Effect of roost size on the emergence behaviour of pipistrelle bats

J. R. SPEAKMAN, N. IRWIN, N. TALLACH & R. STONE
Department of Zoology, University of Aberdeen

(Received 18 November 1997; initial acceptance 28 January 1998; final acceptance 7 June 1999; MS. number: 5705R)

We refined established methods for examining the temporal organization of behavioural events and applied them to the study of bats emerging from roost sites. Previous studies have shown that as roosting numbers of bats increase, temporal patterning (clustering) in their emergence behaviour, as measured by $\chi^2$ or $G$ statistics, also increases. Expressing the extent of temporal clustering using $\chi^2$ or $G$, however, confounds two effects. The statistic may increase because clustering increases. However, an increase will occur, even if the amount of clustering stays constant, if the sample size of observations increases. A positive relation between emergence size and $\chi^2$ or $G$ may not necessarily reflect increased clustering. We developed a method to separate these effects using computer-modelled event streams. Using this method, we examined intra- and inter-roost variation in the temporal patterning of emergences of the pipistrelle bat, *Pipistrellus pipistrellus*, from four roosts in northeast Scotland. The ‘corrected’ expression of clustering decreased as roost size increased, which was opposite to the effect found with the uncorrected expression and also opposite that predicted a priori from the bottleneck hypothesis. This novel result suggests bottlenecks may disrupt clustering rather than promote it. This latter interpretation was supported by observations that clustering was significantly reduced in the middle of large emergences. Variation in clustering between roosts was significant, suggesting exit topology may have an important influence on clustering behaviour.

An important aspect of almost all behaviour is its temporal organization. Studies of temporal organization include foraging (Sibley et al. 1990; Berdoy 1993), preening (Slater 1974), thermoregulatory behaviour such as the periodicity of arousals from hibernation (Strijtstra & Daan 1997; Thomas & Geiser 1997), breathing (Thomas & Cloutier 1992) and vocalization (Koene 1996). One issue concerning temporal organization that is common to all these studies is whether the behaviour in question is performed at random over time, or whether it is nonrandomly structured. The methodology for detecting nonrandomness in behavioural patterns of bats, as they exit their roosting locations, has developed considerably. The principles established in this field of study are applicable across all other fields where temporal organization is important.

Almost all insectivorous microchiropteran bats are nocturnal, spending their days at rest in a ‘day roost’ and their nights either actively foraging, or roosting in ‘night roosts’ (Anthony & Kunz 1997). Bats vary enormously in the range of sites that they select as roost sites and also in the extent of their gregariousness at these sites. A number of species choose enclosed structures, such as caves, church attics of houses and the insides of hollow trees as day roosts (Kunz 1982), all of which may have relatively small entrances. Consequently, at dusk individuals of these species move from inside the roost, through the constricted entrance, to the outside. Particularly when roosting populations are large this ‘emergence’ of bats from their day roost is a predictable and often conspicuous event.

Probably, at least in part, because of the predictability of the emergence event, it has provided the focus of many research studies (e.g. Venables 1943; Church 1957; Erkert 1982; McAney & Fairley 1988; Maier 1992). More recently, attention has been directed towards the exact timing of individual emergences of bats within the entire emerging stream. Kunz (1974) was among the first to report that bats, in this case *Myotis velifer*, appeared to emerge in groups. Swift (1980) also reported that pipistrelle bats, *Pipistrellus pipistrellus*, emerged, not at random, but in groups, which she termed ‘outbursts’.

Brigham & Fenton (1986) and Bullock et al. (1987) applied more rigorous statistical methods to examine the structure of bat emergences. Brigham & Fenton (1986) compared the observed distribution of interevent intervals to that expected from a random emergence, whereas Bullock et al. (1987) used log survivorship analysis. This latter study, on pipistrelle bats, showed that the
emergence stream consisted of ‘clusters’ or ‘outbursts’ of bats, separated by gaps during which relatively few individuals emerged. Wilkinson (1992) also applied the log survivorship technique to show clustering in the emergence behaviour of evening bats, Nycticeius humeralis.

However, the technique of log survivorship analysis, in its application to temporal patterning in behaviour (Slater 1974), makes an implicit assumption that the intensity of the process generating the events is constant. This is also assumed in more direct comparisons of the distribution of the event intervals to that expected at random (Brigham & Fenton 1986). This assumption is violated in the emergence of bats (and also many other patterns of behaviour such as those listed above) because the underlying intensity approximates a normal distribution, with the periods at the start and end of the emergence having relatively low intensities, relative to the middle part. This underlying structure can cause the spurious appearance of clustering in random emergences when they are analysed by log survivorship analysis (Speakman et al. 1992; Speakman 1993). An empirical method has been developed that overcomes this analytical problem (Speakman et al. 1992) and there is software (Clustan) which allows the appropriate analysis to be performed easily and rapidly (Speakman 1993). The software also makes an explicit comparison of the observed distribution of inter-event intervals to that expected if the emergence was random and tests the significance of this difference using $\chi^2$ and $G$ tests. Using this revised method of analysis, clustering has been confirmed to be a statistically significant aspect of the emergence behaviour of several species (little brown bat, Myotis lucifugus: Kalcounis & Brigham 1994; M. B. Fenton, personal communication; Pipistrellus pipistrellus: Speakman et al. 1995; G. Jones, personal communication; serotine bat, Eptesicus serotinus; K. Petrzelkova, personal communication; Tadarida sp.: M. B. Fenton, personal communication).

There are three hypotheses concerning the functional significance of clustering in emergences. The first is that it is an artefact of large numbers of animals moving through a small constriction: the ‘bottleneck effect’ (Bullock et al. 1987). Second, it may be an antipredator behaviour (after Hamilton 1971) whereby, by emerging close to another individual, a bat dilutes its own probability of being predated (Leen & Novick 1977): the ‘selfish cluster’ (Speakman et al. 1995). Finally, bats may emerge in synchrony to coordinate their foraging behaviour (Barak & Yom-Tov 1989; Wilkinson 1992; Wilkinson & Boughman 1998).

Attempts to test these hypotheses have yielded conflicting results. We studied pipistrelle bats emerging at roosts where there were multiple exit holes and found that bats coordinated their emergences across the separate holes, more than would be expected if the two streams were simply two independent bottlenecks (Speakman et al. 1995). We also observed bats clustered more in the first half of the emergence, when it was lighter. Since insectivorous bats are more vulnerable to predation in daylight (Speakman 1991; Fenton et al. 1994; Speakman et al. 1994) the clustering behaviour was most consistent with an antipredator response. However, Kalcounis & Brigham (1994) found little brown bats did not respond to a plastic predator (owl) placed immediately outside the exit hole, combined with owl calls played outside the roost during the emergence. This could indicate that the bats simply did not recognize this ‘predator’ as real. More importantly, Kalcounis & Brigham (1994) also noted that the extent of clustering ($\chi^2$) increased as roost size increased which they suggested was most consistent with the ‘bottleneck’ hypothesis.

There is a problem, however, in interpreting the $\chi^2$ (or $G$) values as indicative of the extent of clustering. This problem stems from the fact that $\chi^2$ or $G$ values, and their associated probabilities, reflect only the deviation of the observed temporal pattern from a random pattern. If one compares two emergences, of the same size, and one yields a greater $\chi^2$ or $G$ value than the other, this indicates clustering is increased in the pattern that has the greater value (assuming the exact pattern of deviation from random is also consistent with clustering). However, if one takes a fixed pattern of clustering, the $\chi^2$ and $G$ values will increase as the number of recorded events increases.

The increase in $\chi^2$ and $G$ with sample size occurs because when the number of observations in a sequence is small there is a high chance that, even if the pattern is clustered, it could have resulted from a random process. However, as the number of observations increases it becomes increasingly unlikely that the pattern could be the result of a random process. The $\chi^2$ and $G$ values increase, as more observations are made, reflecting this declining probability of the extending sequence being random. The associated probabilities express the likely deviation of an observed pattern from a random pattern. However, as a quantitative expression for the extent of clustering, they are valid only when the emergence size is constant. Increasing $\chi^2$ values as a function of emergence size (Speakman et al. 1992; Kalcounis & Brigham 1994) might only reflect the statistical consequences of increasing sample size, rather than an increase in the extent of clustering. The critical values of the goodness-of-fit statistics corresponding to $P=0.05$ and 0.01 are dependent on the number of classes of events being compared rather than sample size. Consequently the increased values of the statistics are associated with increased probabilities of the patterns being different from random. We acknowledge that there was a mistake in previous papers (Speakman et al. 1992; Speakman 1993) where it was suggested that the $\chi^2$ and $G$ statistics reflect the degree of clustering and users were encouraged to use $\chi^2$ and $G$ as measures of clustering.

In this paper we do two things. First, we explore the dependency of the goodness-of-fit statistics ($\chi^2$ and $G$) on sample size for a series of computer-generated event patterns that differ in the extent of their clustering. This modelling allows us to formulate a method for removing the statistical consequences of increasing sample size on the goodness-of-fit statistics. The result is a value that expresses the extent of clustering in emergences, independent of their size. Second, we apply this new method to examine the effects of intra- and inter-roost variation in size of emergence, on the extent of clustering of
emerging pipistrelle bats (55-kHz phenotype: Jones & van Parijs 1993).

METHODS

Computer Modelling

We devised a method for removing the statistical artefact effect of variations in sample size on the expression of clustering. The details of the method can be found in the Appendix.

Field Observations

We observed emergences of pipistrelle bats from maternity roosts in northeast Scotland (57°N) in 1992, 1993 and 1994. We made observations at four roosts, involving 51 emergences. All roosts contained bats of the 55-kHz phenotype (Jones & van Parijs 1993). For one of these roosts the data have been published previously (Speakman et al. 1995: roost A, N=3 nights), for a second roost (B) we used a mixture of previously published observations (N=15 nights: Speakman et al. 1995) supplemented by further observations made in 1994 (N=11 nights). The data for the final two roosts have not been published previously (roost C: N=13 nights; roost D: N=9 nights). In all cases the field observations were made according to the protocols given in Speakman et al. (1995), which will not be repeated in full here. Briefly, individual emergence events were logged, either directly using an event recorder on a laptop computer, or first into a tape recorder, and then on a computer.

The event patterns were analysed for clustering using the Clustan software (Speakman 1993) trimming emergences to remove the effect of variation in underlying intensity (Speakman et al. 1992). We examined the extent of clustering in relation to roost size using a methodology derived from the computer modelling (Appendix). This procedure removed the artefactual effects of sample size that were not related to changes in intensity of clustering. Two analyses were performed. First, we explored the extent of clustering as a function of roost size within a single roost (roost B: N=26 emergences). We removed the statistical artefact effect of emergence size and the effect of roost on the corrected estimate of clustering, using ANCOVA, with size as a covariate.

Intraroost Variation

The observed emergences at roost B included between 164 and 976 bats (N=26). There was a significant positive relationship (Fig. 1) between the $\chi^2$ value from the Clustan analysis and roost size at this roost (N=26 nights of observation). The least-squares fit regression, ln $\chi^2$ value $= 1.61 + 0.595$ ln number emerging, explained 27.9% of the variation in the $\chi^2$ value ($F_{1,24}=9.30, P<0.006$). There was no significant difference in the gradients of the size effects between years (ANCOVA: $F_{1,23}=1.79, P=0.194$) but the year effect was highly significant ($F_{1,23}=37.5, P<0.001$).

We removed the statistical artefact effect of emergence size by correcting these raw values to a mean roost size of 500 bats, using the method detailed in the Appendix. The relationship between the corrected $\chi^2$ values and roost size was still significant, but this time it was negative (Fig. 2). The least-squares fit regression equation, ln corrected $\chi^2$ value $= 9.45 - 0.6648$ ln size of emergence, explained 34.2% of the variation in the corrected $\chi^2$ value ($F_{1,24}=12.46, P<0.002$). Inevitably, given the uncorrected data there was also a significant year effect ($F_{1,23}=36.61, P<0.001$) and the size effect, independent of the year effect, also remained significant ($F_{1,23}=5.71, P=0.025$). This analysis means when larger numbers of bats exited this roost, the emergence was significantly less clustered than when the roost contained fewer bats. Independent of roost size the emergences were more clustered in 1993 than in 1994.
We compared the clustering in the first 100 bats of an emergence with that in the middle 100 bats (centred around the median) and the last 100 bats, for seven emergences from this roost where the total emergence exceeded 400 individuals. Since in these comparisons the sample size was constant we did not make any correction for sample size and compared the raw $\chi^2$ values. The amount of clustering varied significantly between the three phases (Fig. 3: one-way ANOVA: $F_{3,17}=8.83$, $P=0.002$). Using paired $t$ tests we established the clustering at the middle of the emergence was lower than either early ($t_{12}=3.51$, $P=0.013$) or late ($t_{12}=4.34$, $P=0.004$) in the emergence. The mean amount of clustering early in the emergence exceeded that later but the difference was not significant ($t_{12}=1.42$, $P=0.2$). In the middle of the emergences the pattern observed did not differ from that expected at random on 6 of the 7 nights analysed.

Inter-roost Variation

There was a strongly positive relationship between uncorrected $\chi^2$ values and roost size for data from all four roosts (Fig. 4a). The least-squares fit regression equation $\ln \chi^2 = -1.33 + 1.05 \ln$ number emerging, explained 70.1% of the variation in the $\chi^2$ value ($F_{1,49}=117.4$, $P<0.001$). The corrected estimate of clustering (corrected to a size of 350 bats) did not vary significantly across the four roosts (one-way ANOVA: $F_{3,49}=1.24$, NS). However, this does not mean there were no significant differences in clustering between the roosts because they also differed significantly in the mean numbers of emerging bats...
(one-way ANOVA: $F_{1,49}, P=0.001$). The relationship between the corrected $\chi^2$ values across all the roosts ($N=51$) and roost size was significant, and again negative (Fig. 4b). The least-squares fit regression equation, ln corrected $\chi^2$ value $= 6.259 - 0.234 \ln$ number emerging, explained 11.5% of the variation in the corrected $\chi^2$ value ($F_{1,48}=6.37, P=0.015$). The gradient of the relationship between size and clustering varied significantly between roosts (ANCOVA: roost*size interaction term: $F_{1,47}=3.70, P=0.019$). However, the significance of this effect was dependent entirely upon the data for roost A, for which we had only three data points, suggesting the possibility of a small sample size effect. When the data for roost A were excluded and the analysis restricted to the three roosts where we had at least 9 nights of data, the significant interaction effect disappeared. When the analysis was run again, excluding the interaction term, both size of emergence (negative $F_{1,48}=17.98, P<0.001$) and roost ($F_{1,48}=7.04, P=0.002$) significantly influenced the extent of clustering.

These data indicate that there was a significant negative effect of roost size on the extent of clustering, both within and between roosts. The gradient of this effect is probably independent of the roost. However, roost did enter as an independent factor influencing the extent of clustering. Within a single roost studied in 2 years there was a significant year effect independent of the size of the roost.

**DISCUSSION**

**Intraroost Variation**

As found in previous studies (e.g. Kalcounis & Brigham 1994), the uncorrected estimates of the extent of clustering, using the $\chi^2$ statistic, increased with increasing roost size. The significant interyear effects reported here have not been previously described and the cause of this observation is obscure. The significant effect of roost size on clustering formed a major aspect of the argument by Kalcounis & Brigham (1994) that the emergence behaviour reflects a bottleneck. However, removing the statistical artefact effect of emergence size revealed that, at least for pipistrelles, when larger numbers of bats emerged, they were actually less clustered than when smaller numbers emerged. This novel result contradicts the predictions of the bottleneck hypothesis as formulated by Bullock et al. (1987) and tested by Kalcounis & Brigham (1994).

The negative size effect could occur when large numbers of bats overflow through the emergence hole, because of the pressure of numbers trying to push their way out. In large emergences, bats may be forced to exit the roost before they intended. However, when smaller numbers are in the roost, individuals may normally emerge in synchrony, because there is less force of numbers pushing them out at unfavourable times, leading to an increase in the amount of clustering. P. Davidson & T. H. Kunz (personal communication) observed emergences of little brown bats from inside and outside a roost simultaneously and observed some individuals waiting at the exit for other individuals from inside to join them, supporting the suggestion that when the traffic is light bats may have more control over their exit behaviour. The pattern of variation in clustering throughout emergences also supports this interpretation, since at the start, and end, of emergences when the intensity of emerging bats was relatively low, clustering exceeded that at the peak of emergence when the intensity was much higher (Speakman et al. 1992).

If this is true, the prediction that a bottleneck effect leads to increased clustering as roost size increases may be an oversimplification of what is happening at the bottleneck. Clustering may be less likely to occur as roost size increases, when there is a bottleneck, because the force of numbers in larger emergences may prevent bats at the peak of emergence from choosing exactly when to emerge, thereby disrupting the clustering.

There are several other functional interpretations of the decreased clustering in relation to roost size. For example, there may be an effect of predation, linked to variation in the attractiveness of different sizes of roost to potential predators. We do not know whether larger or smaller roosts around Aberdeen attract more predators. In South Africa, Fenton et al. (1994) found that predators accumulated at larger roosts which is opposite to the effect that would be necessary to produce more clustering in smaller roosts.

**Inter-roost Variation**

Clustering varied significantly between the roosts. This effect is predicted by the bottleneck hypothesis because the exact roost topology might be expected to affect the dynamics of the bottleneck. However, this effect could also be caused by other factors. It seems likely, for example, that predation risk might vary between roosts even within relatively small areas and that the risk might also vary with the exact nature of the exit and the immediate flight paths of the bats once emerged. Because of the small sample of roosts involved in the present study ($N=4$) it was not possible to relate variation in the extent of clustering to aspects of the roost architecture or local predation pressure which could not be quantified. Clearly more studies are required, involving larger numbers of roosts and/or more variant exit structures, and also experimental manipulations of exits (as suggested by Kalcounis & Brigham 1994) to understand fully the roles that exit structure and local predation pressure have on the clustering phenomenon.

**Acknowledgments**

We are grateful to the roost owners who allowed us access to their properties to count bats. J.R.S. would like to apologise for the misleading statement in the instructions to the Clustan software that the resultant statistics directly reflect the extent of clustering. At the time these instructions were produced I believed this to be the case and was unaware that this is only strictly true when the sample size of emergence is constant. The statistics derived from the program do reflect the probability that
observed patterns deviate from random. In combination with the procedures outlined in the present paper, a size-invariant expression for clustering can be derived. I hope this software and this paper will continue to assist those people studying these phenomena. The software is available free of charge at the following Web site http://www.abdn.ac.uk/zoology/speakman.htm.

References


APPENDIX

We performed five separate simulations. For each simulation we generated 25 event streams, which differed in their extent of clustering, relative to a completely random pattern, using the Minitab statistical package (Ryan et al. 1987). To generate the 25 event streams in each simulation we first generated a short string of baseline inter-event intervals which were all strongly clustered, but in different ways. The baseline event streams differed for each simulation and were as follows.

Simulation 1:

1 1 4 6 2 1 5 8 3 1 1 2 10 12 5 1 3 8

Simulation 2:

4 5 10 7 1 1 2 1 10 7 9 2 1 1 1 10 8 1 1 1 2

Simulation 3:

2 3 1 1 3 2 1 1 2 1 0 1 2 1 1 2 1 2 1 5

Simulation 4:

2 2 3 10 15 5 2 3 17 2 1 3 8 3 2 2 0 14 2

Simulation 5:

17 3 0.3 8 12 0.6 6 12 6 5 2

3 16 8 0.1 2 4 20 10

These different simulations were designed to mimic a wide range of potential patterns. Hence simulations 1, 2 and 3 include similar intervals separating events within


and between clusters (1–2 s between events within clusters and approximately 7–10 s between events outside clusters), but vary in the number of events within each cluster (2–4 in simulation 1, 4 in simulation 2 and 6–7 in simulation 3). Simulations 4 and 5 vary the intervals between events within the clusters. In simulation 4 the duration between events within clusters is 2–3 s and in simulation 5 it varies between 0.1 and 3 s. In both simulations 4 and 5 the variation in intervals between clusters is 10–20 s.

We made multiple copies of each pattern until we had streams in excess of 2000 interevent times, which repeated over and again the original patterns. To generate the 25 streams within each simulation we then added varying degrees of stochasticity into each baseline stream by generating a parallel stream of 2000 random interevent times, using the Rand command and the Exponential subcommand. In each simulation we added the clustered and random streams together to make a third stream which thus contained aspects of both the clustered and random elements. By varying the mean intensity of the exponential distribution, between 1 and 6, we could introduce varying degrees of randomness into the final combined stream (increasing stochasticity associated with greater values). Within each simulation therefore we generated a series of 25 different patterns which included varying amounts of clustering. As judged from the resultant goodness-of-fit statistics at any given sample size, the simulation patterns included a substantial variation in clustering patterns which far exceeded that observed at natural bat roosts studied to date (across four species: Pipistrellus pipistrellus, Myotis lucifugus, M. thysanodes and Tadarida sp.). Once we had generated the final patterns we completely shuffled the order of the events in each stream.

To sample the events from a given stream we used random numbers between 1 and 1000 to define start points. Starting at a given random number we selected between 50 and 900 interevent intervals. We then reconstructed the sequential timing corresponding to these interevent interval streams of varying sizes. The sequential timings were exported and analysed using Clustan (Speakman 1993). We examined the extent to which the sample size of events affected the derived \( \chi^2 \) and \( G \) values for all the different patterns of event streams.

Figure A1 shows the effects of sample size on the measured \( G \) and \( \chi^2 \) statistics resulting from the Clustan program for five of the 25 computer-generated emergence streams (which differed in the amount of added stochasticity) from one of the simulations (simulation 2; for clarity not all 25 streams are illustrated in the figure). As anticipated, despite the pattern of clustering in each stream being constant, there was an increase in both \( G \) and \( \chi^2 \) as the sample size increased. The value of the statistic corresponding to the 0.01 significance level also increased with sample size because the number of classes of events being compared also increased and thus the degrees of freedom also expanded as sample size increased. Nevertheless, the increase in the \( G \) and \( \chi^2 \) values for all these clustered patterns far exceeded the increase in the fixed probability criterion, because the increase in sample size far exceeded the increase in the number of classes. When these patterns were plotted on log–log axes the gradients were linear, and the variance in the \( G \) or \( \chi^2 \) values explained by sample size for any given pattern always exceeded 98.9%. The same pattern was observed in all five of the simulations and was consequently independent of the extent or pattern of clustering in the starting event patterns.

As is clear from Fig. A1 the gradient of the log–log relationships between the goodness-of-fit statistics and sample size was significantly affected by the amount of clustering in the event stream (\( F_{1.22}=7.08, P<0.001 \)), as assessed by the amount of stochasticity introduced into the repeated data stream. When more stochasticity was introduced, the gradient of the increase in the goodness-of-fit statistic was lower (Fig. A1). The gradients also differed between the two statistics (\( G \) and \( \chi^2 \); not illustrated). The dependence of the gradient relating the goodness-of-fit statistic to sample size, on the degree of stochasticity introduced into the data stream, was unfortunate because this meant the extent of the size artefact was correlated with the extent of clustering. When the pattern was more clustered there was a greater size artefact.

To remove the artefactual effect of increasing sample size on the goodness-of-fit statistic, it was necessary to estimate the gradient of the size artefact effect. The problem was that this gradient depended on the extent of clustering, which was what we were attempting to measure. From any pattern analysis for an actual event stream only two pieces of information are available: the sample size of emergence and the goodness-of-fit statistic. We have already established that when the sample size is constant, increases in the statistics will reflect the extent of clustering. We therefore derived a relationship using the sample size and the values of the goodness-of-fit statistics as predictors, and the gradients derived from the simulations which linked the increase of the...
goodness-of-fit statistics to sample size (hereafter called the artefact gradients). In effect this involved estimating the gradient of the artefact at each point illustrated in Fig. A1 (and for each of the four simulations that are not illustrated), and using the sample size and the value of the goodness-of-fit statistic at that point as predictors of the gradient in a multiple linear regression analysis. For both $\chi^2$ and $G$ statistics both the sample size and the observed value of the appropriate statistic ($G$ or $\chi^2$) significantly influenced the artefact gradients.

Table A1 gives the resultant parameters of the least-squares fit regression equations for each of the five simulations. In each of the separate simulations the forms of the relationships derived were very similar. The constants of the equations were consistent ($\bar{X} \pm \text{SE} = 1.266 \pm 0.035$). In all cases there was a positive effect of the $\chi^2$ value, matched with a negative effect of the sample size ($N$). The coefficients with respect to $\chi^2$ and sample size varied widely between the different simulations (Table A1). However, in each case the magnitude of the coefficient with respect to $\chi^2$ was about twice the equivalent coefficient with respect to sample size. Since these coefficients were in opposite directions, and $\chi^2$ and $N$ were strongly positively correlated for any particular pattern (Fig. A1), the effects effectively cancelled each other, and in all cases the actual derived gradients were always very similar to the constant of the appropriate equation. Since these constants were consistent across equations these simulations each resulted in very similar predicted artefact gradients for any particular combination of $\chi^2$ and sample size. An almost identical pattern was observed in the simulations for the $G$ statistic. Variation in the parameters of the equations was not linked in any obvious manner to the clustering patterns introduced into the initial simulation streams. The average correction equations summed across all of the five simulations were for $\chi^2$: logged artefact gradient=$1.2666+0.000703 \ (\chi^2) - 0.00035 \ (N)$; for the $G$ statistic: logged artefact gradient=$1.245+0.00057(G) - 0.000397(N)$. These simulations indicate that, for an extremely wide variety of clustering patterns, by using the information resulting from an analysis of patterning in an event stream, namely the goodness-of-fit statistics and the sample size, it is possible to estimate the logged gradient of the size artefact effect.

These relationships provide a method whereby the artefact of increasing sample size on the statistics could be removed, generating a size-independent estimate for clustering. To illustrate how this might be achieved, imagine a study had been made of emerging bats. On one night 500 bats emerged and the $\chi^2$ value describing the extent of deviation of their emergence pattern from random was 193 (Fig. A2). On a second night only 200 bats emerged and they yielded a $\chi^2$ value of 117. Which emergence is more clustered? Using the $\chi^2$ values alone it would appear that the larger emergence is more clustered. However, to answer this question correctly we must remove the

Table A1. Parameters of regression equations relating the ‘artefact gradient’ to the sample size ($N$) and $\chi^2$ or $G$ statistics for each of the five simulations

<table>
<thead>
<tr>
<th>Simulation</th>
<th>Constant</th>
<th>Coefficient ($\chi^2$/G)</th>
<th>Coefficient (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$ statistic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.2293</td>
<td>0.000388</td>
<td>-0.000169</td>
</tr>
<tr>
<td>2</td>
<td>1.2827</td>
<td>0.000411</td>
<td>-0.000298</td>
</tr>
<tr>
<td>3</td>
<td>1.1984</td>
<td>0.000142</td>
<td>-0.000079</td>
</tr>
<tr>
<td>4</td>
<td>1.2412</td>
<td>0.000946</td>
<td>-0.000472</td>
</tr>
<tr>
<td>5</td>
<td>1.3817</td>
<td>0.001630</td>
<td>-0.000723</td>
</tr>
<tr>
<td>$\bar{X}\pm\text{SE}$</td>
<td>$1.2666\pm0.035$</td>
<td>$0.000703\pm0.000297$</td>
<td>$-0.00035\pm0.000120$</td>
</tr>
<tr>
<td>$G$ statistic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.198</td>
<td>0.00039</td>
<td>-0.00022</td>
</tr>
<tr>
<td>2</td>
<td>1.257</td>
<td>0.00063</td>
<td>-0.00083</td>
</tr>
<tr>
<td>3</td>
<td>1.205</td>
<td>0.00010</td>
<td>-0.00006</td>
</tr>
<tr>
<td>4</td>
<td>1.203</td>
<td>0.00059</td>
<td>-0.00034</td>
</tr>
<tr>
<td>5</td>
<td>1.363</td>
<td>0.00114</td>
<td>-0.00053</td>
</tr>
<tr>
<td>$\bar{X}\pm\text{SE}$</td>
<td>$1.245\pm0.034$</td>
<td>$0.00057\pm0.00019$</td>
<td>$-0.000397\pm0.000147$</td>
</tr>
</tbody>
</table>

Figure A2. Example of corrections to observed $\chi^2$ values in a study of temporal pattern in behaviour (see text for details). Two observations were made at sample sizes of 200 and 500 and they produced the observed values of $\chi^2$ when analysed using the pattern analysis program (●). The correction gradients for each datum (dashed lines) were calculated and the data corrected to a sample size of 350 (arrowed changes). ○: The resultant corrected data.
artefact effect of sample size on $\chi^2$ using the artefact gradient. We do this by correcting values to an emergence size of 350 bats (midway between the two emergence sizes). For the emergence of 500 bats the predicted artefact gradient = $1.2666 + 0.000703 \times 193 - 0.00035 \times 500 = 1.2272$. Given $\ln 193 = 5.2626$, $\ln 500 = 6.2146$ and $\ln 350 = 5.8579$, the difference between $\ln 500$ and $\ln 350$ is $-0.3567$. If we multiply this by the artefact gradient for this condition (1.2272) we get $-0.436$. Thus the corrected logged $\chi^2$ value equals $5.2627 - 0.4377 = 4.8250$, and the antilogged value (the corrected estimate of $\chi^2$) = 124.58. To correct the emergence of 200 bats we need to derive a second artefact gradient using the sample size of 200 values. Substituting these into the equation gives:

$$1.2666 + 0.000703 \times 117 - 0.00035 \times 200 = 1.2788.$$  
Correcting the emergence of 200 bats using this new gradient to a size of 350 bats gives a clustering value of 239.32.

In this hypothetical example, although the original $\chi^2$ values would lead to the conclusion that the larger emergence was more clustered, removing the size artefact effect revealed that the difference was entirely a statistical artefact of the sample size difference. Once this artefact effect had been removed it was the smaller emergence that was in fact more clustered.

We tested the ability of this method to ‘correct’ for sample size effects in two ways. First, we generated 10 new computer sequences based on completely different base structures, which included a wide variety of nonrandom patterns. We then sampled from these sequences at random between 100 and 900 observations, ran Clustan on them and corrected the goodness-of-fit statistic to an emergence of 500 bats. We then plotted the corrected values of the goodness-of-fit statistics against the original sample size. Since in these cases the effect of sample size was entirely due to the artefact, we would expect that the correction would remove the effect of sample size completely if it was working effectively. In all 10 cases the correction removed any effect of the original sample size on the goodness-of-fit. Second, we took two actual emergences from roost B which involved more than 800 emerging bats. We randomized the interevent interval sequence for these two emergences (to remove any time effects in the patterns of clustering) and then selected sequences of different size (between 50 and 500) for analysis. We corrected these different-sized sequences to an average emergence of 250 bats using the above procedure. In both roost emergences the procedure removed the size effect. The process is capable therefore of removing the size artefact in data from actual roost emergences as well as computer-simulated patterns. Because the artefact gradient approach was effective at removing the effect of sample size on the goodness-of-fit statistic across a range of computer-simulated patterns and two empirical patterns we consider they will be of general use for removing artefact effects of sample size in pattern analysis studies.

Because the artefact gradients never cross (Fig. A1) the choice of an appropriate sample size to which all the data should be corrected does not affect the ranking of the derived goodness-of-fit values. However, choosing a larger value for the correction size will produce greater absolute values and a greater range of values. The choice of the most suitable correction size is arbitrary, but a value close to the mean of the observed distribution of sizes would appear to be most sensible.

Finally, it is important to note that significant values of the goodness-of-fit statistics reflect only the significance of deviation of a given pattern from random and not ‘clustering’ per se. In the above computer-modelled cases we know that the patterns are clustered because we created them to be so. For unknown patterns, the goodness-of-fit statistics may represent clustering but they could also reflect other patterns of deviation from random, such as regularity. To confirm clustering is present it is necessary to demonstrate not only a significant deviation from random, but also that the exact pattern of deviation from random conforms to a clustered pattern. Such deviations involve an overrepresentation of the shortest and longest interevent intervals, reflecting events that occur within clusters and the intervals between them. Intervals of intermediate duration will be correspondingly underrepresented (see Speakman et al. 1992 for further discussion and testing procedures).