Animal Behaviour 154 (2019) e1–e5

Contents lists available at ScienceDirect

Animal Behaviour

journal homepage: www.elsevier.com/locate/anbehav

Forum Article

Accounting for pseudoreplication is not possible when the source of nonindependence is unknown

Paolo Gratton[*], [1], Roger Mundry[**], [1]

Max-Planck Institute of Anthropology, Leipzig, Germany

ARTICLE INFO

Article history:
Received 10 October 2018
Initial acceptance 2 November 2018
Final acceptance 9 December 2018
Available online 11 July 2019
MS number 18-00741

Keywords:
GLMM
LMM
mixed models
pseudoreplication
repeated observations
type I error
unknown subject identity

Repeated observations of the same individuals or other units, which can lead to clustered observations, are common in animal behaviour research, and mixed models are commonly employed to model and account for such clustering in the data and avoid pseudoreplication. However, in some cases, while the data might comprise repeated samples from the same individuals, the precise identity of the individuals from which samples originated is unknown. In a recent paper Garamszegi (2016, *Animal Behaviour*, 120, 223–234) suggested an approach to account for pseudoreplication which is based on repeatedly assigning random subject identities to the samples and then analysing the data using a mixed model or averaged values for each randomly assigned identity. Here we tested this approach using a simulation study. We found that the approach suggested by Garamszegi leads to clearly inflated type I error rates that were essentially the same as those obtained from a naïve linear model simply ignoring individual identity and that only a model based on the correct subject identities roughly produced the nominal type I error rate. We conclude that, currently, there is no method available that allows pseudoreplication to be controlled when subject identities are unknown.

Data sets about animal behaviour frequently encompass multiple observations of the same individuals, for instance due to limited accessibility or availability of study individuals. In such a case one has to take care to avoid pseudoreplication (Hurlbert 1984; Machlis, Dodd & Fentres, 1985). Nowadays, frequently used statistical approaches to avoid pseudoreplication are linear mixed models (LMM) and generalized linear mixed models (GLMM; e.g. Baayen, 2008; Bates, Maechler, Bolker, & Walker, 2015; Bolker et al., 2009). These allow the effects of fixed and random effects predictors to be disentangled. More precisely, for a fixed-effects predictor (such as 'sex', 'age' or 'nutritional state') they determine how much the response changes when these increase by one unit, while for a random-intercepts effect they determine how much the response varies due to variation between the levels of the random-effects factor (e.g. 'individual'), and for a random-slopes effect they determine how much the effect of a fixed-effects predictor varies between the levels of a random-effects factor (e.g. Schielzeth & Forstmeier, 2009; Barr, Levy, Scheepers, & Tily, 2013). Finally, it is also possible to model the correlations between random intercepts and slopes.

A problem, though, arises when the identity of the individuals from which observations were made or samples collected is unknown, as is frequently the case when unmarked and unhabituated animals are investigated (e.g. Hadinger, Haymerle, Knauer, Schwarzenberger, & Walzer, 2015). In a recent paper containing several practical suggestions about dealing with practical limitations typically encountered in behavioural studies, Garamszegi (2016) proposed that the pseudoreplication issue entailed by unknown identity of subjects can be tackled by simulation. Specifically, he suggested two possible solutions. One consists of random assignment of individual identities to samples and the other consists of considering the (spatial or temporal) autocorrelation in the response. Since the latter approach obviously does not work for group-living or highly mobile animals (which rapidly change their spatial configuration to a large extent) the former approach seems more widely applicable. Regarding this, Garamszegi (2016) suggested that after random assignment of individual identities to samples one has two options to avoid pseudoreplication: (1) analyse the averages per (randomly) assigned individual or (2) use a (G)LMM with the (randomly) assigned individual identity included as a random effect. Since each specific random assignment

[*] Correspondence: Max-Planck Institute of Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany.

[**] Correspondence: Max-Planck Institute of Anthropology, Leipzig, Germany.

E-mail addresses: paolo_gratton@eva.mpg.de (P. Gratton), roger_mundry@eva.mpg.de (R. Mundry).

[1] Authors contributed equally.

https://doi.org/10.1016/j.anbehav.2019.05.014
0003-3472/© 2019 Published by Elsevier Ltd on behalf of The Association for the Study of Animal Behaviour.

is just one of many possible assignments, and since there is likely to be uncertainty about the true size of the population from which samples were taken, he further suggested repeating the random assignment multiple times and averaging the results.

A problem arising with the suggested approach is the size of the population from which to sample individual identities. Garamszegi (2016) suggested two possible approaches representing the endpoints of a continuum: (1) identities are assigned by sampling random numbers from one to the number of observations (or even a larger number) or (2) sampling from a population size representing a fair guess of the actual population size. The former rests on the assumption that the probability of sampling the same individual multiple times decreases with an increasing size of the population sampled from, whereas the latter is presumably more realistic when the actual population size is small as compared to the sample size.

Here we investigate whether taking the approach proposed by Garamszegi (2016) allows one to effectively avoid pseudoreplication. To this end we conducted a simulation study. That is, we simulated data randomly sampled from a limited number of individuals whereby we varied the amount of variation between individuals and the population size. We used both methods proposed by Garamszegi (2016; averaging and use of an LMM, both after random assignment of individual identities) and compared them with a 'naïve' linear model (ignoring potential pseudoreplication) and a ‘correctly informed’ LMM controlling for the identity of individuals. We assigned random identities under two scenarios: one where the actual population size is unknown (and random identities are sampled from 1 to the number of observations for each level of the factorial fixed effect of interest) and one where the actual population size is perfectly known (random identities sampled from 1 to the actual number of subjects for each level of the factorial fixed effect of interest). We investigated the performance of each approach in terms of the type I error rate (erroneous significance) and the estimated variation between the levels of the random effect. We expected that only the correctly informed LMM would reveal the nominal type I error rate, since only this should be able to reliably determine the contribution of individual differences to the observations made.

METHODS

We simulated a study investigating the effect of a between-subjects factor (which we defined to be the sex of the subject) on a response variable. We explored three different values for the actual number of subjects in the studied population ($N_{subj} = 15, 30, 100$). For each individual simulation, we sampled N_f females and N_m males (such that $N_f + N_m = N_{subj}$) with equal probabilities for an individual being female or male. All simulations involved $N_{obs} = 100$ observations, with every subject having the same probability of being observed. This sampling strategy implies that not all subjects will necessarily be observed at least once within a given simulation (especially so when $N_{subj} = 100$).

No effect of sex was simulated. The expected average value of the response variable for each subject was sampled from a centred normal distribution (i.e. one with a mean of zero) whose standard deviation was SD_{subj}. The final value of the response variable was obtained by summing to these expected averages residual errors that were randomly sampled from a centred normal distribution with a standard deviation $SD_{res} = 1$. We explored four different values for the magnitude of the variation across subjects, ranging from almost negligible to very strong: $SD_{subj} = 0.25, 0.5, 1, 2$. In terms of repeatability, these values correspond to adjusted intraclass correlations (ICC) of 0.06, 0.2, 0.5, 0.8, respectively. ICC was calculated according to equation 2.7 in Nakagawa, Johnson, and Schielzeth (2017). We simulated 500 data sets for each combination of N_{subj} and SD_{subj} ($3 \times 4 \times 500 = 6000$ data sets).

For each data set, the statistical significance of the effect of sex was determined according to six approaches. First, we fitted a correctly informed LMM comprising a term for the random intercept of subject whereby the identity of the subject was correctly assigned for each observation (response ~ sex + (1 | subject identity)). This is the ‘correctly informed’ model that should reveal the expected type I error rate of 0.05 for the effect of ‘sex’ and serves as a benchmark to compare the other models with. Second, we fitted a naïve linear model in which the identity of subjects was ignored altogether (response ~ sex). This model reveals the type I error rate when the analysis is simply pseudoreplicated (note that this model is fully equivalent to an independent-samples *t* test). Then we followed the two approaches proposed by Garamszegi (2016), assigning random identities to each observation (with the condition that the same identity could only be assigned to individuals of the same sex) and fitting (1) a linear model on the per-subject averages (‘random means’) and (2) an LMM with a random intercept term for subject (‘random LMM’). Both approaches were based on individual identities randomly assigned to the data. Each of these two approaches was carried out in two variants. In the first variant, the actual population size was unknown and the random identities for each observation were sampled from 1 to N_{f_obs} and 1 to N_{m_obs}, respectively, for observations of females and males (with N_{f_obs} and N_{m_obs} being the number of observations of females and males, rather than the actual number of female and male subjects). In the second variant the actual population size, as well as the actual number of female and male subjects, was known (which corresponds to the very favourable case of exactly guessing the actual population size), and the random identities for each observation were sampled from 1 to N_f and 1 to N_m, respectively, for observations of females and males. For both the ‘random averages’ and ‘random LMM’ approach, we generated 100 random assignments of individual identities for each simulated data set.

All simulations and analyses were carried out in the R statistical environment (R version 3.4.4; R Core Team, 2018). Linear models were fitted using the lm function and LMMs using the lmer function of the package lme4 (version 1.1–17; Bates et al., 2015). Statistical significance (*P* value) was assessed using the standard summary function for linear models and the function drop1 for mixed models (which conducts a likelihood ratio test comparing the full model with one lacking sex; Barr et al., 2013). For the ‘random means’ and ‘random LMM’ approach, the *P* value for each simulated data set was obtained by averaging across the 100 random assignments. The significance threshold was set at $\alpha = 0.05$. We evaluated the models' type I error rate, that is, we determined the probability of rejecting the (true) null hypothesis of no effect of sex on the response variable as the proportion of tests revealing $P \leq \alpha$. Note that we based our assessment on averaged *P* values although this approach was not explicitly suggested by Garamszegi (2016), who rather suggested an evaluation of the estimated coefficients as a means to gauge their ‘uncertainty due to the unknown identity of subjects’. Since this is not equivalent to a confidence interval of the effect, and most researchers would probably want a means of inference, we chose to base our investigation on *P* values. The R code for the simulations is available as Supplementary material.

RESULTS

As expected, the ‘correctly informed’ mixed model was able to account for the effect of intersubject variation, as its type I error rate did not show any obvious correlation with the magnitude of the variation between subjects (Fig. 1). Type I errors obtained by this approach were never much higher than the expected α of 0.05.

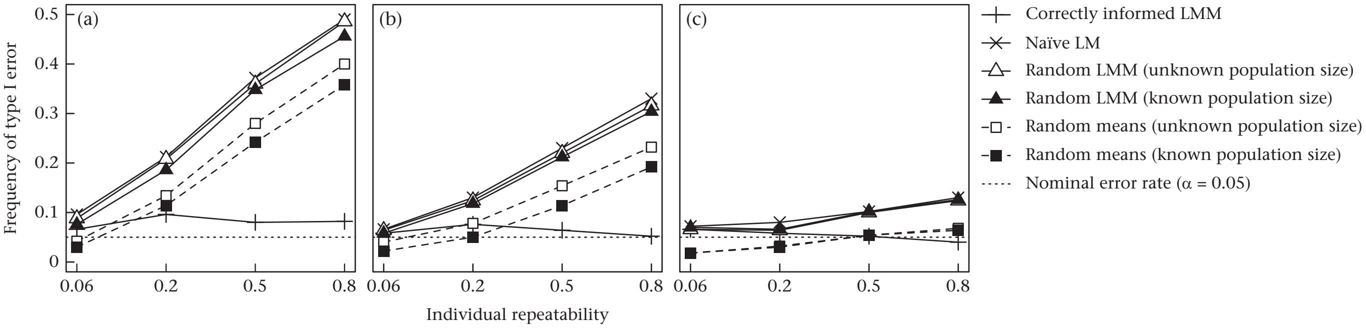

Figure 1. Frequency of type I errors over 500 simulations as a function of individual repeatability (proportion of total variance due to between-subjects variation) for six different statistical approaches (see main text for details). All simulated data sets comprised 100 observations. For 'random means' and 'random LMM' approaches, results from 100 random assignments of subject identities were averaged for each simulation. (a) Actual population size = 15 (mean number of observed individuals = 14.98); (b) actual population size = 30 (mean number of observed individuals = 29.01); (c) actual population size = 100 (mean number of observed individuals = 63.29).

However, our simulations revealed a slight but consistent excess of type I errors when the number of subjects was small (Fig. 1). The 'naïve' linear model (in which the identity of subjects was simply ignored) showed an increasing excess of type I errors when the magnitude of the variation between subjects increased. The excess of type I errors was low or negligible when the variation across subjects was small and/or the actual number of subjects was high in relation to the number of observations (so that the probability of observing the same subject multiple times was low, Fig. 1; see also Mundry & Oelze, 2016). However, a very large excess of type I errors was observed with large intersubject variation and a relatively low number of subjects, with the probability of rejecting the true null hypothesis becoming as high as ca. 50% (Fig. 1).

Type I errors obtained from the 'random LMM' approach proposed by Garamszegi (2016) were essentially identical to those of the 'naïve' linear model, being just marginally lower when the actual population sizes for each sex were both known and low (Fig. 1). Compared to the 'random LMM' approach, the frequency of type I errors was slightly lower in the 'random means' approach. In fact, this difference in frequency of type I errors did not vary with the magnitude of intersubject variation or actual number of subjects (Fig. 1). Consistently, the frequency of type I errors obtained by

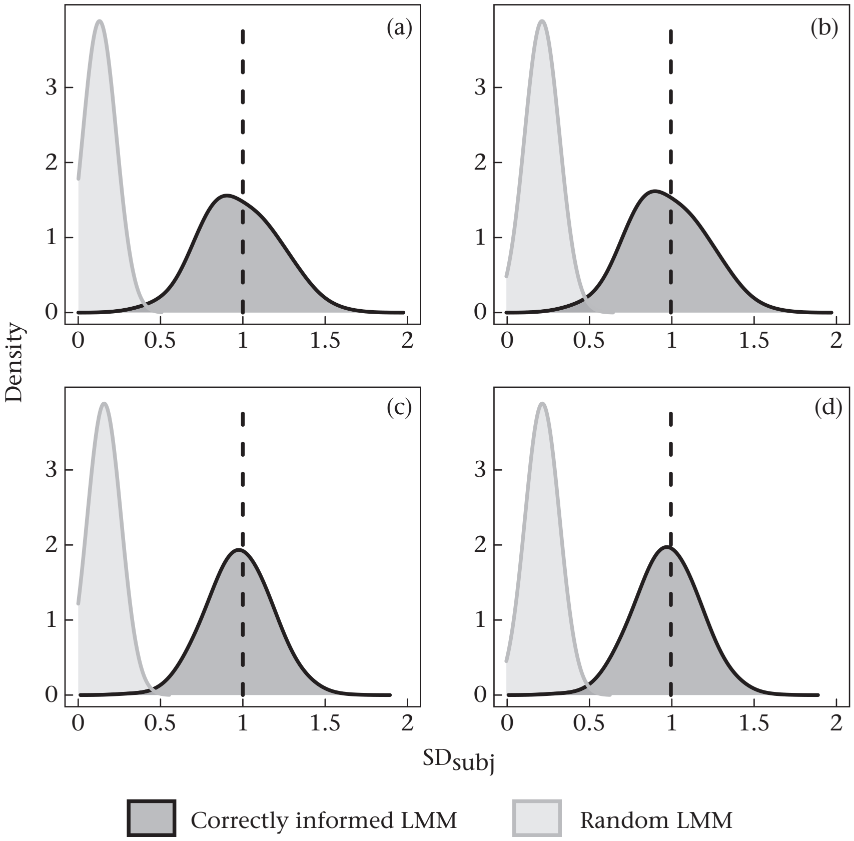

Figure 2. Estimated standard deviations for the random effect of subject (SD_{subj}). The Gaussian kernel densities of estimated standard deviations across 500 simulations are shown for the correctly informed (black line and dark shading) and the random LMM (with known and unknown population size; grey line and light polygon) for a fairly large ratio between intersubject variation and residual error (SD_{subj}/SD_{res} = 1, repeatability = 0.5; the vertical dashed lines show the simulated effect). (a) Actual population size is 15 and known; (b) actual population size is 15 and unknown; (c) actual population size is 30 and known; (d) actual population size is 30 and unknown.

the ‘random means’ approach was lower than α when the simulated data were close to a set of independent observations (i.e. low intersubject variation, high actual number of subjects; Fig. 1). In addition, the ‘known population size’ variant of the ‘random means’ approach resulted in fewer significant tests than the ‘unknown population size’ variant when the actual population size was lower than the number of observations ($N_{\text{subj}} = 15$ and $N_{\text{subj}} = 30$; Fig. 1). In fact, in these cases, the number of data points (randomized subjects) for the ‘known population size’ variant was lower than for the ‘unknown population size variant’ (when the number of observations was equal to the number of subjects the two variants were almost identical).

Regarding the estimated standard deviation of the random effect, we found this to be biased in the random LMM approach. In fact, individual variation was usually greatly underestimated, regardless of population size and whether it was known or not (Fig. 2). For the correctly informed LMM this was not the case.

DISCUSSION

We found that the naïve linear model produced type I error rates that correlated positively with the level of pseudoreplication and the magnitude of variation between individuals, and for larger levels of pseudoreplication and/or variation between individuals these were highly inflated. This result confirms the detrimental effects of pseudoreplication and is in line with a large body of literature reporting similar findings (e.g. Hurlbert 1984; Machlis, Dodd & Fentres, 1985). We also found that the use of the correctly informed LMM led to the nominal type I error rate of 0.05 when the number of subjects was not too small. However, for a small number of individuals, the correctly informed LMM also produced slightly elevated type I error rates of up to almost 0.1. This is probably because the chi-square approximation of the likelihood ratio test statistic (Barr et al., 2013) we used is known to perform well only for large samples (e.g. Bolker, 2008, pp. 13, 194). This interpretation is supported by the finding that for larger samples the type I error rate approached the nominal level.

We found that the LMM with randomly assigned individual identities had a highly inflated type I error rate that could be as large as 0.5 and virtually identical to that of the naïve linear model when the population size was unknown and only slightly reduced when it was known. Hence, using randomly assigned individual identities does not account for pseudoreplication at all. This was also reflected in the results regarding the estimated contribution of the random effect which was clearly underestimated for the LMM with randomly assigned individual identities. Using averages per randomly assigned individual also led to a highly elevated type I error rate; the fact that it was slightly below that of the naïve LMM obviously arose from a reduced sample size due to the averaging. Only when the population size was large compared to the number of samples, leading to a low number of individuals sampled repeatedly (and probably low average numbers of replicate samples per individual), did type I error rates of the approaches proposed by Garamszegi (2016) come closer to the nominal type I error rate of 0.05, but they were still clearly above it and virtually identical to the naïve approach. The only case in which the approach proposed by Garamszegi (2016) led to type I error rates close to the nominal 0.05 was when the amount of variation between individuals was low. This result is not surprising since, in such a case, interindividual variation does not contribute much to the response (i.e. observations are close to being independent). On the other hand, it seems obvious that an LMM conducted on data in which there are differences between individuals but with individual identities randomly assigned to observations will underestimate the contribution of variation between individuals. Hence, neither of the two approaches based on random assignment of individuals to samples can be recommended, since both lead to an inflated type I error rate. Instead, one must rely on a correctly informed LMM, which is obviously only possible when individual identities are known. When these are not known, the approach proposed by Garamszegi (2016) does not alleviate the problem of pseudoreplication.

The simulation we used here assumed the predictor of interest to vary between individuals but not within them, and hence the question arises what one would expect for a predictor of interest varying within subjects. First, our results showed that only the correctly informed LMM could reliably disentangle variation due to differences between individuals and the influence of a fixed effect and there is no reason why this should be different when the fixed effect varies within rather than between individuals. However, Mundry and Oelze (2016) recently reported that in such a case pseudoreplication can lead to an inflation of both type I and type II error rates. Hence it seems likely that the same is the case when using the approaches proposed by Garamszegi (2016) in such a situation.

To summarize, for data that are likely to comprise repeated observations of the same individuals but with the identity of the individuals from which data were collected unknown, there is currently no method available that allows one to avoid the consequences of pseudoreplication (i.e. a potentially drastically inflated type I error rate). In fact, the methods proposed by Garamszegi (2016) produced type I error rates that were largely identical to those of a model that was simply pseudoreplicated (i.e. the 'naïve' model). Actually, the slightly reduced type I error rates observed with the ‘random means’ approach (still clearly above the nominal type I error rate of 0.05) depend on the reduced sample size and must be expected to come at the cost of increased type II error. Hence, such data currently do not allow a robust analysis. However, even though it seems unlikely to us, an approach building upon Garamszegi's (2016) proposition to base inference on the variation in the estimated coefficients might reveal a different pattern. Hence, future research is warranted.

Acknowledgments

This work was supported by the Max Planck Society. P.G. is supported by the Max Planck Society through the Pan African Programme (panafrican.eva.mpg.de).

Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.anbehav.2019.05.014.

References

Baayen, R. H. (2008). *Analyzing Linguistic Data*. Cambridge, U.K.: Cambridge University Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.

Bates, B., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Bolker, B. M. (2008). *Ecological models and data in R*. Princeton, NJ: Princeton University Press.

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., et al. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution, 24*(3), 127–135.

Garamszegi, L. Z. (2016). A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons. *Animal Behaviour, 120*, 223–234.

Hadinger, U., Haymerle, A., Knauer, F., Schwarzenberger, F., & Walzer, C. (2015). Faecal cortisol metabolites to assess stress in wildlife: Evaluation of a field method in free-ranging chamois. *Methods in Ecology and Evolution, 6*(11), 1349–1357.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs, 54*(2), 187–211.

Machlis, L., Dodd, P. W. D., & Fentress, J. C. (1985). The pooling fallacy: Problems arising when individuals contribute more than one observation to the data set. *Zeitschrift für Tierpsychologie, 68*(3), 201–214.

Mundry, R., & Oelze, V. M. (2016). Who is who matters – the effects of pseudoreplication in stable isotope analysis. *American Journal of Primatology, 78*, 1017–1030.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface, 14*(134), 20170213.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology, 20*(2), 416–420.