Fixed or random? On the reliability of mixed-effect models for a small number of levels in grouping variables

Authors and Affiliations:

Johannes Oberpriller[1*], Melina de Souza Leite[1, 2], Maximilian Pichler[1*]

[1] Theoretical Ecology, University of Regensburg, Universitätsstraße 31, 93053 Regensburg, Germany

[2] Department of Ecology, University of São Paulo, Rua do Matão 321, Travessa 14, São Paulo, SP CEP 05508-090, Brazil

* These authors contributed equally

Keywords:

Mixed-effect models, generalized linear models, multilevel models, hierarchical models, fixed effects, random effects

Abstract

1. Biological data are often intrinsically hierarchical. Due to their ability to account for such dependencies, mixed-effect models have become a common analysis technique in ecology and evolution. While many questions around their theoretical foundations and practical applications are solved, one fundamental question is still highly debated: When having a low number of levels should we model a grouping variable as a random or fixed effect? In such situation, the variance of the random effect is presumably underestimated, but whether this affects the statistical properties of the fixed effects is unclear.
2. Here, we analyze the consequences of including a grouping variable as fixed or random effect and possible other modeling options (over and underspecified models) for data with small number of levels in the grouping variable (2 - 8). For all models, we calculated type I error rates, power and coverage. Moreover, we show the influence of possible study designs on these statistical properties.

3. We found that mixed-effect models already for two groups correctly estimate variance for two groups. Moreover, model choice does not influence the statistical properties when there is no random slope in the data-generating process. However, if an ecological effect differs among groups, using a random slope and intercept model, and switching to a fixed-effect model only in case of a singular fit avoids overconfidence in the results. Additionally, power and type I error are strongly influenced by the number of and the difference between groups.

4. We conclude that inferring the correct random effect structure is of high importance to get correct statistical properties. When in doubt, we recommend starting with the

simpler model and using model diagnostics to identify missing components. When having identified the correct structure, we encourage to start with a mixed-effects model independent of the number of groups and only in case of a singular fit switch to a fixed-effect model. With these recommendations, we allow for more informative choices about study design and data analysis and thus make ecological inference with mixed-effects models more robust for low number of groups.

Introduction

Many biological data have a hierarchical grouping structure that introduces dependencies among observations (McMahon & Diez 2007; Bolker *et al.* 2009; Zuur *et al.* 2009; Harrison *et al.* 2018). When conducting statistical analyses, models have to reflect these dependencies in their assumptions for correct statistical properties (Arnqvist 2020), a task for which linear and generalized mixed-effects models (LMMs or GLMMs) were designed (Laird & Ware 1982; Chen & Dunson 2003). Due to their ability to simultaneously analyze variance on different hierarchical levels (Krueger & Tian 2004; Boisgontier & Cheval 2016) and their better statistical properties when data is missing (Baayen *et al.* 2008), mixed-effects models have also replaced ANOVAs as the common tool for variance analysis (Wainwright *et al.* 2007; Boisgontier & Cheval 2016). Thus, mixed-effects models have become one of the most popular methods in ecology and evolution (Bolker *et al.* 2009) and other fields (e.g. psychology, see Meteyard & Davies 2020).

The ability of mixed-effects models to adapt to different data structures (i.e. their flexibility to handle many different hierarchical structures, see Box 1), which makes them in the first place so powerful (Wainwright *et al.* 2007), raises also discussions about their proper application (Nakagawa & Schielzeth 2013). These issues include data-related problems such as overdispersion (Harrison 2014, 2015), robustness to wrong distributional assumptions (Schielzeth *et al.* 2020), and their correct evaluation (Nakagawa & Schielzeth 2013). Despite these rather technical challenges, there are also application-oriented issues (Harrison *et al.* 2017; Meteyard & Davies 2020), such as the question whether to use the most complex random effect structure (Barr *et al.* 2013; but see Matuschek *et al.* 2017) or in which situations we should model a grouping variable as random or fixed effect (Harrison *et al.* 2018).

A priori, modeling a grouping variable as fixed or random effect are equally well suited for multilevel analysis (Townsend *et al.* 2013) and strict rules don't exist because the best strategy generally depends on the goal of the analysis (Gelman & Hill 2007, see Box 2). There are, however, technical differences between these two options (Millar & Anderson 2004). The most prominent difference is that modeling a grouping variable as a random effect implicitly makes the assumption that the individual levels of the grouping variable are realizations of a common distribution, usually a normal distribution, for which the standard deviation is unknown and need to be estimated (e.g. DerSimonian & Laird 1986). This assumption shrinks the estimates of each random effect level to the mean of the underlying distribution. In contrast, modeling a grouping variable as a fixed effect makes no distributional assumptions about the individual level estimates. Thus, compared to modeling the grouping variable as a fixed effect, the random effect model uses fewer degrees of

freedom (e.g. Schelldorfer *et al.* 2011), which leads to higher power to detect significant effects on the remaining fixed effects (p-values are inversely proportional to degrees of freedom) at the cost of a small bias towards zero on the random effect estimates (e.g. Johnson et al. 2014).

Based on these properties, random effect modeling seems preferable over fixed-effect modeling if the estimates of the grouping variable are not of particular interest, or if the modeler is willing to accept the additional assumption of the normal distribution on those estimates. As the variance of the random effect has to be estimated, however, it is not clear if these advantages remain when the number of groups is small (cf. also Harrison *et al.* 2018). Additionally, a small number of levels might cause a wrong estimate of the random effects' standard deviation (Harrison, 2018), which might influence other estimates of the mixed-effects model (e.g. the fixed effects, Hox *et al.* 2017).

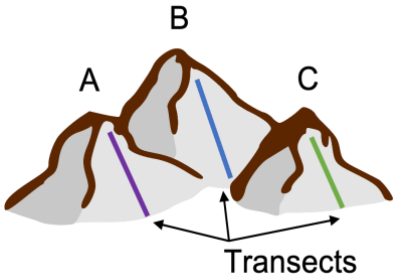
The ecological literature suggests, as a rule of thumb, to only fit a grouping variable as random effect if it has a minimum of five, sometimes eight, levels to properly estimate the standard deviation of the random effect (Bolker 2015; Harrison 2015; Harrison *et al.* 2018). With four or fewer levels in the grouping variable, fitting it as a fixed effect is recommended as the preferred alternative (Gelman & Hill 2007; Bolker *et al.* 2009; Bolker 2015). In other disciplines other thresholds are proposed, e.g. 10-20 in psychology (McNeish & Stapleton, 2016) or 30-50 in sociology (Maas & Hox, 2005). To our knowledge, however, none of these values were based on a systematic analysis of how the choice of the grouping variable affects the statistical estimates (i.e. type I error rate, power, and coverage) of other effects of interest in the model (i.e. the slope effect of some explanatory variable, henceforth called the ecological effect).

Here, we simulated a hypothetical study on reproductive success and height of a plant on a temperature gradient (altitudinal gradient among different mountains (levels), see Box 1) with two in complexity increasing data-generating processes to compare statistical properties with a varying number of levels (two to eight mountains) and varying data complexity. To represent the challenge of an analyst to correctly specify the model structure and the consequences if not, we additionally tested miss-specified models (too complex or too simple versions of the fixed and mixed-effects models). To investigate the consequences of these modeling choices on the ecological effect with different number of levels in the grouping variable, we compared: type I error rates (how often a non-existing temperature effect would be interpreted as existing), statistical power (the rate at which a truly existing temperature effect would be interpreted as existing), and statistical coverage (i.e. how often the true effect of temperature falls into the 95% confidence interval of the model). Based on our results, we give practical recommendations on when to include grouping variables as random effects and in which situations it might be more beneficial to include them as fixed effects.

Box 1: Example of an ecological study design with hierarchical/grouping variables

Sampling design. We want to understand the effect of temperature on the reproductive success of a plant that grows in mountains. We hypothesize that: (H1) increasing temperature (lower altitude) increases the probability of flowering (reproduction), and also (H2) the height of flowering plants. To do so, we establish altitudinal transects in many mountains (populations) and collect information from a certain number of plants.

Sampling design

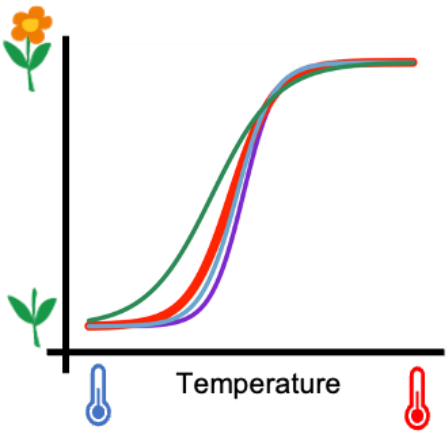


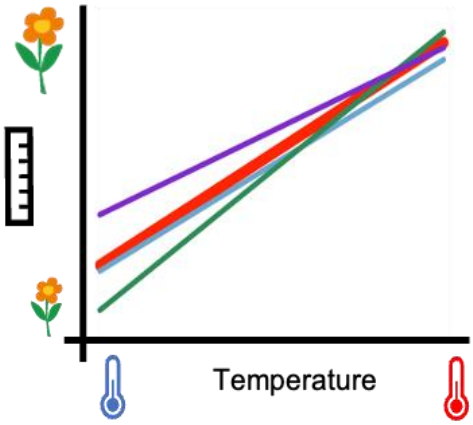
Problem. The transects are not in the same geographical alignment, the type of soil varies in each mountain, and the plants are genetically very distinct among populations. All these factors introduce differences among populations that are not exactly of our interest (given our hypotheses), but statistically, plants of the same mountain are non-independent observations. The mountains can be considered as grouping, blocking or control factors.

Modeling options. We may use a mixed-effects model with a random intercept and slope (Box 2) for mountain to account for the differences among populations (colored lines in H1 and H2), while still modeling the relationship of interest as fixed effects (red lines). An alternative may be to use a simple fixed-effects model, i.e., to include mountain as a categorical predictor (Box 2).

Hypothesis 1 - The reproductive success (flowering or not) of plants increases with temperature.

Hypothesis 2 - The height of flowering plants increases with temperature.





Box 2: Differences between modeling a grouping variable as random or fixed

Fixed or random effects The question of whether to include a hierarchical/grouping variable as a random or fixed effect in the analysis depends on several factors. Fixed effects are usually used when the analysts are interested in the individual level estimates of a grouping variable (Bolker *et al.* 2009) and these are independent, mutually exclusive, and completely observed (e.g. control and treatment in experiments, male and female when analyzing differences between sex) (e.g. Hedges & Vevea 1998; Gunasekara *et al.* 2014). Random effects, in comparison, are vital modeling choices when the variance between the different levels is of interest (Bolker *et al.* 2009) and not the exact estimates of the different levels (e.g. DerSimonian & Laird 1986). Additionally, random effects can be used when not every realization of the underlying mechanism can be observed (e.g. species across a number of observational sites in different geographic areas) but the analysts want to control for its influence (i.e. pseudo-replication, see Arnqvist 2020).

Technical differences between random and fixed effects When specifying a grouping variable as fixed effect, the model estimates the effect of one reference level and, for the default contrast method (see Schielzeth 2010), differences between the reference level and possible linear combinations of other levels (Fig. B1a,c). Fixed-effect models don't estimate mean effect over groups, but it can be calculated using uncertainty propagation (see Supporting Information S1.3). Mixed-effect models, nevertheless, estimate the grand mean effects and their standard deviation additionally to each individual level effects (Fig. B1b, d). Grouping variables may have not only different intercepts (Fig B1 a,b), but also different slopes (Fig B1 c, d - the temperature "ecological" effect). In fixed-effects models, this is done by introducing an interaction between the ecological effect and the grouping variable, and in mixed-effects models by estimating random slopes for each group. The choice of modeling different slopes for each group is related to the complexity of the data and may have impact on modeling structure and inference.

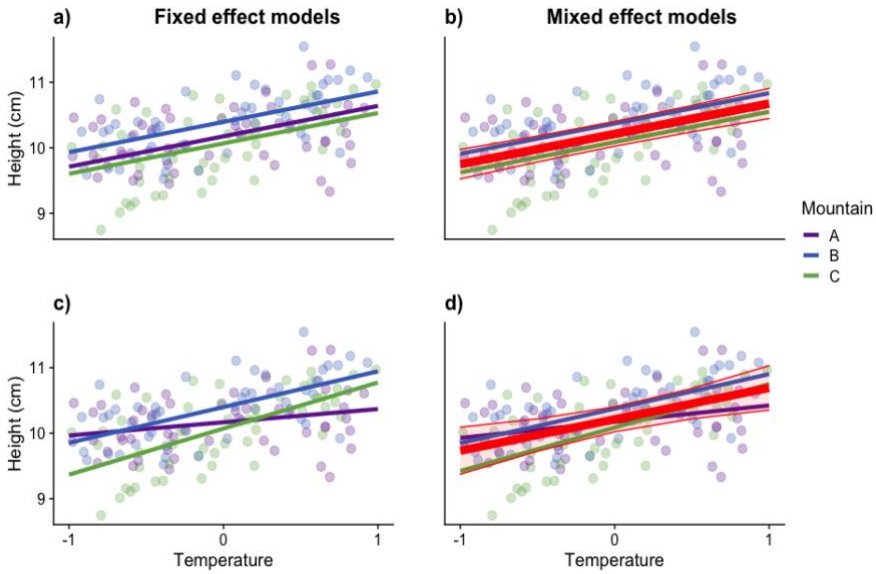


Figure B 1: Fixed- and mixed-effects models fit to simulated data with random intercept (a,b) and random intercept and slope (c,d) for each mountain in the example from H2 Box 1. Dots represent the data points, lines the individual estimates for each mountain. The big read line is the mean and the shaded red interval is its estimated standard deviation for mixed-effect models.

Models in R. Here are the formula R syntax[1] of fixed and mixed-effect models fitted in Figure B1

Scenario A random intercept only

a) Height ~ Temperature + Mountain (fixed-effect model)
b) Height ~ Temperature + (1| Mountain) (mixed-effect model)

Scenario B random intercept and slope

c) Height ~ 0 + Mountain + Temperature:Mountain (fixed-effect model)
d) Height ~ Temperature + (1|Mountain) + (0 + Temperature| Mountain) (mixed-effects model[2])

1 From lm() or glm() base R functions, lme4 (Bates et al. 2015) and glmmTMB (Brooks et al. 2017) packages.
2 Uncorrelated random intercept and slope. For correlated models, the formula for the random terms is (1+Temperature|Mountain) or (Temperature|Mountain).

Methods

Simulated example

To compare random and fixed-effect modeling of a hierarchical variable with small number of groups, we simulated data based on our hypothetical example from Box 1. We hypothesized first that a higher temperature improves the reproductive success (either yes or no) of a plant species (H1), and second, that a higher temperature also increases the average height of the reproductive plants (H2). To test these hypotheses, we simulated a data collection in a randomly chosen number of mountains from a certain number of plants (200 plants per mountain for H1 and 50 plants per mountain for H2) along altitudinal transects. We varied the number of mountains (groups) from two to eight and simulated 5,000 datasets for each case. We investigated performance for models with binomial (GLMM and GLM, H1) and normal distribution (LMM and LM, H2). The different choices of observations per group was made to meet the higher data request of binomial models compared to linear models.

Scenarios of data complexity and model fitting

Table 1: Data-generating and tested models for each scenario of data simulation: Scenario A random intercept for each mountain and B random intercept and slope for each mountain. For the fixed-effects models, we used R syntax for model formula in *lm()* function and for the mixed-effects models we used syntax from *lmer()* or *glmmTMB()* functions from *lme4* and *glmmTMB* packages. Response variables are reproductive success (RS) for the binomial distribution models (H1, Box 1) and height of flowering plants for the normal distribution models (H2, Box 1). T is the temperature effect on both response variables.

		Scenario A Random intercept only		**Scenario B** Random intercept and slope	**Description**
Data-generating model	(1)	Height/RS ~ T + (1|mountain)	(6)	Height/RS~ T + (1|mountain) + (0+T|mountain)	Effect of intercept (and slope in B) vary across mountains

Tested models					
Fixed-effect models	(2)	Height/RS ~ T	(7)	Height/RS ~ T	Temperature only main effect – too simple model
	(3)	Height/RS ~ T + mountain	(8)	Height/RS ~ 0 + mountain + T:mountain	Main effects of temperature and mountain (and interaction in B) – correctly specified models
Mixed-effect models			(9)	Height/RS ~ T + (1|mountain)	Temperature and mountain both vary – too simple models
	(4)	Height/RS ~ T + (1|mountain)	(10)	Height/RS ~ T + (1|mountain) + (0+T|mountain)	Effect of intercept (and uncorrelated slope temperature in B) vary across mountains – correctly specified models
	(5)	Height/RS ~ T + (1|mountain) + (0+T|mountain)	(11)	Height/RS ~ T + (T|mountain)	Effect of intercept and slope temperature (correlated effects in B) across mountains – too complex models

Scenario A - random intercepts per mountain

In scenario A, we assumed mountains differ in their intercepts (mean height/reproductive success, Table 1, Eq. 1). For this scenario we tested two different mixed-effects model structures: a correctly specified model which corresponds to the data generating model (Table 1, Eq. 4) and a too complex model (Table 1, Eq. 5) with an additional random slope for each mountain. Since in real data analysis the true underlying data generating process is unknown, it is useful to understand if a too complex model can correctly estimate a zero (or nearly zero) random slope and, thus, approximate the true model structure (Table1, Eq. 1).

As fixed effect alternatives, we tested the correctly specified model with mountain as fixed intercept together with temperature as slope (Table1, Eq. 3), and a too simple model omitting mountain at all (Table 1, Eq. 2). This last model corresponds to a mixed-effects model that estimates the random intercept of mountain as zero (Table1, Eq. 4).

Scenario B - random intercepts and random slopes per mountain

In scenario B, we assumed a random intercept and a random slope (without correlation among them) for each mountain (Table 1, Eq. 6) as data-generating process. We tested three different mixed-effects model structures: a correctly specified model corresponding to the data generating model (Table 1, Eq. 10), a too complex model with a correlation term for the random intercept and random slope (Table 1, Eq. 11), and a too simple model with only a random intercept for each mountain (Table 1, Eq. 9). We used the too simple model to test the effect of not accounting for important contributions to the data-generating process.

As fixed effect alternatives, we tested the correctly specified model with the main effects of temperature and mountain, and the interaction term (Table 1, Eq. 8), and the too simple model dropping mountain as predictor (Table 1, Eq. 7). We tested the last model, because mixed-effect models, with standard deviation estimates of zero for both random effects, correspond to fixed-effects models omitting the grouping variable.

Model fitting

We fitted mixed-effects models to our simulated data with two of the most commonly used packages in R: lme4 (Bates *et al.* 2015) and glmmTMB (Brooks *et al.* 2017). We present here lme4 results because it reports optimization issues as singular fit, i.e., when some of the parameters of the variance-covariance Cholesky decomposition used in the optimization are exactly zero. Results with glmmTMB can be found in Supporting Information S1.

For LMMs we used the restricted maximum likelihood estimator (REML). For GLMMs we used the maximum likelihood estimator (MLE) since REML is not supported in lme4 for

GLMMs (for a comparison of REML and MLE see Supporting Information S1). All results presented in scenario A and B are for the datasets without singular fit convergence problems.

Statistical properties and simulation setup

We compared the modeling options for both data generating scenarios based on important statistical properties of the temperature effect (ecological effect) in our simulated data: type I error rate, statistical power, and coverage. Type I error rate is the probability to identify a temperature effect as statistically significant although the effect is zero. Statistical power is the probability to detect the temperature effect as significant if the effect is truly greater than zero. Coverage is the probability that the true temperature effect falls into the estimated 95% confidence interval. For a correctly calibrated statistical test, the type I error is expected to be equal to the alpha-level (in our case 5%), the coverage to be equal to the confidence level (in our case 95%), and the higher the statistical power the better the test.

To investigate the behavior of type I error, we simulated data with no slope effect, i.e., the effect of temperature on plant reproduction (average reproductive success of 0.5) and on height (average 10 cm) is zero. To additionally investigate statistical power and coverage, we simulated an example with a weak effect, i.e., an effect that is barely statistically significant. In our example, a weak effect corresponds to an average increase in size per unit step of the standardized temperature (linear scale) of 0.4 cm (Hypothesis 2) and 0.4 gain in reproductive success (Hypothesis 1) at scale of the linear predictors.

For scenario A and scenario B, the individual effects for each mountain were drawn from a normal distribution with standard deviation of 0.1 around the average effects: 10.0 cm

average height (intercept), and 0.4 cm average increase in size or 0.4 (logit link scale) gain in reproductive success with temperature (slope).

Standard deviation of random effects and singular fits

To understand how the number of levels affected the estimation of the standard deviation of the random effect in mixed-effect models, we recorded and compared standard deviation estimates for random intercepts and slopes from the correctly specified mixed-effects model in scenario B (Table 1, Eq. 10). We also compared optimization routines (REML and MLE) in terms of estimating zero standard deviations (singular fits, see below) for both LMMs and GLMMs (see supporting information S1).

Technically, singular fits occur when at least one of the variances (diagonal elements) in the Cholesky decomposition of the variance-covariance matrix are exactly zero or correlations between different random effects are estimated as -1 or 1. As singular fits represent models that have problems in the optimization process, we have classified our simulated datasets into the ones that showed singular fit warnings or not for mixed-effect models in lme4 package for the correctly specified model in scenario B (Table 1, Eq. 10). We calculated the above-mentioned statistical properties separately for the two groups of data: singular and non-singular fit.

Quantifying the influences of study design on power and type I error

The statistical properties of the ecological effect may depend not only on the number of groups (mountains) but also on the standard deviation of the random effect and the overall number of observations per group. To further quantify their impacts on the statistical properties of the ecological effect, we additionally ran 300 iterations (each with 1,000 non-singular model fits) with the data generating model from scenario B for hypothesis 2 (height of plants, uncorrelated random slope and random intercept). Thereby, we sampled the number of mountains from 2 to 20, the number of observations per mountain from 10 to 500 and the standard deviation from 0.1 to 2.

We fitted the correctly specified LMMs and LMs (Table 1, Eq 8), with the same structure as the data generating model (Table 1, Eq 10), and calculated type I error rate and statistical power of the temperature effect. Based on if the simulations had more than 10 groups, we split the results to better distinguish the effects of different study design parameters. We then fitted a linear regression with the statistical property (power and type I error rate) as response and standard deviation, number of groups and number of observations per group and their two-way interactions as predictors. All predictors were standardized to allow for a correct comparison of effect sizes across variables.

Results

Scenario A - random intercepts per mountain

When the data-generating model had a random intercept, irrespectively of the number of groups (mountains), all models except for the too complex model (random intercept and slope) showed a type I error rate of 5% and coverage around 95% (Fig. 1) as long as we

made sure that they converged (no singular fit). Power increased (Fig. 1A) with the number of mountains for LMMs from 85% (2 mountains) to 100% (5 to 8 mountains), and for GLMMs (Fig. 1B) from 35% (2 mountains) to 90% (8 mountains). Note that the model omitting the grouping variable mountain presented similar properties as the other models. However, when increasing the standard deviation of random intercept in the simulation, this model showed much lower power (Fig. S6).

For the too complex model for LMMs, we found on average a lower type I error rate of around 1-2% (Fig. 1A), lower statistical power to detect the temperature effect for a small number of mountains, and higher coverage (of around 98-99%) than the other models (Fig. 1A). Binomial datasets with small number of observations per mountain (25, 50, 100) presented similar results regarding type I error and coverage but as expected, very low power among models (Figure S8).

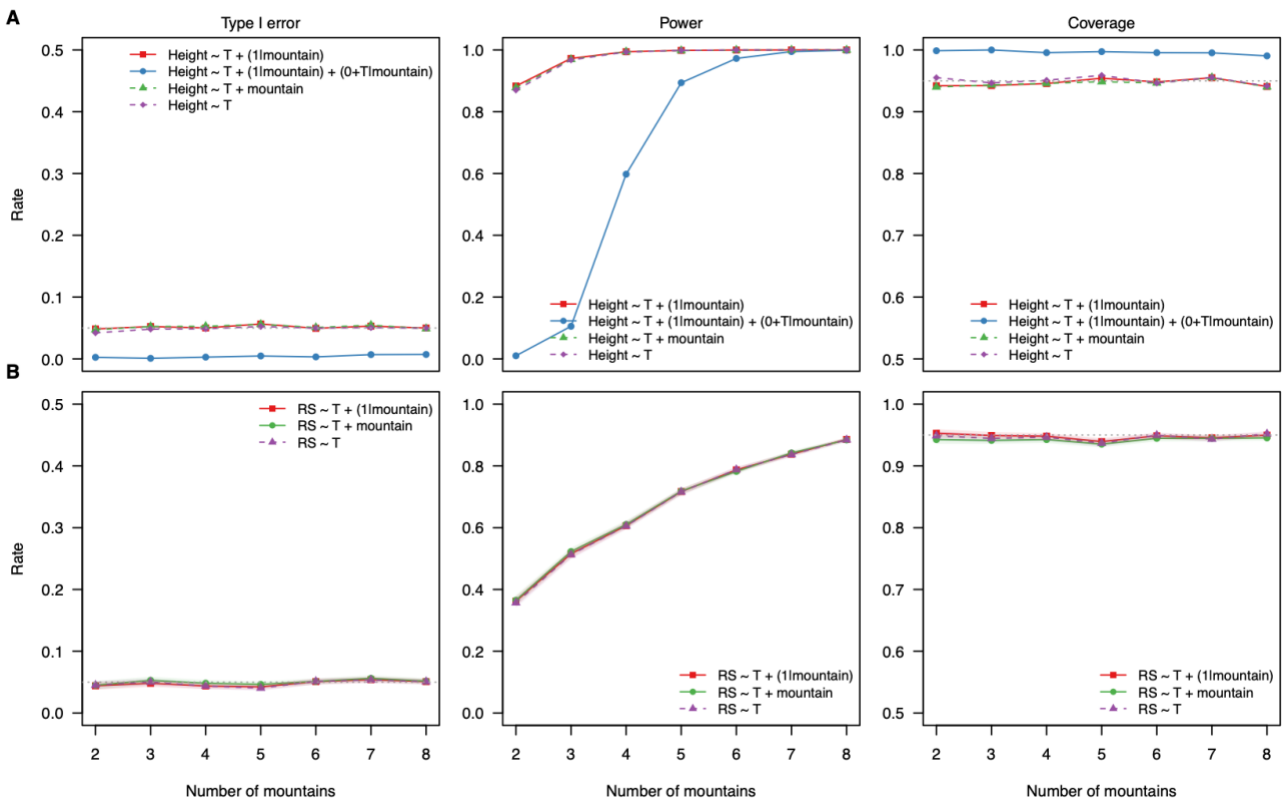


Figure 1: Type I error rates, power, and coverage for linear fixed and mixed effect models (A) and generalized linear fixed and mixed-effect models (B), fitted to simulated data with 2-8 mountains (random intercept for each mountain - Scenario A) and 50 observations per mountain. For each scenario, 5,000 simulations and models were tested. These results are only for datasets in which mixed-effect models converged without presenting singular fit problems.

Scenario B - random intercepts and slopes per mountain

In scenario B, where mountains differed in their average height/reproductive success (intercept) and their response to increasing temperature (slope of the ecological effect), the differences among models were greater (Fig. 2). We found that type I error rate of the correctly specified mixed-effects model (Table 1, Eq. 10) increased, and coverage decreased with the number of groups towards the respective nominal values (0.05 and 0.95, respectively). The too complex model with correlated random intercept and random slope (Table 1, Eq. 11) presented similar properties, but with slightly increased type I error and decreased power (Fig. 2). For the correctly specified fixed-effect model, type I error (≈ 2%) and coverage (≈ 98%) stayed constant with the number of groups (Fig. 2). For the correctly specified fixed- and mixed-effects models, power increased with the number of mountains, but the mixed-effects model showed overall higher power than the fixed-effects model irrespective of the number of mountains. For both normal and binomial models, the too simple model omitting the grouping variable resulted in a higher type I error rate (> 0.10), too low coverage (<0.90), but higher power than the other models (Fig. 2).

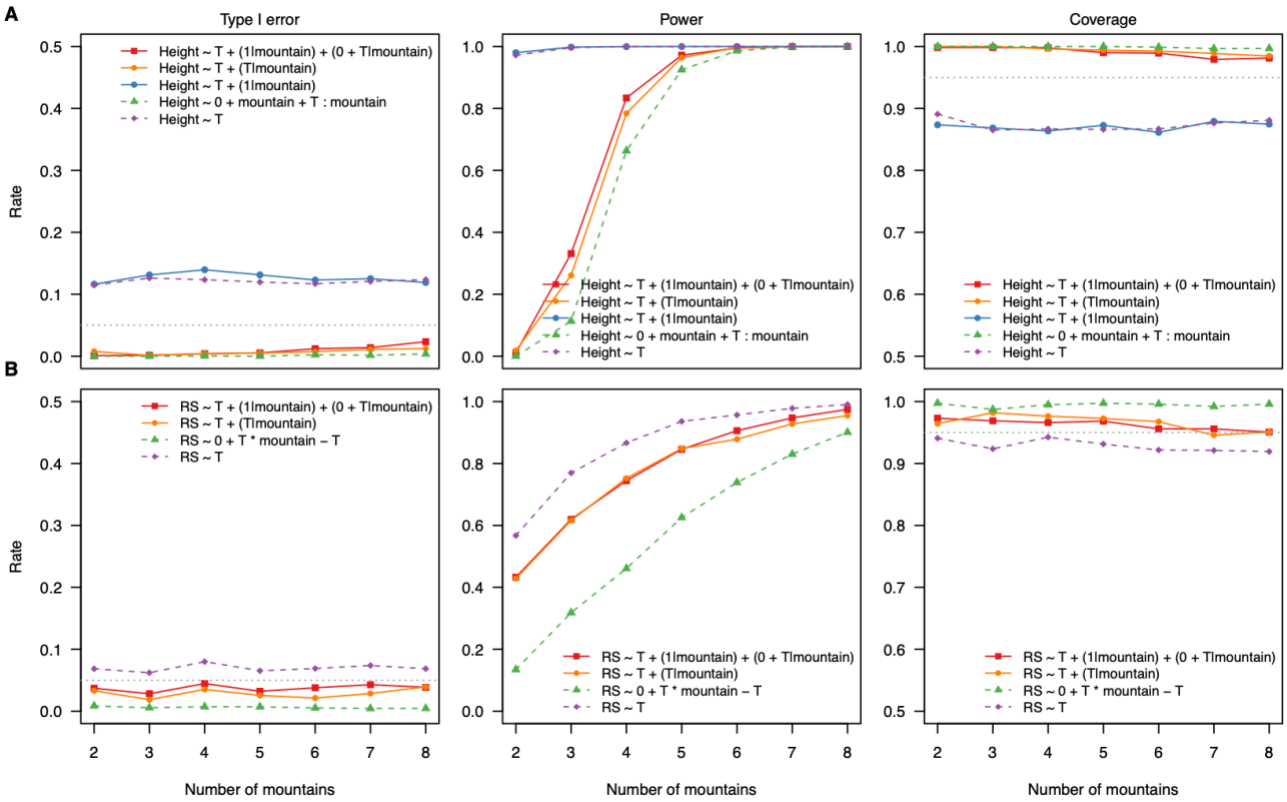


Figure 2: Type I error rates, power, and coverage for linear (mixed effect) models (A) and generalized linear (mixed-effects) models (B), fitted to simulated data with 2-8 mountains for scenario B (random intercept and random slope for each mountain range). For each scenario, 5,000 simulations and models were tested. These results are only for datasets in which mixed-effect models converged without presenting singular fit problems.

Standard deviation of random effects and singular fits

We found for LMMs (singular and non-singular fits combined, see Methods) in Scenario B (random intercept and slope) that the standard deviations of the random effects of the correctly specified model (Table 1, Eq. 10) showed bimodal distributions with one peak around the correct value (0.1) and one peak around zero (Fig. 3A, B). The peak around zero decreased in height, i.e., less models estimated a standard deviation of zero with an increasing number of mountains (Fig. 3A, B, see also Table S1). When looking at models without singular fits (Fig. 3 C and D), the distribution of standard deviations is unimodal

around the correct value. Results for glmmTMB were consistent with lme4 when using a threshold of 10^{-3} to classify a standard deviation as zero (singular fit) (Fig. S2).

By comparing the fitting algorithms, we found that using MLE led to more standard deviation estimates of zero (Fig. S3, S4) than REML. Additionally, using MLE, the distribution of standard deviation estimates of the non-singular fits are not at the correct value (Fig. S3, S4), but the bias decreases with increasing number of groups. For both optimization routines, increasing the number of levels reduced the number of singular fits (Table S1).

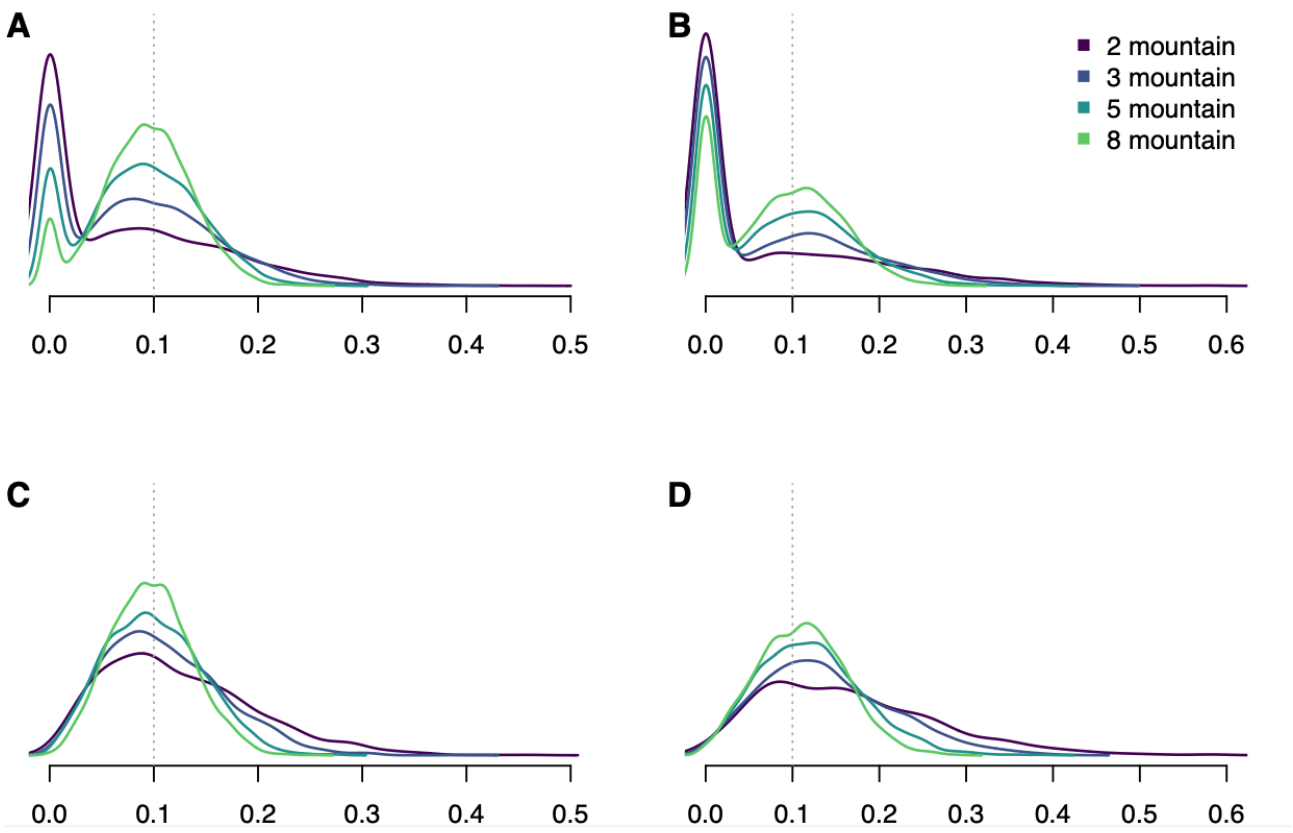


Figure 3: Standard deviation estimates of the random intercepts (A, C) and random slopes (B, D) for linear mixed-effect models (LMM, Table 1. eq. 10) in Scenario B, fitted with lme4 using REML to simulated data with 2-8 mountains. Figures A and B show the results for all models (with and without singular fits) and figures C

and D show the results for the models without singular fits. For each scenario, 5.000 simulations and models were tested. The grey lines represent the true standard deviation used in the simulation (0.1).

We found that singular fits had a high influence on type I error and power (Fig. 4) of mixed-effects models. For a singular fit, the type I error rate of the correctly specified mixed-effect model was constant around 10% (similar to the model omitting the grouping variable), while with non-singular fits it was 1% for two groups and increased towards the correct value of 5% with increasing number of groups (Fig. 4A). In comparison, the fixed effect model had similar type I error rates for data that presented singular and non-singular fits in mixed-effects models, both increasing towards the nominal value with increasing number of groups (Fig. 4C).

We also found differences in power for the models between singular and non-singular fits (Fig. 4B, D). For the correct LMM and correct LM, power increased with the number of groups, but the increase for the LM was stronger. The power of the mixed-effect model with correct structure is higher for singular than non-singular fits especially for a low number of mountains (Fig. 4B). For the fixed-effects model, the difference is less strong, however the model has still higher power with singular fit than for non-singular fits (Fig. 4D).

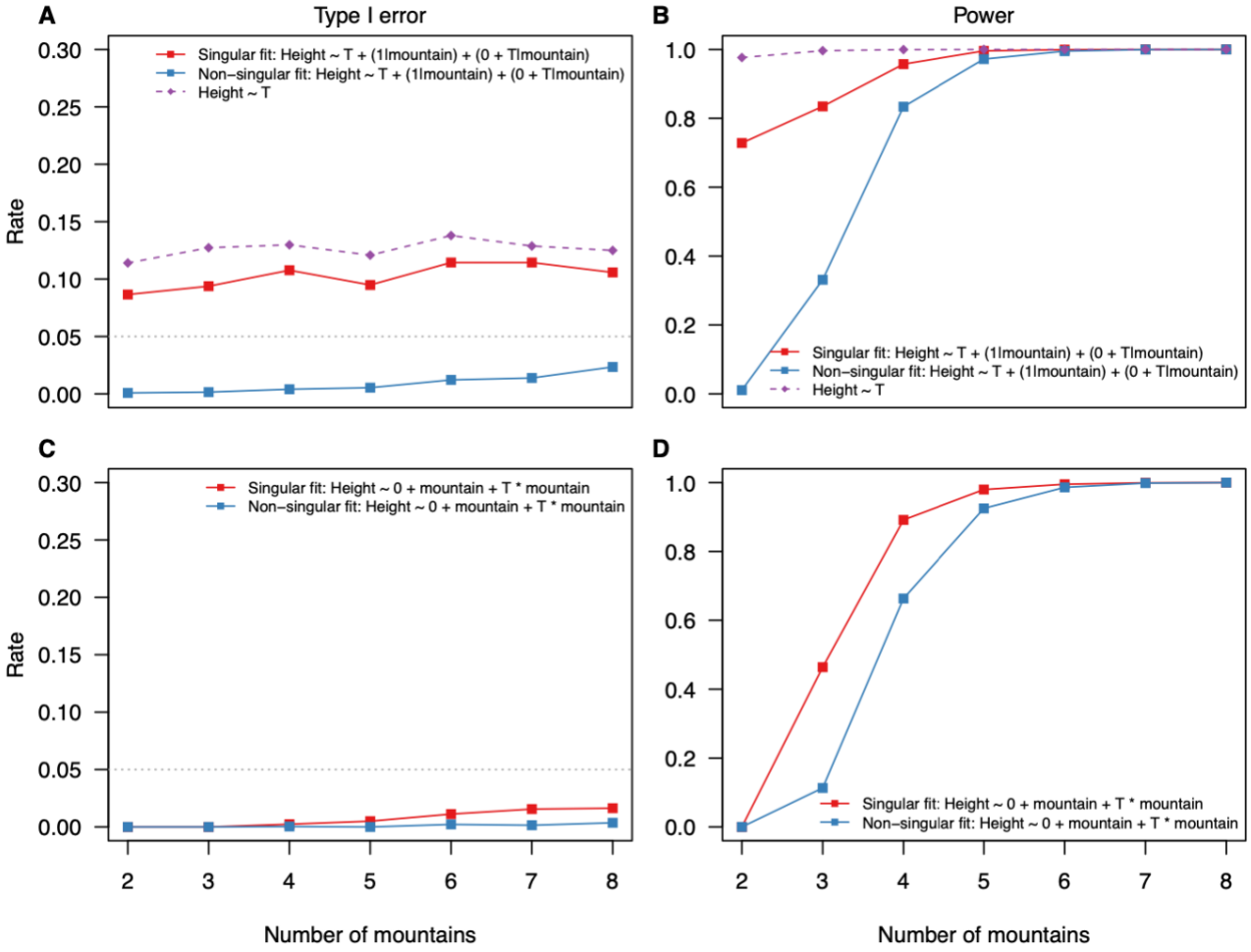


Figure 4: Type I error rate and power of the correctly specified linear fixed and mixed-effects models in scenario B separated by datasets that presented singular fit (red lines) and non-singular fit (blue lines) when fitted with lme4 package. Figure A and B are results for the linear mixed-effects models, and C and D for the linear fixed-effect models. For comparisons, we show also results for the fixed-effect model that omits the grouping variable (mountain).

Quantifying the influences of study design on power and type I error

With less than ten mountains (Fig. 5A, B), we found that the correctly specified LM started with a higher deviation to the expected type I error rate (0.05) for the ecological effect (Fig. 5A, intercept) than the correctly specified LMM. The LM was also more strongly affected by

the standard deviation of the random effects and the number of observations, i.e., increasing standard deviation/number of observations decreased the deviation to the nominal type I error rate. For both models, increasing the number of mountains equally led to a more correct type I error rate (Fig 5A). Additionally, for both models we found a tradeoff between the standard deviation of random effects and the number of mountains and observations per mountain (interaction terms in Fig 5A). When increasing random effects standard deviation, both the number of mountains and observations per mountain must also increase to obtain similar type I error rates. For the power, we found that the LM and LMM have similar values, and both are similarly affected by the study design factors (Fig. 5B). Increasing the standard deviation decreased power, while increasing the number of mountains increased power (Fig. 5B). Notably, we also found a strong negative interaction between the standard deviation and the number of mountains.

For more than 10 levels, the LM had a higher average deviation of the observed to the expected type I error rate but was stronger positively affected by the standard deviation (in the sense that the deviation was decreasing), number of mountains, and the number of observations (Fig. 5C). Power was equal for both models. Overall, the influence of the study design factors became less strong with more than 10 levels (Fig 5D).

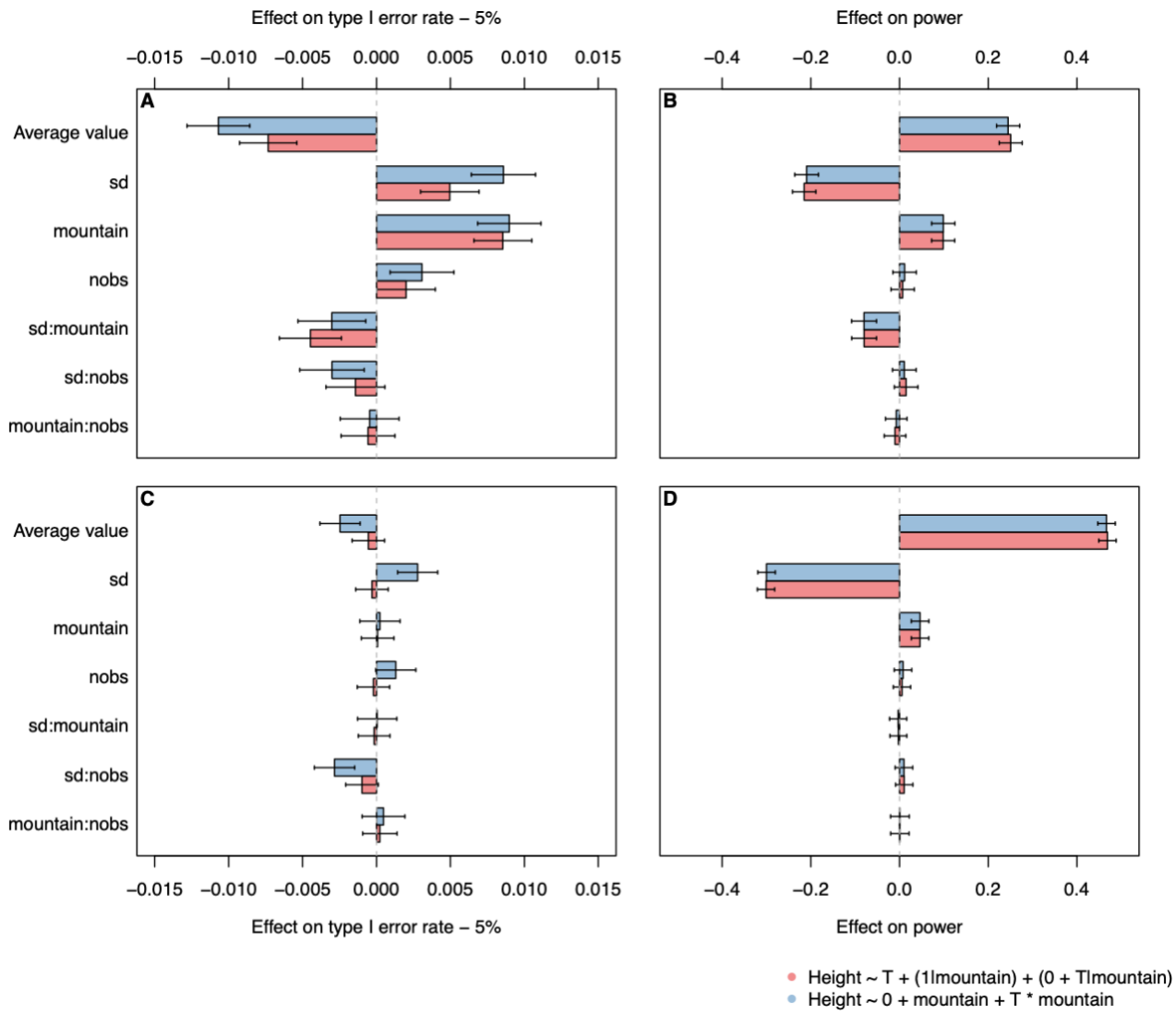


Figure 5: Comparing the influence of study design factors on the type I error rate (A, C) and power (B, D) for less than 10 levels (A, B) and more than 10 levels (C, D) of linear mixed- and fixed-effects models. We found that the standard deviation of the random effects and the number of mountains are the most important values to get correct statistical properties and improve the change to find a significant effect. For this analysis, we used the plant height example (Hypothesis 2, Box 1) for Scenario B (random intercept and random slope).

Discussion

Ecological data collections or experiments often produce data with hierarchical structures and mixed-effects models can account for these dependencies. When such data, however,

has a small number of levels in the grouping variable, analysts have to make a decision: Should they model the grouping variable as a fixed or a random effect? How does this decision influence the ecological effect of interest? Here, we showed with simulations that mixed-effect models with small number of levels in the grouping variable are more robust than previously assumed (Fig. 3) and that the decision between fixed and random effect modeling matters most when the data-generating processes included a varying slope for each group (Fig. 2).

When groups (mountains) differed only in their intercept (scenario A), almost all models independent of the number of groups presented the same statistical properties for the ecological effect (temperature) (Fig. 1), The only exception, the too complex model, had too low type I errors (close to zero, Fig. 1) and lower power. We speculate this is caused by the lower degrees of freedom and that the model was unable to correctly estimate the random slopes to zero.

Notably, for scenario A, the too simple model omitting the grouping variable presented correct statistical properties (Fig. 1). However, power decreased strongly with increasing differences between groups (Fig. S6), which confirms the importance of including grouping variables to correctly partition the variance among the different predictors (Gelman 2005; Gelman & Hill 2007; Bell *et al.* 2019).

When the intercept and slope varied per group (Scenario B) and the model fitted correctly, the model choice had a great influence on the statistical properties. The mixed-effects model had a better coverage, type I error, and a slightly higher power than the fixed-effect model, especially for a higher number of mountains (Fig. 2). A higher power may be linked to the

distributional assumption of the random effects (see Gelman & Hill 2007): with increasing numbers of mountains this assumption reduces the “effective” number of parameters needed to be estimated for the variations between mountains. In summary, mixed-effect models are less conservative but have higher power than fixed-effect models for low number of levels.

Too complex mixed-effect models for both scenarios presented slightly lower type I error and power compared to the correctly parameterized mixed-effects model (Fig. 1, 2). This behavior shows the trade-off between type I error and power reported by Matuschek *et al.* (2017) for different model complexities. Overall, the more complex (higher parametrized) models are more conservative but have less power than the simplified models.

In scenario B, too simple models in scenario B exhibited inflated type I errors (in line Schielzeth & Forstmeier 2009; Barr *et al.* 2013; Bell *et al.* 2019), too low coverage but very high power (Fig. 2). We speculate that the fixed-effect slopes explained additional variance coming from the different levels making the model overconfident.

Standard deviation of random effects and singular fits

The standard deviation estimates of the random effects of a correctly fitted model (no-singular fit) were around the correct value, showing that mixed-effect models were able to correctly estimate the random effect variance for a low number of levels (McNeish 2017), although with a higher uncertainty (Fig. 3B). However, in case of singular fit problems, the standard deviation estimates of the random effects were around zero (Fig. 3).

A singular fit in our simulations corresponded to a zero-variance estimate as we excluded the other possibility (we have set the correlations between random slope and intercept to zero, see Methods). In such situations, the correctly specified mixed-effect model had similar statistical properties as a fixed-effect model dropping the grouping variable (Fig. 4). This behavior becomes clearer when we think about the actual consequences of a variance estimate of zero: no difference between the levels and thus no influence of the grouping variable in the model (as a fixed-effect model without the grouping variable). However, the models still differed in their number of parameters (and degrees of freedom) which could explain the slight differences in statistical properties (Fig. 4) or the fact that only variance estimate of the random intercept and not the slope showed a singular fit.

The frequency of such singular fits decreased with the number of groups, suggesting that with more groups the estimates of the random effects' variance-covariance matrix became more stable (Fig. 3B, Table S1). When switching to fixed-effect models for singular fits in the random effect, the type I error rate and power were similar to the random effect model with non-singular fits (Fig. 4), suggesting that in case of a singular fit, the fixed-effect model with the correct structure is favorable over the mixed-effect model because of its more conservative estimates.

Connection to study design

Earlier studies reported mixed recommendations about important study design factors. While some studies stressed the importance of the number of observations for each level (Martin *et al.* 2011; Pol 2012), we found in accordance with Aarts *et al.(*2014) that the

number of levels and the standard deviation between the levels have higher influence on type I error rates and power.

The influence of the standard deviation on the statistical properties is mixed. On the one hand, increasing the standard deviation had a positive effect on the type I error for both models but the fixed-effect model was more strongly affected. We speculate the fundamentally different assumptions about the distribution of the levels could explain this different behavior: The mixed-effect model assumes the levels to be normally distributed and estimates flexibly the standard deviation of the levels, whereas the fixed-effect model makes no assumption about the distribution of the level which is equivalent to the assumption of an infinite standard deviation. On the other hand, increasing the standard deviation decreased the power of both models because more variance is explained by the different levels and thus the slope effect estimate is more uncertain.

To avoid the tradeoff between the influence of increasing standard deviation on type I error rate and power, we encourage to design a study with at least 10 levels because in this situation the effects of the study design on the statistical properties were small (Fig. 5). However, we are aware that financial or time resources are limited, so we believe that sampling similar grouping variables (i.e. mountains with a small standard deviation) is more beneficial as this leads to too conservative type I errors, but higher power. Moreover, the impact of study design on type I error and power stresses the importance of pre-experiments and power analyses (e.g. Johnson *et al.* 2015; Green & MacLeod 2016; Brysbaert & Stevens 2018) to maximize the meaningfulness of a study.

Practical suggestion

Coming back to our original question, we found no harsh threshold for the minimum number of levels in a grouping variable necessary to model it as a random effect (Fig. 2), but rather found that a singular fit in the random effect model indicates switching to the fixed-effect model is beneficial (more conservative type I error rates, Fig. 2).

We thus recommend, independent of the number of groups, to start with the mixed-effect model and only in case of a singular fit to switch to a fixed-effect model (more conservative type I error rate). For a random intercept and slope in the data-generating process, we recommend starting with correlated random slope and intercept (following Barr et al. 2013). If this model converges, the resulting estimates have approximately the correct nominal type I error. When obtaining a singular fit, one should first try uncorrelated random effects (following Matuschek *et al.* 2017) and only in case of a recurring singular fit switch to a fixed-effect model. These recommendations are summarized in Fig. 6.

The previous recommendations assumed that we know whether only the intercept, or both slope and intercept differ across levels. In practical data analysis, we, however, are never totally sure about the underlying data-generating process. In such situations, specifying a too complex model (including a random slope if there is none) leads to too low type I error rates and less power, while a too simple model (not including a random slope if there is one) has inflated type I error rates and higher power. Thus, it is essential to include random slopes if they are present because this prevents us from being overly confident (Schielzeth & Forstmeier 2009), but because of the power issue, we doubt that always starting with the maximally complex random effect structure (as suggested in Bell *et al.* 2019) is generally a

good advice. We therefore need some decision criteria for deciding on the complexity of the random effect structure.

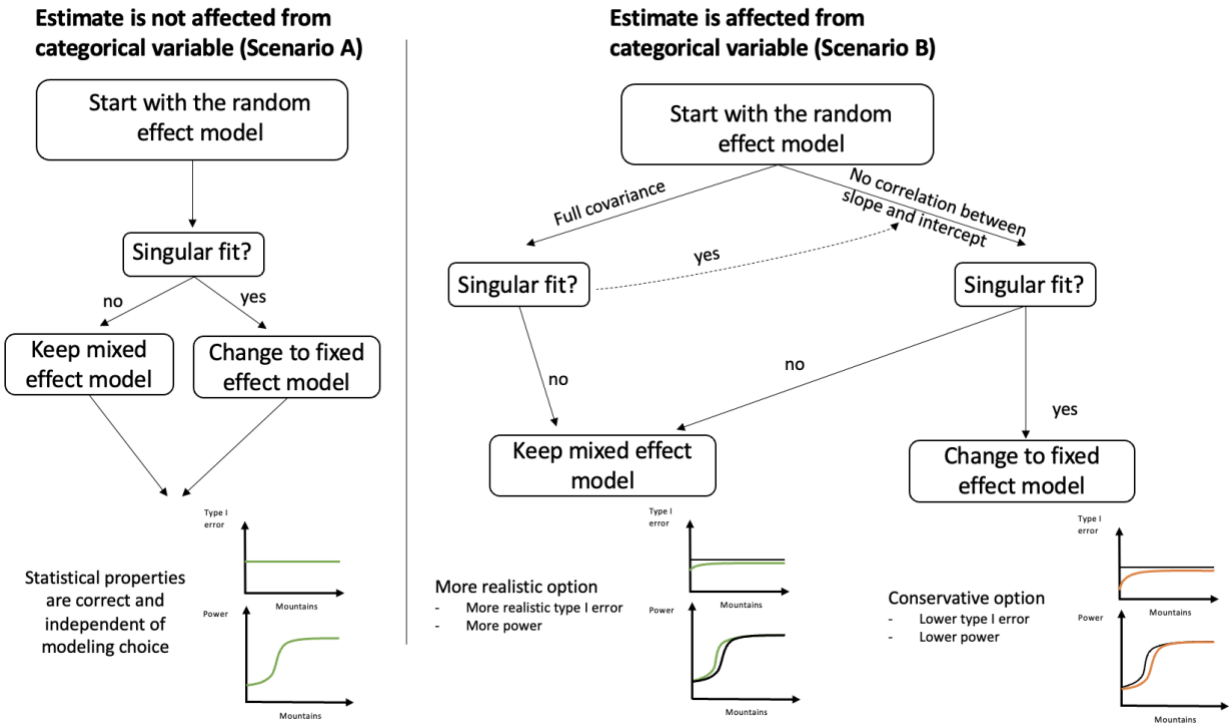


Figure 6: Consequences of mixed-effect models with less than 5 levels in a random effect. Two extreme events can be observed for the variance estimate of the random effect. On the one hand (left side), the variance estimate can be close to zero, which can result (if the random effect is not changed to a fixed effect) in a higher probability of a type I error for the associated fixed effects. On the other hand (right side), the variance estimate can be overestimated, which can result in a higher probability of a type II error for the associated fixed effects.

Unfortunately, model selection on random effects is relatively complicated, because the exact degrees of freedom (Kuznetsova *et al.* 2017, only approximations exist (for low number of levels a Kenward-Rogers correction is preferred McNeish 2017)), that are used by a random effect, are unclear. Although this prevents a naïve use of AIC or likelihood ratio

tests (LRT), other methods such as simulated (restricted) LRTs (Wiencierz *et al.* 2011) can be used to decide if adding a random slope is justified. Thus, we recommend starting with the simpler structure (typically random intercepts) and then use residual checks (e.g. Hartig 2019) or appropriate model selection criteria (e.g. Matuschek *et al.* 2017) to decide if a random slope should be added.

Another, somewhat perpendicular option would be to modify the assumption that the individual levels stem from a normal distribution. The normal is commonly chosen due to mathematical convenience (Beck & Katz 2007) and corresponds to a L2 (RIDGE) regularization on the random effects (the concept of regularization is common over all statistical fields and techniques, e.g. in Bayesian statistics regularization is done via shrinkage priors). An L2 regularization purposefully biases parameter estimates towards zero but it is unable to produce sparse estimates (Zou & Hastie 2005) i.e., the random effect estimates in the too complex model will typically not be shrunk to exactly zero.

From regularization theory we know that a l1 (LASSO) regularization, which corresponds to a Laplace distribution (Park & Casella 2008; Hans 2009; Tibshirani 2011) is able to produce more sparse estimators. Because mixed-effect models are relatively robust to distributional assumptions (Bell *et al.* 2019; Schielzeth *et al.* 2020), switching RE distributions to LASSO assumptions will make random effects more sparse while maintaining most of their statistical advantages. Thus, using random effect models with L1 or a combination of L1 and L2 regularization (elastic net, see Zou & Hastie 2005) could possibly allow to start with the maximal complex random effect structure, as the stronger shrinkage towards zero from the L1 structure ameliorates the loss of power observed here.

Conclusion

In conclusion, we showed that, contrary to common opinions, mixed-effect models correctly estimate the variance components of the random effects also for small numbers of random effect groups. We also found that the statistical properties of the ecological effect are robust against the model choice when the grouping variable does not share variance with the ecological effect in the data-generating process. When in doubt about the data-generating process, we encourage starting with a simplified model (random intercept only) because of power-issues with too complex models and consulting model diagnostics and likelihood ratio tests to check for evidence of random slope effects. When finding evidence for random slopes in these tests, we recommend going first the mixed-effect model and switching only to a fixed-effect model in case of a singular fit problem. These recommendations ensure conservative type I error rates and avoid overconfidence. Moreover, we demonstrated that the statistical properties of both choices, the fixed-effects and the mixed-effects model, depends on the study design, particularly on the number of levels and the magnitude of difference between levels further supporting an informative decision or the study design itself. With this work, we provide a practical guideline, which helps ecologists in the study design, the data analysis and thus making ecological inference more informative and robust.

Data availability statement:

No empirical data was used in this study. Code to run and analyze the experiments can be found at https://github.com/JohannesOberpriller/RandomEffect_Groups.

Author Statement:

MP, JO and MSL designed the study. MP and JO ran the experiments, analyzed the results and wrote a first draft. All authors contributed equally to revising the manuscript and interpreting and discussing results.

Acknowledgement

The idea of the manuscript originated from a discussion in the Theoretical Ecology seminar and was further developed in the Coding Club at University Regensburg. We thank Rainer Spang, Carsten Dormann, Magdalena Mair, Björn Reineking, Sean McMahon, Andreas Ettner and Florian Hartig for comments and discussions on earlier versions of the manuscript. JO was funded by the Bavarian Ministry of Science and the Arts in the context of Bavarian Climate Research Network (bayklif). MSL was funded by the Smithsonian Predoctoral Fellowship.

References

Aarts, E., Verhage, M., Veenvliet, J.V., Dolan, C.V. & van der Sluis, S. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience*, 17, 491–496.

Arnqvist, G. (2020). Mixed Models Offer No Freedom from Degrees of Freedom. *Trends in Ecology & Evolution*, 35, 329–335.

Baayen, R.H., Davidson, D.J. & Bates, D.M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, Special Issue: Emerging Data Analysis, 59, 390–412.

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

Beck, N. & Katz, J.N. (2007). Random Coefficient Models for Time-Series—Cross-Section Data: Monte Carlo Experiments. *Polit. anal.*, 15, 182–195.

Bell, A., Fairbrother, M. & Jones, K. (2019). Fixed and random effects models: making an informed choice. *Qual Quant*, 53, 1051–1074.

Boisgontier, M.P. & Cheval, B. (2016). The anova to mixed model transition. *Neuroscience & Biobehavioral Reviews*, 68, 1004–1005.

Bolker, B.M. (2015). Linear and generalized linear mixed models. *Ecological Statistics: Contemporary theory and application*, 309–333.

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, 127–35.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., *et al.* (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R journal*, 9, 378–400.

Brysbaert, M. & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *J Cogn*, 1.

Chen, Z. & Dunson, D.B. (2003). Random Effects Selection in Linear Mixed Models. *Biometrics*, 59, 762–769.

DerSimonian, R. & Laird, N. (1986). Meta-Analysis in Clinical Trials. *Controlled Clinical Trial*, 7, 177–188.

Gelman, A. (2005). Analysis of variance—why it is more important than ever. *The Annals of Statistics*, 33, 1–53.

Gelman, A. & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*.

Green, P. & MacLeod, C.J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7, 493–498.

Gunasekara, F.I., Richardson, K., Carter, K. & Blakely, T. (2014). Fixed effects analysis of repeated measures data. *Int J Epidemiol*, 43, 264–269.

Hans, C. (2009). Bayesian lasso regression. *Biometrika*, 96, 835–845.

Harrison, X.A. (2014). Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2, e616.

Harrison, X.A. (2015). A comparison of observation-level random effect and Beta-Binomial models for modelling overdispersion in Binomial data in ecology & evolution. *PeerJ*, 3, e1114.

Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C., *et al.* (2017). *Best practice in mixed effects modelling and multi-model inference in ecology* (No. e3113v1). PeerJ Inc.

Harrison, X.A., Donaldson, L., Correa-Cano, M.E., Evans, J., Fisher, D.N., Goodwin, C.E.D., *et al.* (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, 6, e4794.

Hartig, F. (2019). DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. *R package version 0.2*, 4.

Hedges, L.V. & Vevea, J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504.

Hox, J.J., Moerbeek, M. & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Johnson, P.C.D., Barry, S.J.E., Ferguson, H.M. & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, 6, 133–142.

Krueger, C. & Tian, L. (2004). A Comparison of the General Linear Mixed Model and Repeated Measures ANOVA Using a Dataset with Multiple Missing Data Points. *Biological Research For Nursing*, 6, 151–157.

Laird, N.M. & Ware, J.H. (1982). Random-Effects Models for Longitudinal Data. *Biometrics*, 38, 963–974.

Maas, C.J.M. & Hox, J.J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology*, 1, 86–92.

Martin, J.G.A., Nussey, D.H., Wilson, A.J. & Réale, D. (2011). Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. *Methods in Ecology and Evolution*, 2, 362–374.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H. & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, 94, 305–315.

McMahon, S.M. & Diez, J.M. (2007). Scales of association: hierarchical linear models and the measurement of ecological systems. *Ecology Letters*, 10, 437–452.

McNeish, D. (2017). Small Sample Methods for Multilevel Modeling: A Colloquial Elucidation of REML and the Kenward-Roger Correction. *Multivariate Behavioral Research*, 52, 661–670.

McNeish, D.M. & Stapleton, L.M. (2016). The Effect of Small Sample Size on Two-Level Model Estimates: A Review and Illustration. *Educ Psychol Rev*, 28, 295–314.

Meteyard, L. & Davies, R.A.I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112, 104092.

Millar, R.B. & Anderson, M.J. (2004). Remedies for pseudoreplication. *Fisheries Research*, 70, 397–407.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4, 133–142.

Park, T. & Casella, G. (2008). The Bayesian Lasso. *null*, 103, 681–686.

Pol, M. van de. (2012). Quantifying individual variation in reaction norms: how study design affects the accuracy, precision and power of random regression models. *Methods in Ecology and Evolution*, 3, 268–280.

Schelldorfer, J., Bühlmann, P. & Geer, S.V.D. (2011). Estimation for High-Dimensional Linear Mixed-Effects Models Using ℓ1-Penalization. *Scandinavian Journal of Statistics*, 38, 197–214.

Schielzeth, H. (2010). Simple means to improve the interpretability of regression coefficients. *Methods in Ecology and Evolution*, 1, 103–113.

Schielzeth, H., Dingemanse, N.J., Nakagawa, S., Westneat, D.F., Allegue, H., Teplitsky, C., *et al.* (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, n/a.

Schielzeth, H. & Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behav Ecol*, 20, 416–420.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73, 273–282.

Townsend, Z., Buckley, J., Harada, M. & Scott, M. (2013). The Choice between Fixed and Random Effects. In: *The SAGE Handbook of Multilevel Modeling*. SAGE Publications Ltd, 1 Oliver's Yard, 55 City Road, London EC1Y 1SP United Kingdom, pp. 73–88.

Wainwright, P.E., Leatherdale, S.T. & Dubin, J.A. (2007). Advantages of mixed effects models over traditional ANOVA models in developmental studies: A worked example in a mouse model of fetal alcohol syndrome. *Dev. Psychobiol.*, 49, 664–674.

Wiencierz, A., Greven, S. & Küchenhoff, H. (2011). Restricted likelihood ratio testing in linear mixed models with general error covariance structure. *Electronic Journal of Statistics*, 5, 1718–1734.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67, 301–320.

Zuur, A., Ieno, E., Walker, N., Saveliev, A. & Smith, G. (2009). *Mixed effects models and extensions in ecology with R*.