

# **Record of the Tracking Mammals Partnership Workshop on Statistical Analyses of Population Trend Data**

**Held at the Joint Nature Conservation Committee,  
Monkstone House, Peterborough  
12 July 2005 11.00-4.00**

## **Background**

The Tracking Mammals Partnership (TMP) consists of 24 organisations with an interest in carrying out surveillance on mammals to assess population and distribution change over time. The nature of the partnership has resulted in a variety of schemes being set-up, with the common approach to carry out surveillance, but with several differences in coverage of species, time-frames and geographical scales and different ways of analysing the data. All the schemes contribute interpreted data and results to an overview of general trends for a number of mammal species.

The first TMP report, published in March 2005, provided this overview, but also highlighted the problems associated with attempting to make meaningful comparisons of results between different surveys providing information on the same species. In some cases there may be up to six schemes potentially delivering information on one species, providing very large overall sample sizes. However, the sample sizes in individual schemes are often not sufficient to provide statistically significant trends and comparison of results is problematic because of different data collection and analysis methods.

There are many advantages in having a diverse approach to mammal surveillance and it is not envisaged as being practical, realistic or even desirable to unify all the schemes and have one survey method or overarching survey for mammals. However, there may be ways of producing a more standardised approach to data collection and analysis within the schemes, which allows for greater confidence in the comparison of the results. One of the ways of improving co-operation between the organisations in the TMP, in order to produce better surveillance information, is to standardise, where possible, the approach to statistical analysis of data and share best practice and statistical expertise. The purpose of this workshop is to facilitate exchange of ideas and information between statisticians, which hopefully will produce benefits for the individuals concerned and also for the TMP generally.

## **Objectives**

1. To achieve a more unified approach to the analysis and display of time-series results for TMP survey schemes.
2. To share expertise and knowledge between organisations concerning the advantages and disadvantages of using various statistical methods, approaches to dealing with problems encountered, etc.

## **Those attending:**

Nicholas Aebischer (GCT) NA

Stuart Ball (JNCC) SB

Jessa Battersby (JNCC) JB

Phoebe Carter (The Mammal Society Surveys Officer) PC

Colin Catto (BCT- NBMP Director) CC

Steve Freeman (BTO statistician) SF

Steve Langton (BCT Statistical Advisor) SL

Naebischer@gct.org.uk

Stuart.ball@jncc.gov.uk

Jessa.battersby@jncc.gov.uk

Pcarter@mammal.org.uk

Ccatto@bats.org.uk

Steve.freeman@bto.org

steve@slangton.co.uk

Martin Newman (IT consultant, PTES, BCT and TMS) MN

Stuart Newson (BTO - analysis of BBS data) SN

Martin@martin-newman.co.uk

Stuart.newson@bto.org

Simon Poulton (The Mammal Society's Statistical Advisor) SP  
David Roy (CEH, Butterfly Monitoring Scheme) DR  
David Wembridge (PTES – Mammals on Roads co-ordinator) DW

Simon@bioecoss.co.uk  
Dbr@ceh.ac.uk  
David@ptes.org

## **Discussion points covered:**

**1. Survey and analysis methods adopted for different schemes - why has a particular approach been taken? Is there an agreed best approach? Does it change depending on the number of years surveys have been running? What is the effect of outlier data points on trend estimates obtained?**

**See Appendix I for overview**

### **Steve Freeman, BTO Surveys**

Most BTO schemes involve generation of a site by species matrix. Because site monitoring is done by volunteers, sites tend to drop in and drop out of a scheme from year to year. Analysis is mostly concerned with trends in abundance with time, not with site to site variation. However, some analysis is done looking at differing trends by region and habitat. One of the main assumptions of analysis is that the trend is parallel on a log scale across sites – which is probably not realistic! The models used for mammals are the same as those used for birds, usually General Additive Models (GAM) or General Linear Models (GLM) with Poisson error distribution (appropriate for count data). GAMs are used to smooth trends, whereas GLMs will show full variation between years. There was no point in using GAMs with anything less than 10 years of data. Confidence limits will, in general, be closer together the longer the time series. Data for many species are not well fitted by these models, often showing too many zero counts for a simple distribution and hence (or otherwise) being overdispersed with respect to the Poisson.

Special methods are used in some cases, e.g. the Heronry Survey, which approaches a total census since the location of heronries are well known. In this case trend data can be combined into a full demographic model which merges survival data from ringing and productivity data from nest records. Such a population model gives much more scope for ecological interpretation and investigation of the reasons behind trends. The possibilities of merging data from ringing studies and nest records into population models for a wider range of species is an active area of research.

### **Stuart Newson - The Breeding Bird Survey (BBS)**

#### **(See BBS mammal analyses.pdf)**

About 2,000 sites are surveyed annually by volunteers. The sites are 1km squares which were chosen according to a stratified random design which attempts to account for differences in the density of observers. The method involves a line transect across the 1km square which is visited twice to make counts and once to record habitats. Mammal recording is a combination of counts of individuals observed for some species (e.g. hare, deer) and recording of signs for other species (e.g. fox droppings). Temporal trends have been calculated for mammal species which have been recorded in at least 40 different squares in at least two different years. Analysis is by a GLM (Log-linear with Poisson errors,

weighted for stratified sampling design). Mammal analysis suffers from over-dispersion, which is corrected and observed counts are used to predict missing counts. The trend is extracted by comparing the predicted index for the first and last year and significance is judged by looking for non-overlapping 95% confidence intervals. This causes problems if the first and last years are atypical. Overall trends are calculated and also trends for Government Regions and Environmental Regions. Presence absence data are looked at as well.

## **Steve Langton - The National Bat Monitoring Programme (NBMP) GAM approach and outliers**

**(see outliers. pdf )**

In the early years of the NBMP mixed models were fitted with either Poisson or log-normal errors and annual estimates produced. GAMs have been fitted to NBMP data since 2004. The amount of smoothing is generally set to the default of 0.3 times the number of years, as suggested in the original paper, but this is checked and changed if it is obviously inappropriate. Results are normally displayed as the estimated, smoothed trend with confidence intervals derived by bootstrapping. The index value in the first year is conventionally set to 100 and the smoothed estimate for each subsequent year will change every time the model is rerun as new data becomes available.

Although the smoothing function used in the GAM is quite robust to outlying data points, it does inevitable suffer from outliers in the first or last year. Such terminal outliers tend to cause the smoothed curve to turn up or down too much at the ends. Since the first year is set to an index value of 100 and all subsequent year's estimates are relative to this, an outlier in the first year can have a noticeable and continuing effect on the results. Similarly, if a trend is derived by comparing the first and last estimate, an outlier in the last year can bias the reported trend. This was demonstrated by simulation studies.

It is not uncommon for the first year's results in a survey to be atypical:

- the methodology is not yet well established (teething problems);
- observers are learning the ropes;
- it is not uncommon for fieldwork to start late because of the difficulties in getting funding and recruitment sorted out in time.

This can be countered by discarding the first year's results, and this is often done in analyses once a scheme is well established, but it is a difficult decision in the early years. Another possibility is to use the second or even the third year as the "base year" on which to base the 100 index against which all other estimates are shown.

Bias due to an outlier in the last year is less serious because its results are transitory. The smoothing functions used by GAM are known to be resilient to outliers in intermediate years, so the effect will disappear as subsequent year's data accumulates.

**2. Importance of distinguishing between (1) the estimation of annual indices of abundance (with confidence limits), which define the pattern of change from one year to the next at the same time interval as the data have been collected, and (2) the estimation of the amount of change (with confidence limits) that has taken place over**

**fixed time periods, e.g. 5, 10, 20 years, which integrate change over several years and provide alert limits.**

### **Nicholas Aebischer - Comparing GAM and GLM**

GAM and GLM both model trends over years by assuming a parallel pattern of change between sites. They essentially fit the same model to each site allowing the intercept to vary from site to site. This assumption is not necessarily true and is known to be untrue in a number of cases. The GAM applies smoothing to the results before the analysis and will produce different results from a GLM for the same dataset.

GAM produces a continuous line and a continuous envelope representing confidence intervals. GLM produces a series of annual estimates, each with its own confidence interval. This is a presentation difference. It is possible to get the annual estimates from a GAM and it is always wise to examine the annual estimates from any model and consider whether it is behaving reasonably.

It is relatively easy to compare GLM results for the same species resulting from different surveys. It amounts to asking the question “are the estimates parallel?” which is a well understood problem. The same cannot be said for GAM. As yet, it is not known whether models can be compared in this way.

SF noted that BTO know the parallelism assumption is false, but can't factor in causes of asymmetry (eg habitat, regional differences). Site factors use up degrees of freedom in analysis, so including them is a trade-off.

SL said that other models may be more appropriate, e.g. REML models with random site effects. These models may not be applicable for schemes with self-selected sites. There is an advantage of applying a standard model to all species, even though the best model will be different for different species.

### **GAM smoothing**

GAM smoothes and analyses the data points. This is a useful attribute of GAM because we are generally interested in the trend over a period of years, not in the detailed year to year variation. Various smoothing functions can be used, but the amount of smoothing is controlled by an “effective degrees of freedom” parameter. A value of 1 df results in a linear trend line (i.e. all the variation between years is smoothed out) whilst a value for df equal to the number of years will attempt to fit a curve which passes through each annual point (i.e. it preserves all the variation). What is required is a value somewhere in between. In practice, a value of  $n/3$  (where  $n$  is the number of years of data available) has been found to be about right. However, when the run of data gets long enough (e.g. the combined CBS/BBS data is up to 35 years),  $n/3$  can get too large and does not provide sufficient smoothing, so it is probably necessary to cap this value.

Suggested approach: *The basic GAM approach should be the standard for trend analysis, but other models should be considered, particularly where particular datasets have complications that do not fit easily into the standard framework. For purposes other than trend analysis, other models may well be more appropriate; for example in the NBMP we use mixed models for looking at the effects of covariates since these better account for the complex structure of the dataset.*

**3. How should we treat zeroes (or null entries) in survey data? Such entries can mean either that the species is absent, or that it is present but was not recorded. This usually means that assumptions have to be made before analysis can proceed, with consequences for the outcome of the analysis.**

When we visit a site to survey for a species, there are essentially two outcomes – we record it or we don't. But if we fail to record it, there are two possible reasons:

- It was absent,
- It was present, but we failed to detect it.

Therefore the probability associated with “0” has two components – the actual probability that the species is absent from the sites under survey and the probability of detection. Recent work has suggested ways of estimating these components (Zero-inflated Poisson Model) which could be incorporated into other models. However, it does require lots of data to work well – initial investigation suggest that at least three surveys of the same site per year are needed.

*There are actually two separate issues; firstly there is the distributional problem that Poisson models don't fit well due to an excess of zeros. Secondly there is the bias from undercounting, and in many ways this applies to all counts, not just zeros. ZIPs help with the first, but maybe not the second.*

*Simulations indicate that the standard model doesn't give any particular problems in this situation and so it should be used for the moment. However, ZIPs could potentially give improved results so there is a need for further research on this topic.*

**4. The possibility of combining different data sets to produce one trend for a species. How different can the surveys be and still be able to combine the data?**

**Steve Langton - Combining results from different surveys**

**(See combining datasets.pdf )**

There are many cases where two or more surveys cover the same species. Ideally we would like to pool the results in order to get longer time series, better precision of estimates and more statistical power. Is it possible?

There are usually no real problems in the technicalities of running a combined analysis, the problems are biological and methodological. Clearly the most important question is whether the survey methodology is compatible: “are they measuring the same thing?”. If they are, then it may be possible to fit one big model, although this is not necessarily straightforward.

Ask the questions:

- Are the two surveys estimating the same thing?
- Are there obvious differences between them?

If the answers to these two questions are “yes” and “no”, then it is worth trying to combine them in a single model.

Investigated the possibility of a single model for Daubenton's Bat by combining waterway counts and hibernating site census.

### **Discussion:**

JB noted that there is a need to report a single picture for each species, even when multiple surveys exist and that applying common standards for analysis and presentation wherever possible was very important.

SF felt it was easier to compare GLMs, as has been done for CBC – BBS comparison, but harder to compare GAMs.

NA noted that testing for survey x species interaction may be a solution.

SB noted that interpretation was a problem when multiple surveys exist, e.g. lapwings have ~9 surveys reported by BTO, with the 'best' survey given greater prominence in presentation. The new wildlife statistics website presents all surveys available and leaves interpretation to the users.

CC questioned whether surveys could be 'kite marked' in some way to aid comparison of results and provide guidance on which survey might be considered 'the best' for a particular species.

Different possible approaches to combining datasets:

1. The first approach, an approximation of the BTO joint indices produced for CBC and BBS data was proposed by SL. In this, annual indices are produced from the two (or more) separate surveys, from which the averages are taken weighted by the individual trend variances. Note that whilst the sampling variances are accounted for, the serial correlations between the annual trends estimates are not. However, it was felt that this was unlikely to be a serious shortcoming in most situations, and may be a worthwhile sacrifice if it saves considerable computer time.
2. The second approach, proposed by SF has already been used to combine the CBC and BBS. With this a joint model is produced for the CBC and BBS data simultaneously by multiplying together the separate likelihoods of GLMs fit to CBC and BBS separately. Many of the problems associated with combining very different surveys discussed in the workshop have been examined previously for CBC and BBS. A copy of the BTO research report based on this work to be circulated to all participants.

There were some concerns that while combining data sets may be possible, the question of whether it was desirable still remained. There were also concerns expressed about possibly 'cherry picking' surveys for combining and not applying stringent criteria. There was no point in combining surveys looking at different time periods, (other than for reasons of comparison over time such as between CBC and BBS) or very different samples, but there might be advantages to combining datasets covering the same time period and sampling generally the same populations of the same species.

Combining datasets should not replace the existing separate analysis of surveys, but should be carried out additionally and only where appropriate. It was also noted that such analyses could be carried out experimentally and need not be continued if it was deemed to have no added value. In terms of presentation of the data, though, JB felt it was important to standardise methods of analysis and presentation where possible.

**5. Alternative ways of analysing and displaying time-series data. Methods for interpolating information to extrapolate results. Are there other, novel ways of analysing the existing datasets to produce valuable information? Analysis of mammal**

**data against other datasets including habitat data and the implications for collecting and storing data.**

(See [BBS Mammal Analyses.pdf](#))

Stuart Newson (BTO) demonstrated attempts to produce a complete distribution map of various mammal species covered by the BBS by a combination of spatial interpolation and correlation with environmental layers (CoKriging). Stuart Ball reported that he had been attempting to map the “real” distribution of various insect species based on the rather *ad hoc* and random samples of “presences” reported by various National Biological Recording Schemes using a machine learning system (GARP – Genetic Algorithm for Rule-set Production). This technique attempts to identify a set of rules which predicts the presence of a species according to a combination of environmental variables.

This led to a discussion of the availability of environmental datasets too which these sort of modelling techniques could be applied. Some are listed in Appendix II.

## **6. Ways of assessing percentage change in presence/absence data as opposed to abundance data.**

When trends are based on population counts, or areas of range, a trend can easily be expressed and understood as a percentage change. It would be desirable for all surveys to make results available in this fashion. However, many surveys are essentially analysed in relation to the chance of a species being observed in a sample and the results reported are in terms of the change in “odds ratio”.

The odds ratio is the number of times an event occurs divided by the number of times it does not occur. For example, say we survey 300 sites and see Snarks at 100 of them (i.e we don't see Snarks at 200): Then the odds ratio =  $100 / 200 = 0.5$  If, over time, we find that the index produced by the survey has halved (suggesting that the odd ratio is now 0.25), how many sites now have Snarks?

If the number of sites had halved to 50, then the odds ration would be  $50/250 = 0.2$  NOT 0.25. In fact to get an odds ratio of 0.25, we need  $60/240$  – so the number of sites with Snarks has actually decreased by 40%. Say we had started by seeing Snarks in 200 sites out of the 300 (odds ratio =  $200/100 = 2$ ) and had subsequently got the same trend suggesting this had halved. To get an odds ratio of 1, we need  $150/150$  so the number of sites with Snarks has only decreased from 200 to 150 – a 16.67% decrease.

So, the actual change in the number of sites with Snarks which causes the odds ratio to halve depends on the proportion of sites that originally had Snarks and that proportion has NOT necessarily halved. So we cannot necessarily go straightforwardly from the trend reported by a survey to a percentage change in abundance of the organism.

### **Steve Freeman – odds ratio analysis of BBS presence/absence data**

For the analysis of mammalian presence/absence data from the BBS or WBBS, the approach adopted by the BTO has been to use logistic regression, fitting the same site + year model as for abundance data, and using GAMs to produce a smoothed trend. Statistically this is the best course of action, because it ensures that the underlying probability estimates remain between 0 and 1. However, it leads

to presentational problems because the year effects are measured in relative odds ratios on a logarithmic scale. It is possible to calculate the percentage change over a given time period, but the change is in the odds on a species being present. Because this is a quantity that is not immediately obvious to people other than gamblers or statisticians, it runs the risk of being confused with change in abundance.

In the ensuing discussion, it was suggested that an avenue worth exploring would be to use the fitted model to predict presence/absence at the beginning and end of the relevant time period across all 1x1-km squares in the UK, and to express the change as the percentage increase/decrease in number of occupied squares.

### **Main action points from meeting:**

1. Set-up email discussion group
2. Meet as a group periodically to discuss data analysis issues – maybe annually.
3. Consider use of GAMs in trend analysis, if not already using them, to provide more standardised presentation. Agree standard graphical presentation for reporting purposes.
4. Consider not using first year of data collection as baseline.
5. Consider possibilities of combining certain datasets, where appropriate and desirable.
6. Consider combining datasets for interpolated mapping.

### **References**

Freeman, S.N., Noble, D.G., Newson, S.E. & Baillie, S.R. 2002 *Modelling bird population changes using data from the Common Birds Census and the Breeding Bird Survey*. BTO Research Report 303, BTO, Thetford.

Kruger, H. B., 1969: General and special approaches to the problem of objective analysis of meteorological variables. *Quart. J. Roy. Meteor. Soc.*, **95**, 21-39.

## Appendix I

### INDEX METHODS FOR BAT POPULATIONS

*Note: this is an unpublished document which I started writing in 2004 when we first applied GAM models to National Bat Monitoring Programme (NBMP) data. I never finished it, but I've added various bits for the current meeting, based on our experiences fitting the models in 2004 and 2005.*  
Steve Langton 8/7/05 & 9/8/05

Initial analyses of NBMP data used mixed models with an overdispersed Poisson error structure for colony counts and a lognormal distribution for field surveys. Asymptotic standard errors were used, rather than bootstrapping at the site level. These methods were adopted to explore the data in the early period of the study when there were only a few years of observations available. As the data has accumulated we used the same models to produce indices by simply expressing the annual estimates from the model as a percentage of the value in a base year (generally the first year with a decent number of sites). With more years of data now available it is advantageous to move to an analysis specifically designed to produce indices, such as the Poisson Generalised Additive Models used by the BTO. However, this raises a number of questions, some specifically related to bats, others more general, which it would be useful to explore before adopting the BTO models.

1. **Overdispersion.** Bat data are heavily overdispersed relative to Poisson for good biological reasons – are the methods reliable and efficient with such data?
2. **Speed of fitting.** The models become very slow to fit with large numbers of sites using standard statistical packages. The BTO get around this by using specialist software, but the process is still slow and is also labour-intensive to set up initially, as an interface must be written between the stats software and the specialist software.
3. **Replicate counts.** The bird trends are generally based on a single figure at each site in each year. Where two counts are available, the maximum value is taken. We have two surveys which, in the case of the field surveys, each consist of 10 or 12 spot counts. How far should this structure be reflected in the analysis or can it be at least partially ignored? This is related to the previous question, since fitting speed may limit the ability to handle individual counts.
4. **Alternative models.** Other models might cope better with the overdispersion – for example a GLM with negative binomial errors, or a zero-inflated Poisson model. Should we be using random effects models, rather than a simple GLM?
5. **Outliers.** Is the method robust against outliers – either individual outliers (which relates to item 1) or sites following a different trend to the majority.
6. **Testing for trend.** The Fewster et al. paper suggests that testing should be based on the bootstrap confidence limits, but is this the best test, particularly if year to year variation is present in the counts (either due to variation in breeding etc. or due to variation in survey effectiveness)?

7. **Regional differences.** The paper suggests a deviance test is appropriate, or the usual F-test alternative where overdispersion is present – how do these tests perform.

## Methods

To explore these questions simulated data has been generated for 10 years of observation from a population of 20,000 roosts. The simulated data was produced in such a way as to resemble common pipistrelle roost counts from the NBMP in its overall mean, proportion of missing values and overall level of overdispersion. To do this it was assumed that the true number of bats in each colony really did follow a Poisson distribution, but that the overdispersion was created by the fact that sometimes some or all bats would not emerge to be counted, either because they remained in the roost or because they had moved to a different location. This was achieved by initially simulating data from a Poisson distribution with a different mean for each roost; the roost means themselves being generated from a log-normal distribution with variance estimated from a REML analysis of log counts from the NBMP data. The proportion of bats actually emerging was then simulated from a beta distribution with parameters  $a=0.5$  and  $b=0.3$ , which produces a bimodal distribution, with a lot of values close to 1 (i.e. most bats emerging), a few values in the mid-range and another large group close to 0 (i.e. most or all bats not emerging, perhaps due to moving roosts). No attempt was made to simulate any temporal correlation in the data at this stage.

To investigate the properties of the GAM index methods, a subset of 50 of the 20,000 roosts was selected at random to represent the roosts actually observed and the model fitted to this subset only. To simulate missing values within the observations for the selected roosts, the observed pattern of missing data in the real NBMP pipistrelle counts was superimposed on the simulated data. This process of selecting subsets was repeated a large number of times so that the estimates and their confidence limits could be compared with the true change in the population of all 20,000 roosts. The choice of 50 roosts, which is rather less than in most of our datasets was designed to ensure that the simulation didn't take too much computer time. Using 10 years data, rather than a longer period, had similar advantages, and is a reasonable time span to consider, given that the NBMP only began in the late 1990s.

The simulations were performed both without any temporal trend and with a decline equivalent to the red data alert value of 50% decline over 25 years.

## Results

### *Overdispersion*

Table 1 shows the results of simulations with and without overdispersion.

**Table 1: results of simulations to study the performance of the GAM method with overdispersed data.** Results are based on 2000 different random subsets each containing 50 of the 20,000 sites in the population, except for the coverage of the confidence limits which is based on 2000 random subsets. The significance of the bias is tested by means of a t-test against the null hypothesis that the mean is equal to the true mean.

Over-dispersion	Trend	Mean absolute error (mae)	Mean of estimates	True mean	Sig test for bias	Coverage 80% limits
No	No	2.42	100.1	99.95	P=0.018	71.5%

Yes	No	17.67	101.9	100.3	P=0.002	80.5%
No	Yes	1.97	77.92	77.85	P=0.234	74.5%
Yes	Yes	14.20	81.30	80.19	P=0.007	75.5%

The mean absolute error is clearly much higher with the overdispersed data, but this is merely because the data is inherently more variable, rather than representing any failure of the model. There does appear to be slightly more bias with the overdispersed data, but the level of around 1 percentage point is not sufficiently large to cause a serious problem. Coverage of the confidence limits is slightly non-conservative, with or without overdispersion, but this is not unexpected with bootstrap limits.

### *Speed of fitting*

Fitting the model can be very slow, which causes problems if bootstrap confidence limits are to be calculated, since the process needs to be repeated several hundred times. The problem is caused not so much by the spline function of time, but by the number of parameters that must be fitted to allow for the site effects. This is important because in data collected by volunteers the runs of data tend to be comparatively short, so that the total number of sites in a study may be several times greater than the maximum number in any single year.

One potential alternative is to fit a model with year as a factor in order to obtain estimates of mean counts in each year and then to fit the GAM as a separate model to these annual estimates, weighted by the inverse of their variances if the number of sites differs substantially between years. This does not save much computing time in itself, but in packages such as Genstat the annual means model can be fitted as a ‘within groups’ regression, with site as a grouping factor, whereas this is not possible with the GAM. The within groups regression produces identical results to the ordinary approach but takes a fraction of the time for large numbers of sites. Unfortunately, not all variances and covariance terms are estimated, which means that the variances of the annual means are not known, but approximations can instead be used for weighting the GAM model. Various approximations to the weights have been tried but the most successful is to use the variances of annual estimates taken from a simplified model without fitting the site terms. The sequence of models is thus:

1. Fit a within groups Poisson GLM with site as the grouping factor and year fitted as a factor. Use the parameter estimates from this model to calculate fitted means (on the log-scale) for each year.
2. Fit a Poisson GLM with year as a factor but no site terms and extract the variances of the fitted annual means.
3. Fit a GAM to the fitted means from model 1, weighted by the inverse of the variances from model 2.

**Table 2: results of simulations to compare the published GAM method with the faster alternative.** Results are based on 2000 different random subsets each containing 50 of the 20,000 sites in the population. The significance of the bias is tested by means of a t-test against the null hypothesis that the mean is equal to the true mean.

Model	Trend	Mean absolute error (mae)	Mean of estimates	True mean	Sig test for bias
Published	No	17.67	101.9	100.3	P=0.002
Fast version	No	17.88	102.0	100.3	P=0.001
Published	Yes	14.20	81.30	80.19	P=0.007

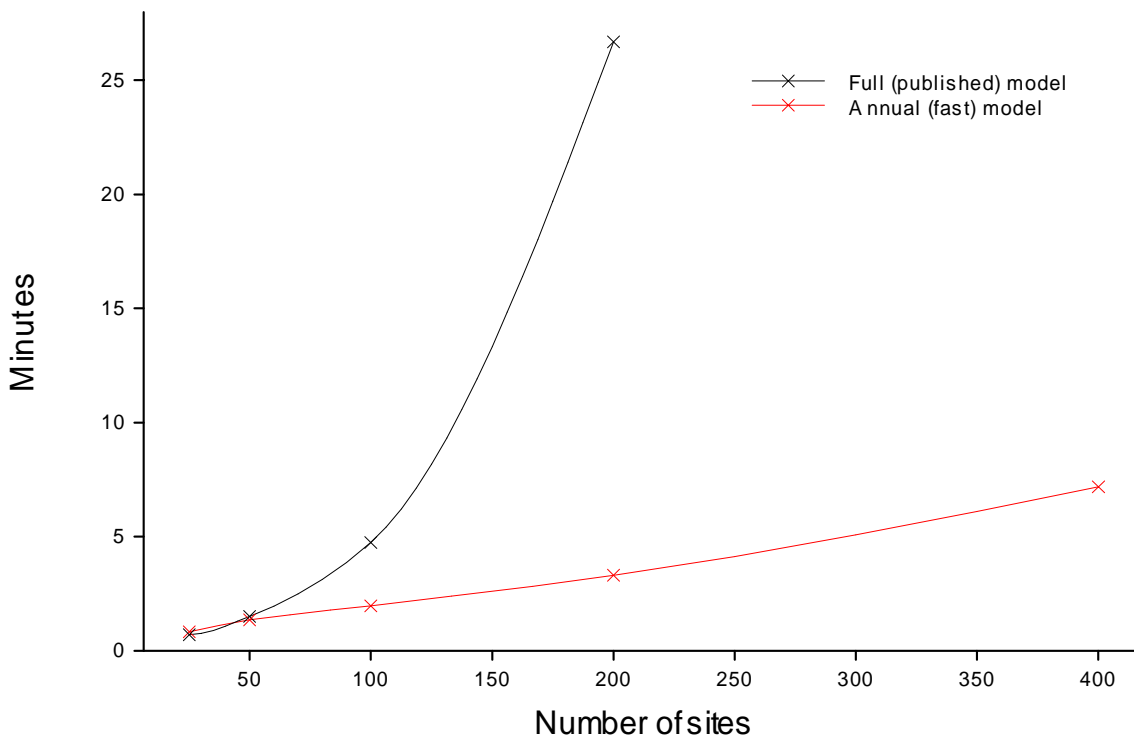
Fast version	Yes	14.40	81.41	80.19	P=0.004
--------------	-----	-------	-------	-------	---------

Table 2 shows results of simulations to compare the published method with the alternative; it can be seen that there is a minimal loss of precision from using the faster method. Figure 3 indicates the difference in speed between the two methods and shows that the computing time required to fit the model increases exponentially for the published method, whereas the fast method shows only a modest increase in time with increasing numbers of sites.

Besides the obvious advantages that the fast approach has in terms of speed of fitting to large datasets, it also allows a variety of alternative models to be adopted. The published method can only be applied using a model that can accommodate the GAM approach, whereas the fast version can be used with any estimation approach that generates unbiased annual means with variance estimates (or an approximation to them) for use in the GAM fitting stage.

Whilst the fast method generally produces very similar estimates to the full analysis, substantial differences are sometimes found when the pattern of missing values is very non-random, e.g. if different sites were recorded in different years. In this situation the covariances between the annual estimates become important and they are lost when the fast method is used.

**Figure 1: speed for various numbers of sites for 400 bootstrap samples.** All points are based on 10 years of simulated data, using the 8<sup>th</sup> edition of Genstat for Windows running on a PC with a Pentium 4 processor.



*Replicate counts*

I haven't done any simulations on this problem. However, experience with the roost count data suggests that using the maximum of the two counts at each site in each year often gives shorter bootstrap confidence intervals for the index values than using both values. This is particularly true for species such as Common Pipistrelle, presumably because it lessens the impact of zero counts, whereas the Lesser Horseshoe data produced slightly wider limits using the maximum count. I haven't tried using the maximum count with survey data, but my suspicion is that using the maximum would lead to some loss of efficiency, and hence wider limits, in this case.

*Alternative models*

Limited work done on this. Mixed models tend to give bias in the estimation of annual means and aren't therefore good for this purpose, although I still use them for exploring the effects of covariates.

*Outliers*

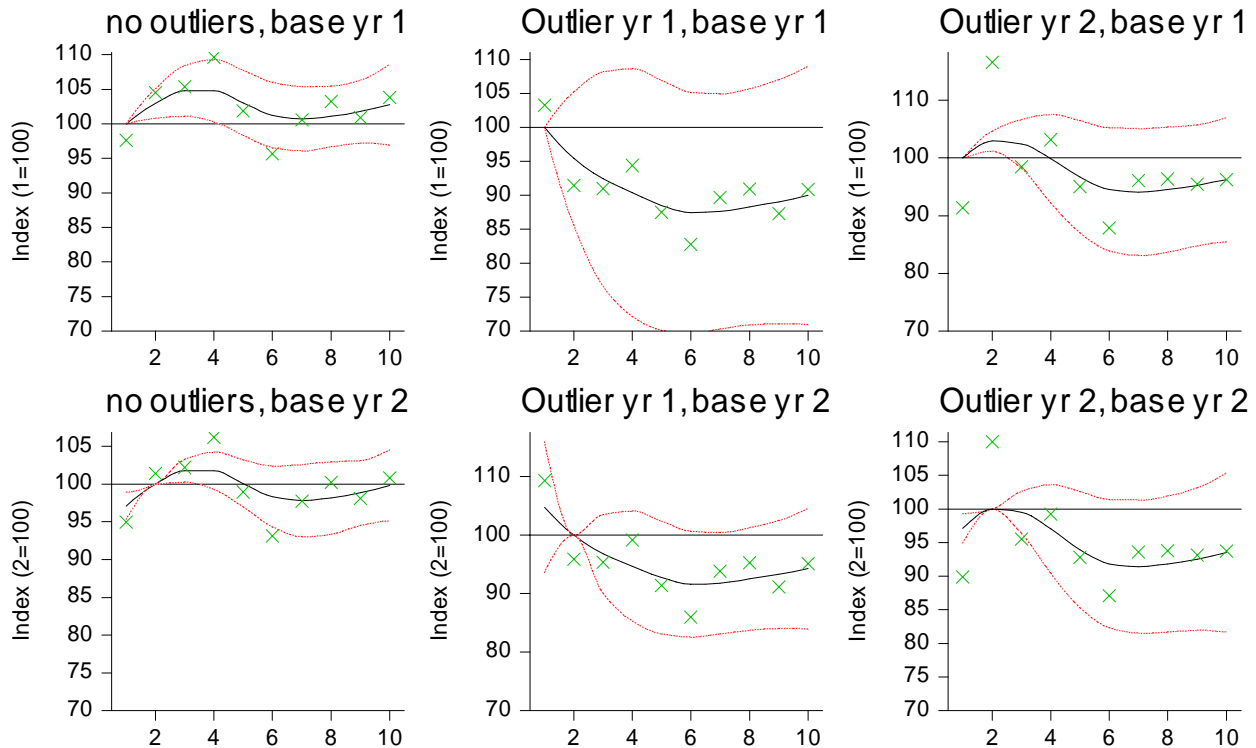
Generally the methods seem fairly robust to outlying data values. In particular, the smoothing effect of the GAM ensures that outliers (or indeed, atypical years) in the middle years of a survey have little impact on the trend estimate. Outliers in the first or last year of the data can have more impact as the GAM curve can turn up or down to accommodate them, without the stabilising effect of the years either side. This poses a particular problem when (as is usually the case) the first year is taken as the base year of the survey.

To investigate this, outliers were added to simulated Poisson data by doubling the value of the highest observation in the appropriate year. Table 3 shows the mean absolute error from 2,000 simulations using various combinations of the base year for the index and the year in which an outlier is added. Adding an outlier to year 1 when this is also the base year has by far the largest impact. Using year 2 as the base year ensures that the outlier does not have such a large impact, whether it occurs in year 1 or year 2. Figure 2 shows the first of these simulated datasets to give a visual impression of why this is happening; when the first year is the base year it has a major impact on the shape of the curve, as the curve is free to bend up towards the outlying value. By contrast, when year 2 is the base year, the curve is constrained by the values in years 1 and 3, with the result that an outlying year 2 value (or indeed an outlying year 1 value) does not have such a large impact.

**Table 3: The effect of outliers in the base year. Mean absolute error refers to the mean difference between the observed value in year 10 and the expected value of 100 (there was no trend in the data before addition of the outlier).**

<b>Outliers</b>	<b>year used as base</b>	<b>mean absolute error</b>
none	year 1	2.384
none	year 2	2.054
Outliers in year 1	year 1	12.11
Outliers in year 1	year 2	5.942
Outliers in year 2	year 1	6.512
Outliers in year 2	year 2	6.565

**Figure 2: The effect of outliers in the base year.** For illustration purposes, these graphs show one of the 2000 simulations contributing to the results in Table 3. Red lines are 95% limits.



The other point to note from Table 2 is that even without the outliers, a considerably lower mean absolute error for the index at year 10 is obtained when the second year is treated as the base. It therefore looks like it is wise not to use the first year of data as the base year, even if there is no reason to suppose that outliers are present.

### Testing for trend

The bootstrap testing procedure essentially tests the consistency of the trends at different sites – in a GLM context it is equivalent to testing the trend against the site.year error term. Where year-to-year differences are substantial, an alternative would be to test trend against the between years variation – otherwise the between site bootstrapping would tend to indicate significant trends that were actually just due to a random sequence of ‘good’ or ‘bad’ years. This is unlikely to be a problem with bats, but might be an issue with more r-selected species, such as shrews.

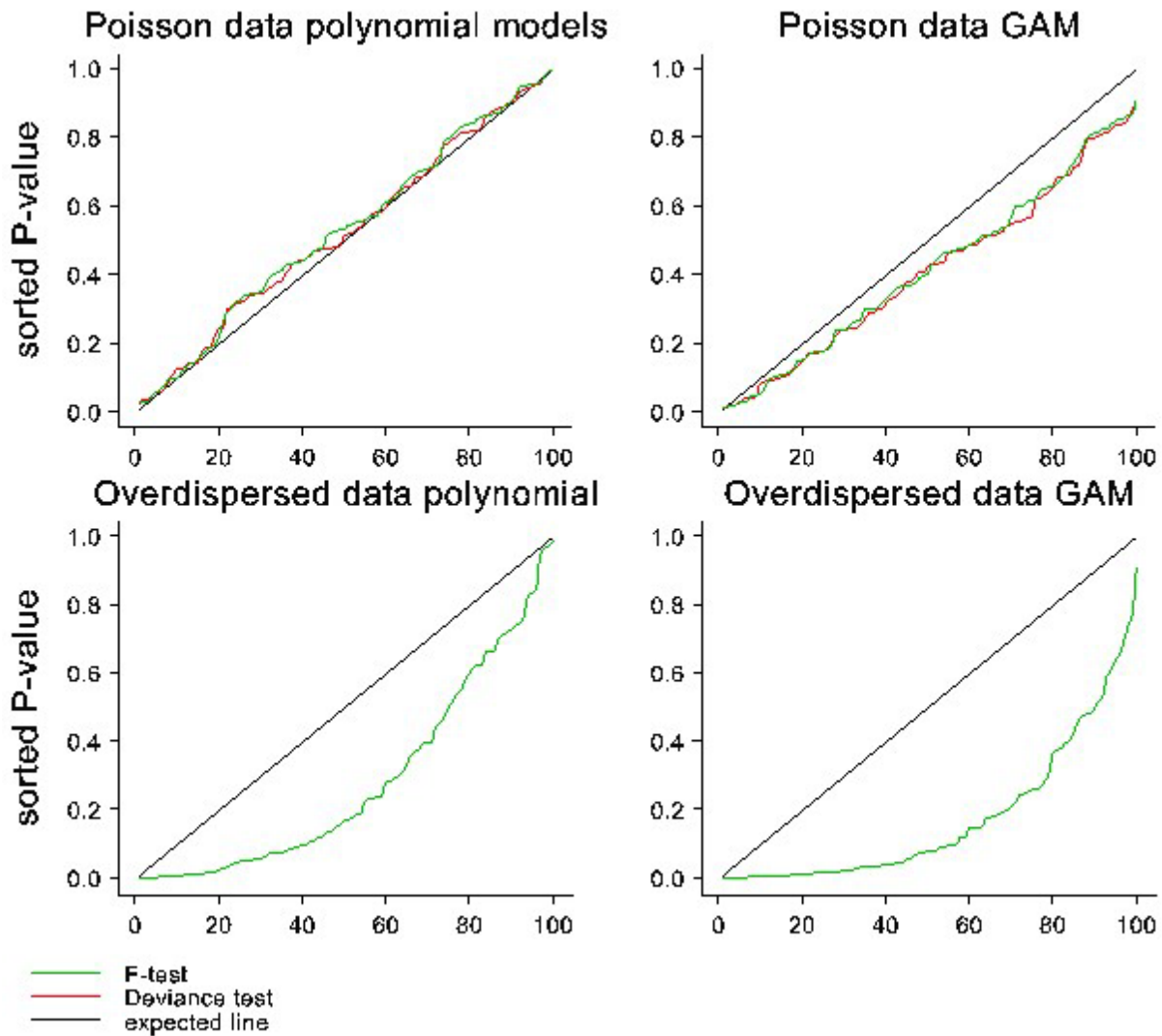
Year-to-year variation also raises issues regarding the base year. The base value in the GAM approach is not the base year itself, but rather the smoothed fit of the GAM at the base year. For the reasons discussed in the previous section, the smoothing effect will be greater if the base year is not at the end of the sequence (i.e is not the first year of the survey).

### Regional differences

The tests proposed in the paper were assessed using the same simulations, but selecting 200 sites at random from the ‘population’. At each loop these sites were assigned at random to 3 regions, one with 100 sites and the other two with 50 each. Figure x plots the sorted P-values for tests of the interaction between year and region, using an F-test and, in the case of simulated data from a Poisson distribution with no overdispersion, the deviance test. When the temporal trend is fitted as a polynomial, rather

than a GAM, both tests perform well as expected. With the GAM the tests are slightly non-conservative, but are quite adequate for model assessment; this confirms the recommendations in Fewster et al. However, when extreme over-dispersion is present, the deviance test cannot be used, and the F-test is severely non-conservative, particularly with the GAM model. I have therefore continued to use Wald tests from mixed models to check for regional differences. *Note: this result is surprising and I need to check it to ensure that it is not the result of a programming error.*

**Figure 3: simulation to check the proposed tests for interactions between temporal trend and region (c:\data\bats\indexsim\regionsim)**



## Appendix II

### Environmental datasets too which interpolated mapping techniques could be applied

#### Weather

The UK Climate Impacts Programme (<http://www.ukcip.org.uk/>) makes available datasets on a 5 x 5km grid for each month for the years 1960 to 2000 for a wide variety of weather variables. These are available free for research use – you just need to sign up to an agreement concerning the use of the data and the citation of UKCIP in any resulting publications.

Results for 2020, 2050 and 2080 are also available from models for four climate change scenarios produced by the Intergovernmental Panel on Climate Change. These are also on a 5 x 5 km grid, but cover a more limited range of variables.

#### Land cover

The CS2000 land cover maps are available via the Countryside Information System (<http://www.cis-web.org.uk/home/>). The software is freely available via a request made via the web-site and includes the CS2000 maps as part of the core system. They are at 1 x 1km resolution and show the number of hectares of each Countryside Survey land cover class in each 1km square. The data can be downloaded from CIS in text format and readily loaded into another system. ITE Land Class also forms part of the core data supplied with CIS.

Note that the CS90 Land Cover data is also available in older versions of CIS, but the classification system is different and the methods of deriving the maps from satellite images changed drastically – so the two land cover maps are not comparable! Work is ongoing at CEH to produce comparable coverages and statistics about changes.

#### Topography

The OS Strati dataset is included with CIS. This includes the mean, minimum, maximum, 10-percentile and 90-percentile altitude per 1 x 1 km square and a measure of slope per 1km square. Also includes areas of town, villages, woodland, “open countryside”, lengths of motorways, A, B and “other” roads, rivers, canals and railways per 1km square.

#### Agriculture

Access to the UK Agricultural Census data on a gridded basis is being made available via Edinburgh University (<http://edina.ac.uk/agcensus/>). This is still under development so it is not entirely clear what resolutions will be available, but 10km square gridded data is mentioned. A large number of variables including acreages of a wide range of crops and numbers of stock are covered. The access situation is complex with categories for academic and other researchers, policy users and individuals – so you probably need to look at it and assess it yourself!

Data by administrative areas is available from the agriculture departments in England, Scotland, Wales and NI. See [http://www.defra.gov.uk/esg/work\\_html/publications/cs/farmstats\\_web/default.htm](http://www.defra.gov.uk/esg/work_html/publications/cs/farmstats_web/default.htm) for English data. Local data (including holding level information) is also available to researchers subject to confidentiality agreements. CSL are currently working on a 1km dataset derived from IACS

(subsidy) data which will be more precise than the EDINA data which is based on interpolated parish-level June survey data. Contact Steve Langton for more detail on any of these issues.

### Soil

The National Soil Resources Institute, Cranfield University makes available soil survey data on 1, 2 x 2 and 5 x 5 km grided basis (<http://www.silsoe.cranfield.ac.uk/nsri/services/cf/gateway/ooi/natmap.cfm>) but it is expensive!

### Geology

The British Geological Survey make geological map data available in raster format (<http://www.bgs.ac.uk/products/digitaldata/licencefee.html>) but it is expensive!

### Population

The 2001 census data is publicly available, but relates to census regions which are defined as polygons and are highly variable in size. Converting this to a gridded dataset is not trivial! The National Office of Statistics would be prepared to do this – but at cost (we haven't even bothered asking how much!).

## Appendix III

### Using fitted values to express trends in presence/absence data

**Steve Langton**

*Summary:* If we can estimate the absolute proportion of sites occupied at some point in time, the trend from a GAM can be applied to this figure and the results can be presented in a similar way to indices of count data without any mention of odds ratios, etc.

Consider the following small example, derived from a subset of common pipistrelle records:

year	1	2	3	4	5	6	7
site							
1	1		1	0	1		1
2	1	1	1	1		0	
3	1		1	0	0		
4	0	1	0	0			0
5		1	1	0			0
6	0	1	0	0		0	
7	1	1	1	0			
8		1	0		1	1	
9	0	1		0		0	0
10	1		1	1			0
11	1			1	0	1	1
12			0	1	1	1	0
13	1		0	1			0
14	1			1	1	0	1
15			0		1	1	1
16	1			1		0	1
17				1	0	0	0
18	1		1			0	0
19		1		1	0		0
20		1	0	1	1		1
Mean proportion present	0.77	1.00	0.50	0.59	0.60	0.36	0.40

**Table; 1 example data (1 is presence, 0 absence, blank no data)**

If we then fit a logistic regression model to the data, here using a GAM with 2 d.f. for the trend, plus fixed site effects, we can produce a fitted proportion for each site:

year	1	2	3	4	5	6	7
site							
1	0.98	0.96	0.91	0.83	0.73	0.63	0.56
2	0.97	0.93	0.85	0.74	0.62	0.51	0.43
3	0.85	0.73	0.54	0.36	0.24	0.17	0.13
4	0.47	0.29	0.15	0.08	0.05	0.03	0.02
5	0.89	0.79	0.61	0.44	0.30	0.22	0.17
6	0.46	0.29	0.15	0.08	0.05	0.03	0.02
7	0.92	0.85	0.70	0.53	0.39	0.29	0.23
8	0.97	0.94	0.87	0.77	0.65	0.54	0.46
9	0.51	0.33	0.17	0.09	0.05	0.04	0.03
10	0.97	0.93	0.86	0.74	0.62	0.51	0.43
11	0.99	0.97	0.93	0.87	0.79	0.71	0.64
12	0.96	0.92	0.83	0.71	0.58	0.47	0.40
13	0.87	0.76	0.58	0.40	0.27	0.20	0.15
14	0.99	0.97	0.93	0.87	0.79	0.71	0.64
15	0.98	0.97	0.92	0.86	0.77	0.69	0.61
16	0.98	0.96	0.91	0.83	0.73	0.63	0.56
17	0.87	0.76	0.57	0.39	0.27	0.19	0.15
18	0.90	0.81	0.65	0.48	0.34	0.25	0.19
19	0.92	0.85	0.70	0.53	0.39	0.29	0.23
20	0.98	0.96	0.91	0.83	0.74	0.64	0.56
Mean fitted proportion	0.87	0.80	0.69	0.57	0.47	0.39	0.33

**Table 2: fitted proportions from the logistic regression model, fitting site effects and 2 d.f. smoothed trend for years.**

People often try to convert such data back to presence/absence by applying a threshold, with any site above the threshold being present. This has two disadvantages:

- The choice of threshold is problematic, and the natural choice of 0.5 is often not appropriate.
- It loses some information, although this may not matter, depending on the purpose. For example a probability of 0.99 is clearly very different to 0.51, but with a threshold of 0.5 they both count as present.

A better alternative for our purposes of trying to illustrate the fitted GAM trendline is to consider the problem from a randomisation and simulation perspective. We simulate presence and absences from the fitted proportion, so that, for example, if the fitted proportion is 0.40, we assign it to be present with probability 0.4. This might produce the following:

year	1	2	3	4	5	6	7
site							
1	1	1	1	1	1	1	1
2	1	1	1	1	0	0	0
3	0	1	0	1	0	0	0
4	0	0	1	0	0	0	0
5	1	1	1	0	0	0	0
6	0	0	0	1	0	0	0
7	1	1	1	1	0	0	0
8	0	1	1	1	1	0	1
9	1	1	0	0	0	0	0
10	1	0	1	1	0	1	0
11	1	1	1	1	0	1	1
12	1	1	1	1	1	1	0
13	0	0	1	0	1	0	0
14	1	1	1	1	0	1	0
15	1	1	1	1	0	0	0
16	1	1	1	1	0	0	0

17	1	1	1	0	0	0	0
18	1	1	0	0	1	0	1
19	1	0	1	1	1	1	1
20	1	1	1	1	1	1	0
Mean simulated proportion present	0.75	0.75	0.80	0.70	0.35	0.35	0.25

**Table 3; one simulated randomisation from the fitted model**

Notice that I have simulated values even where we have real data (e.g. site 8, year 6 is simulated as absent, even though it was really present). This is because we are trying to simulate what the data would look like under the fitted smooth trend line, not what it really is like. Now the example above is just one possible outcome and will be subject to chance distortions, so rather than relying on one randomisation to display the trend, we do several hundred simulations and use the mean proportion present in each year, averaged over all sites from all simulations to illustrate the fitted trend from the model. The table below shows averages from 2,000 simulations:

year	1	2	3	4	5	6	7
Mean simulated proportion present	0.87	0.80	0.69	0.57	0.47	0.39	0.33

**Table 4; average proportions from 2,000 simulations from the fitted model.**

Thus under the fitted smoothed trend (GAM) model we estimate that the proportion of sites occupied has fallen from 87% in year 1 to 33% in year 7. I would suggest that this explanation has considerably more intuitive appeal than any approach based on odds ratios and is also largely analogous to what we do with count data.

Whilst it is helpful to go through this randomisation approach to explain the rationale behind what we are doing, it is not necessary to do this in practice because comparison of the table above with the mean fitted proportions from Table 2 shows that they are identical. In other words the mean of the fitted values from the model can be interpreted as being the average proportion of sites that would be occupied based on a large number of simulations from the fitted model (yes, I know this is probably self-evident to statisticians, but is probably not to others!).

In the above, I have just dealt with a small number of sites and in reality the number would be much larger (probably at least several hundred). If these sites can reasonably be regarded as being representative of all possible sites in the country, then the mean fitted proportions are a sensible way of describing the national trend. In some cases some weighting might be applied when averaging the fitted values to make them more representative of the national picture, as is sometimes used with counts data. Alternatively, if the presence/absence at the observed sites can be modelled using site specific covariates, then it should be possible to model for all sites in the country – for example, if the sites are surveys of 1km squares, we can predict for all squares in the country.

If we can generalise to the whole country in this way, then we can calculate a number of sites occupied in the first and last year of the survey and a simple percentage decline between the two figures, so that the whole trend can be summarised by saying that ‘The numbers of sites where pipistrelles were detected has declined by 62%  $((0.87-0.33)/0.87)$  over the 7 years of the study’ (note the use of ‘were detected’ to allow for the fact that there will be sites where they were missed). Indeed there is no reason why the number of sites occupied could not be displayed as an index, thus making the presentation look identical to the indices from counts data. Bootstrapping could be applied to the process to produce confidence limits as with the GAMs for counts.