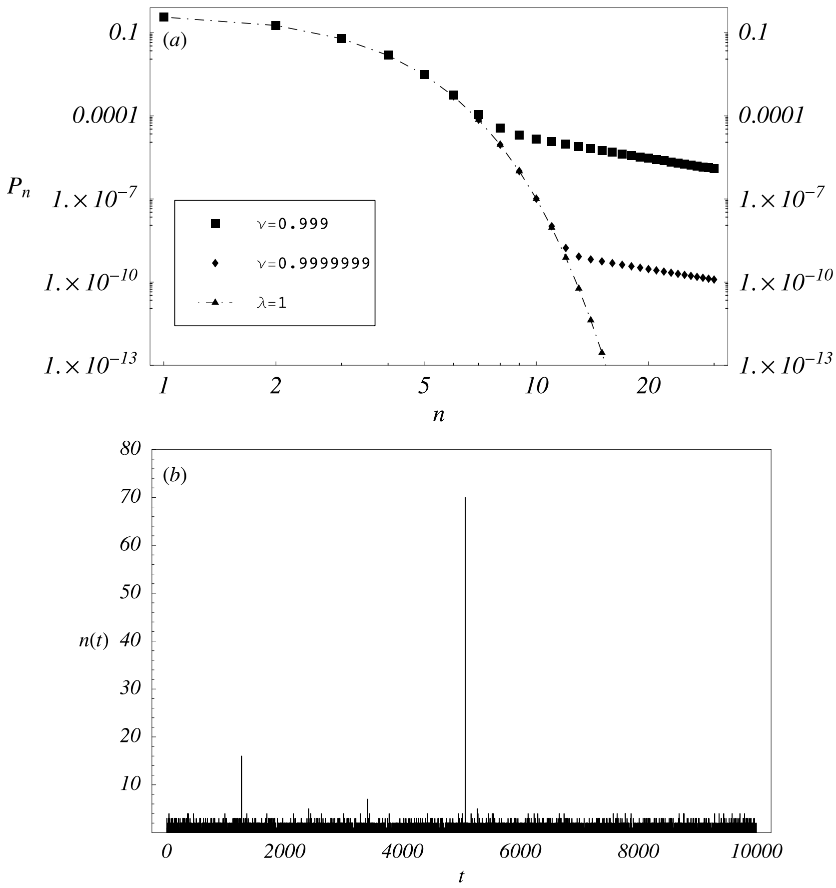


Figure 9. (*a*) Poisson distribution together with two examples of discrete-stable distributions that have indices very close to 1. (*b*) Realization of a discrete-stable process with $v = 0.999$.

in the population size punctuated by rare but large events. Figure 9(*b*) shows a realization of a discrete-stable process with $v = 0.999$. The majority of the realization follows that of a Poisson distribution with the exception of two extreme events. The probability of these events occurring in a Poisson process are approximately $1/16!$ (4.8×10^{-14}) for the event at $t \sim 1250$ and $1/70!$ (8.3×10^{-101}) for the event at $t \sim 5000$. The probability that these events occur in the stable process with index $v = 0.999$ are 4.9×10^{-6} and 2.1×10^{-7} respectively. Events such as these might be termed 'outliers' in the context of a Poisson process, but are nevertheless an integral part of the discrete-stable process and occur with much greater frequency.

5. Summary and conclusion

This paper has studied the accuracy of measurements of an external series of events generated by individuals leaving a population whose fluctuations are governed by the class of discrete-stable distributions [9]. In the analysis of such a process, characterization is problematic due to the presence of a power-law tail in the probability distribution. This means that no moment-based statistics such as the mean or autocorrelation function exist. By implementing a simple nonlinear signal processing technique known as 'clipping', we have shown how the external series of events can be regularized so that the moments and correlation functions of the new 1-bit process are finite. These provide new options for the determination of parameters such as the power-law index that characterize the original process and we have extended the approach used in earlier work to investigate the statistical accuracy of such measurements. In particular, the characteristic index of the process can be deduced from the integration time dependence of estimates of the clipped mean and we have calculated the variance of these.

Unlike the case of a classical birth–death–immigration process, the variance of estimates of the clipped mean of a series of events generated by a discrete-stable process increase without limit as the number of samples in a fixed experiment time is increased. In the birth–death–immigration case the chance of finding more than one event in an integration time becomes so small for sufficiently large N that clipping has no effect, whilst correlation means that the effective number of independent samples saturates. Thus the statistical error converges to that of the unclipped case as the experiment time is divided up into more and more samples. In the case of events generated by a stable process, however, the number of events decreases more slowly with integration time so that clipping continues to discard information whilst the samples become more correlated. This leads to the divergence of the error with increasing number of samples shown in figure 4 and according to equation (21) means that not only is there is an optimum number of samples into which any fixed experiment time should be divided in order to obtain the best estimate of the characteristic power-law but that the accuracy of such measurements will deteriorate without limit if the data resolution is improved beyond this optimum value. When the data are optimally sampled the relative variance of estimates of the power-law improves more slowly with experiment time than a simple reciprocal scaling (figure 7). This is similar to results found previously for measurements on Gaussian random fractals

The optimum value of N is influenced by the correlation properties of the underlying process. We have found that these correlations can be persistent for events generated by a stable process. For example, the degree of correlation for small power-law indices is small but extends over long periods so that the optimum number of samples is large. Conversely, as $\nu \to 1$ the population and counting processes become ostensibly more Poisson-like in nature. In this limit emanations from the population comprise small, Poisson-like fluctuations, punctuated by extreme rare events relating to the residual tail of the discrete-stable distribution. These extreme events have the characteristics customarily associated with outliers; however, they are an integral part of a discrete-stable process distinct from the Poisson case $\nu = 1$.

In conclusion, we have demonstrated some of the subtleties that may be encountered in making measurements on correlated random series of events characterized by power-law probability distributions. This paper concentrated on the accuracy of estimates of the mean number of clipped events generated by a discrete-stable process with the clipping level set at unity (equation (11)). Improvements in accuracy could be sought by increasing the threshold of the clipping level so that less information is discarded. We have not investigated the implications of our results for the determination of particular parameters of the model in detail although equation (21) makes predictions for the power-law index itself.

The accuracy analysis can be employed to consider a wider class of scale-free fluctuations than the discrete-stable class considered here. A nonlinearly coupled population model is investigated in [23] which, for a particular choice of model parameters, gives scale-free distributions with generating function

$$Q(s) = \frac{1}{1 + As^{\nu}}. \tag{23}$$

Although not of the discrete-stable class, the tails of the corresponding distributions are the same as that of the class of discrete-stable distributions. The accuracy of measurements is a key consideration in the analysis and evaluation of data enabling different stochastic models to be distinguished and matched to physical processes.

Acknowledgments

J O Matthews acknowledges the support of the Leverhulme Trust and G B Siviour was supported by a Nuffield Foundation Undergraduate Research Bursary.

Appendix

The clipped autocorrelation function is defined as

$$\frac{\zeta(T, \Delta t)}{\bar{c}^2} = \frac{\langle c(0,T)c(T+\Delta t, T)\rangle}{\langle c(T)\rangle^2} = \frac{(1 - 2Q_c(0,1;T) + R(1,1;T+\Delta t, T))}{(1 - Q_c(0,1;T))^2} \tag{A.1}$$

where

$$\begin{aligned} R(z, z'; t, T) = {} & Q_2(0, z'; T) Q_1(\Phi(0, z'; T); t - T) Q_2(f(\Phi(0, z'; T); t - T), z; T) \\ & \times Q_1(\Phi(f(\Phi(0, z'; T); t - T), z; T), \infty), \end{aligned} \tag{A.2}$$

and where the generating functions with equation (A.1) are in the table below for the birth–death–immigration and discrete-stable population models, respectively.

	Birth–death–immigration model	Discrete-stable model
Q_1	$\left(1 + \frac{\bar{N}}{\beta} s(1 - \exp(-\bar{\mu}t))\right)^{-\beta}$	$\exp\left(-\frac{as^\nu}{\nu\mu}(1 - \exp(-\nu\mu t)\right)$
f	$\frac{s\exp(-\bar{\mu}t)}{1+\frac{\bar{N}}{\beta}s(1-\exp(-\bar{\mu}t))}$	$s\exp(-\mu t)$
Q_2	$\left(\frac{u_+ - u_-}{(s-u_-)\mathrm{e}^{\lambda u_+ T} - (s-u_+)\mathrm{e}^{\lambda u_- T}}\right)^\beta$	$\exp\left(-\frac{a}{(1+\nu)\eta z}\left(\Phi(s,z;T)^{\nu+1}F\left(1+\nu, 1, 2+\nu; \frac{\bar{\mu}\Phi}{\eta z}\right) - s^{\nu+1}F\left(1+\nu, 1, 2+\nu; \frac{\bar{\mu}s}{\eta z}\right)\right)\right)$
Φ	$\frac{u_+(s-u_-)\mathrm{e}^{\lambda u_+ T} - u_-(s-u_+)\mathrm{e}^{\lambda u_- T}}{(s-u_-)\mathrm{e}^{\lambda u_+ T} - (s-u_+)\mathrm{e}^{\lambda u_- T}}$	$\frac{\eta z + (\bar{\mu}s - \eta z)\exp(-\bar{\mu}T)}{\bar{\mu}}$

The constant λ is the birth rate of the process and

$$u_\pm = \frac{-\bar{\mu} + \lambda \pm [(\bar{\mu} - \lambda)^2 + 4\lambda\eta z]^{1/2}}{2\lambda}, \tag{A.3}$$

$$Q_c(s, z; T) = Q_1(\Phi(s, z; T), \infty) Q_2(s, z; T) \tag{A.4}$$

and $F(a, b, c; x)$ is the hypergeometric function [19]. A full derivation can be found in [9].

References

[1] Albert R and Barabási A L 2002 Statistical mechanics of complex networks *Rev. Mod. Phys.* **74** 47–97

[2] Jakeman E and Jordan D L 1990 Statistical accuracy of measurements on Gaussian random fractals *J. Phys. D: Appl. Phys.* **23** 397–405

[3] Lévy P 1937 *Théorie de l'Addition des Variables Aléatories* (Paris: Gauthier-Villars)

[4] Krapivsky P L and Redner S 2001 Organisation of growing random networks *Phys. Rev.* E **63** 066123

[5] Bornholdt S and Ebel H 2001 World wide web scaling exponent from Simon's 1955 model *Phys. Rev.* E **64** 035104

[6] Barabási A L, Albert R and Jeong H 1999 Mean-field theory for scale-free random networks *Physica* A **272** 173–87

[7] Newman M E J, Strogatz S H and Watts D J 2001 Random graphs with arbitrary degree distributions and their applications *Phys. Rev.* E **64** 026118

[8] May R M and Lloyd A L 2001 Infection dynamics on scale-free networks *Phys. Rev.* E **64** 066112

[9] Matthews J O, Hopcraft K I and Jakeman E 2003 Generation and monitoring of discrete stable random processes using multiple immigration population models *J. Phys. A: Math. Gen.* **36** 11585–603

[10] Shepherd T J and Jakeman E 1987 Statistical analysis of an incoherently coupled, steady-state optical amplifier *J. Opt. Soc. Am.* B **4** 1860–9

[11] Jakeman E and Shepherd T J 1984 Population statistics and the counting process *J. Phys. A: Math. Gen.* **17** L745–50

[12] Van Vleck J H and Middleton D 1966 The spectrum of clipped noise *Proc. IEEE* **54** 2–19

[13] Jakeman E and Pike E R 1969 Spectrum of clipped photon-counting fluctuations of Gaussian light *J. Phys. A: Math. Gen.* **2** 411–2

[14] Jakeman E 1970 Theory of optical spectroscopy by digital autocorrelation of photon-counting fluctuations *J. Phys. A: Math. Gen.* **3** 201–15

[15] Jakeman E, Oliver C J, Pike E R and Pusey P N 1972 Correlation of scaled photon-counting fluctuations *J. Phys. A: Gen. Phys.* **5** L93–6

[16] Woolf M, Arrowsmith D K, Mondragón-C R J and Pitts J M 2002 Optimization and phase transitions in a chaotic model of data traffic *Phys. Rev.* E **66** 046106

[17] Jakeman E, Hopcraft K I and Matthews J O 2003 Distinguishing population processes by external monitoring *Proc. R. Soc.* A **459** 623–39

[18] Shepherd T J 1984 Photoelectron counting—semiclassical and population monitoring approaches *Opt. Acta* **31** 1399–407

[19] Abramowitz M and Stegun I A *Handbook of Mathematical Functions* 9th edn. (New York: Dover)

[20] Renshaw E 1991 *Modelling Biological Populations in Space and Time* (Cambridge: Cambridge University Press)

[21] Bartlett M S 1966 *An Introduction to Stochastic Processes* (Cambridge: Cambridge University Press)

[22] Evans M, Hastings N and Peacock B 1993 *Statistical Distributions* 2nd edn (Chichester: Wiley)

[23] Jakeman E, Hopcraft K I and Matthews J O 2005 Fluctuations in a coupled population model *J. Phys. A: Math. Gen.* **38** 6447–61