Appendix D. Assessing the impact of zeros on GEE analyses

Steve Langton
Defra Environmental Observatory

Data collected in this project contains a significant proportion of zeros, even for the more common species. For the A1 Dishforth to Leeming dataset, the proportion of zeros is 33% considering all passes, and in excess of 50% for all the individual species apart from *Pipistrellus pipistrellus*. When the data is analysed using REML or GEE, this gives rise to a residual plot with the line of points representing the zeros (Figure C1a), although a histogram of residuals can look approximately normal (Figure C1b).

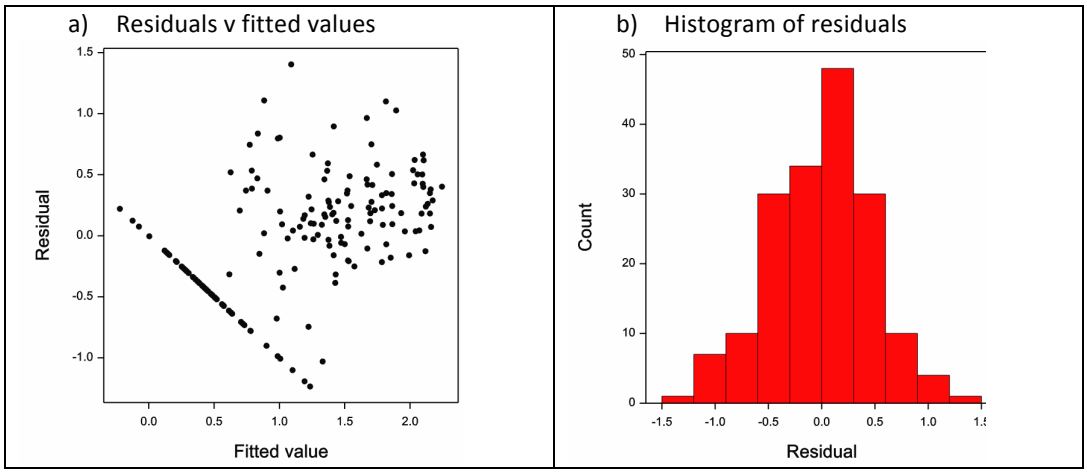

Figure C1. Residual plots from a REML analysis of log passes+1 for the A1 dataset (all species combined).

This raises the question of whether GEE or REML analyses are robust against this type of distributional issue and, in particular, can these analyses produce reliable results for the less abundant species where the proportion of zeros is very high? This piece of work uses simulated data to explore these questions.

Methods

Data was simulated based on the 'all passes' data for the A1 Dishforth to Leeming section. A REML model was fitted to the log-transformed passes+1, fitting survey routes and days (replicates) within routes as random effects and habitat score, days and linear distance as fixed effects. The model was fitted with and without an autoregressive error term to estimate the spatial correlation between survey points.

Simulated datasets were then generated from a normal distribution using the variance components and correlation from the REML model. The normal simulated data was then back-transformed and rounded to the nearest integer. The lognormal distribution very occasionally produced unrealistically high counts of passes and so an upper limit of 999 passes was applied (about 20% above the maximum of the real data). The distribution of the simulated data was then compared with the real data using cumulative probability plots. The simulated data had a lower proportion of zeros than the real data (Figure C2), but otherwise matched the distribution well.

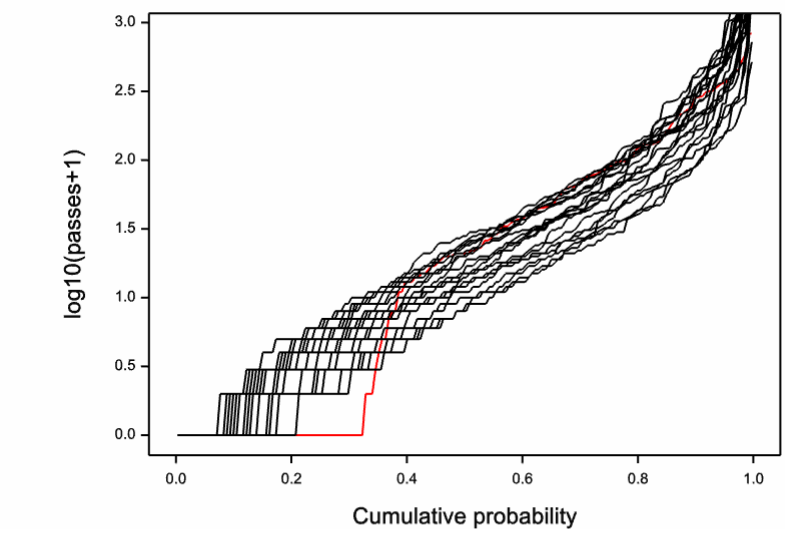

Figure C2. Cumulative probability plot of the real data (red line) and 20 sets of simulated data, before any adjustment to increase the proportion of zeros.

In order to simulate the rarer species and to ensure a high proportion of zeros, two adjustment factors were applied:

- An overall adjustment reducing average abundance to a tenth or a fiftieth of the all passes (i.e. subtracting 1 or 1.699 from the log data)

- Adjustment on a sliding scale (linear on log-scale), with a reduction to a tenth or a fiftieth as above for zero values, down to no reduction for counts of 100 or more. This increases the proportion of zeros appreciably, whilst having less impact on the overall mean and the numbers of high counts.

This approach of adjusting the 'all passes' simulated data was preferred to directly simulating from the rarer species data due to the difficulty in obtaining robust estimates of variance components and correlations with very sparse data.

A thousand sets of simulated data were generated and then analysed using either REML or GEE. All analyses were conducted in Genstat for Windows (www.vsni.co.uk/software/genstat/) apart from the GEE analysis which used the same R program used for the analyses in the main report.

Results

The first sets of simulations were run without any distance effect being included, so that the estimated linear effect of distance should average zero, with approximately 5% of simulations being statistically significant at P=0.05. Results are shown in Table C1a. REML tests of significance tend to be slightly conservative, with less than the expected 5% of significant values at P=0.05, whilst GEE is very close to the nominal value, perhaps becoming very slightly non-conservative with very high proportions of zeros. Mean estimated effects are always close to, and never significantly different from, zero.

Table C1. Results of simulations. Each row is based on 1000 simulations. Mean distance effects are shown multiplied by 1000 to reduce the number of zeros.

a) with no added difference effect

No.	Overall adjustment	Sliding adjustment	Mean passes	Prop'n zeros	% sig at P=0.05		Mean distance effect x 1000	
					REML	GEE	REML	GEE
A	0	-1.0	71	0.33	2.3%	4.4%	0.0002	0.0016
B	-1.0	-1.0	9	0.63	2.9%	5.4%	0.0005	0.00096
C	-1.699	-1.699	1.7	0.81	2.7%	6.4%	-0.0005	-0.0004

b) with distance effect of 0.0006

No.	Overall adjustment	Sliding adjustment	Mean passes	Prop'n zeros	% sig at P=0.05		Mean distance effect x1000	
					REML	GEE	REML	GEE
D	0	-1.0	74	0.33	33.9%	44.2%	0.42	0.42
E	-1.0	-1.0	10.0	0.63	31.5%	39.0%	0.26	0.25

With a distance effect added with values close to the estimated effect for all passes both REML and GEE detect the effect in 30-45% of simulations, with GEE performing rather better. There is a downward bias in the estimates of the distance effect.

Conclusions

The significance tests from REML and GEE are not unduly affected by the presence of multiple zeros in the data, at least in terms of their type 1 error rates. There may be an issue with the analyses failing to detect distance effects when they are present, and downward bias in estimation of the effect; a comprehensive power analysis study would be needed to assess this. Nevertheless, for the purposes of the current study, the GEE analysis is satisfactory and is unlikely to indicate a distance effect, unless one is present.