## Modelando a variabilidade: o maravilhoso mundo dos modelos mistos\*

Melina de Souza Leite

\* Esse texto foi extraído e levemente modificado da introdução da minha tese de doutorado ([disponível aqui](https://www.teses.usp.br/teses/disponiveis/41/41134/tde-22032023-120555/pt-br.php)). Se quiser citar algo do texto, por favor cite a tese (Leite, 2022) ao invés dessa postagem.

“Statistics is concerned with the variability

that is evident in any body of data.”

Searle et al. 2006

Percebendo rápida disseminação do uso de modelos mistos[[1]](#footnote-1) atualmente na ecologia, é de se pensar que sejam uma descoberta recente no campo da estatística. Ledo engano! Searle et al. (2006) nos conta a história do desenvolvimento do interesse estatístico nos componentes de variação (ou variância) dos dados e de como surgiram os precursores de modelos de efeito aleatório e mistos. Curiosamente, a primeira formulação de um modelo de efeitos aleatórios foi em 1861, no qual o astrônomo Airy assume que diferentes noites de observação telescópica devem possuir diferentes estimativas por efeitos de “circunstância atmosférica e pessoais” (Searle et al., 2006). Temos então *noite* como a primeira variável de efeitos aleatórios da história[[2]](#footnote-2)! Mas foi a Análise de Variâncias (ANOVA), desenvolvida por Ronald Fisher entre as décadas de 1920 e 1940, um dos primeiros métodos a resumir a variabilidade nos dados, permitindo particionar a variância observada entre categorias de um fator ou mais fatores (variável explicativa) da variabilidade residual dos dados. Posteriormente, o desenvolvimento dos modelos mistos substituiu o uso da ANOVA como a ferramenta mais comum para análises de variância e partição de variabilidade nos dados (Boisgontier & Cheval, 2016; B. M. Bolker et al., 2009; Wainwright et al., 2007). Mas afinal o que são modelos mistos e quais suas utilidades na ecologia?

Modelos mistos começaram a ganhar destaque nas ciências ecológicas apenas a partir dos anos 2000 (e.g., McMahon & Diez, 2007; Bolker, 2009; Cressie *et al.*, 2009; Zuur *et al.*, 2009). Ainda assim, acredito que pouco do potencial destes modelos está sendo aplicado e, portanto, este texto é também uma tentativa de promover o uso e estudo destes modelos na ecologia. Há quem diga que os modelos mistos deveriam ser a regra e não a exceção na aplicação de modelos lineares (e.g., McElreath, 2020) e que vieram para revolucionar a pesquisa ecológica, removendo restrições que forçaram ecólogos a simplificar em demasia os modelos estatísticos e a ignorar importantes distinções entre erros de medida, erros de processo e incerteza de modelos (Cressie et al., 2009). Uma das razões é que dados ecológicos são frequentemente agrupados/hierárquicos, necessitando de um modelo com estrutura aleatória para evitar pseudoreplicação (Arnqvist, 2020). Entretanto, ainda temos muito a aprender sobre o comportamento destes modelos, suas utilidades e limitações.

O que diferencia um modelo linear “simples” de um modelo misto é a distinção entre **efeitos fixos e aleatórios** na tomada de decisão de como uma variável categórica deve ser modelada. Isso depende primordialmente do objetivo do estudo, do contexto dos dados e desenho amostral, e de quais propriedades se espera obter do modelo e da variável. Uma das formas de decisão mais simples é avaliar se o interesse está nas estimativas de cada categoria (*narrow-sense inference*, McLean *et al.*, 1991), se as categorias são independentes, mutualmente exclusivas e completamente observáveis. Nesse caso, deveriam ser estimados por efeitos fixos, por exemplo, sexo ou tratamentos diferentes em experimentos. Efeito aleatórios são uma escolha geralmente quando a variabilidade entre as categorias e não as estimativas exatas de cada uma são de interesse (*broad-sense inference*), quando os níveis podem ser considerados como uma amostragem aleatória de uma população de valores, e o interesse é controlar por esta variabilidade (também podendo ser considerada pseudoreplicação caso não controlado), por exemplo blocos em experimentos ou outras unidades amostrais que agrupam muitas observações (sub-amostras). As decisões e formas de pensar um modelo com variáveis de efeitos fixos ou aleatórios e as diferenças técnicas entre eles não se esgota aqui e podem ser estudadas em Gelman & Hill (2007), Harrisson (2018), Oberpriller et al. (2022). No entanto, mesmo que se queira conceitualmente incluir uma certa variável categórica como efeito aleatório, outras questões devem ser analisadas do ponto de vista estatístico. Discutiremos mais sobre isso na próxima seção.

## Fixo ou aleatório? O que fazer quando há poucas categorias

Quando o pesquisador decide que vai modelar uma variável categórica como um efeito aleatório, existe uma segunda pergunta a ser feita: quantas categorias essa variável tem? Ter poucas categorias significa maior incerteza ou dificuldade em se estimar a variância entre elas. E quais as consequências para interpretação do modelo quando isso acontece? Qual o número mínimo de categorias para que uma variável seja modelada como um efeito aleatório? Essas perguntas, por simples que pareçam ainda não tinham sido investigadas a fundo, e eu me sentia frustrada toda vez que me deparava com estes questionamentos entre colegas. Na ecologia, não é difícil encontrar situações em que a variável possui poucas categorias, por exemplo em experimentos de campo no qual os blocos de experimentos são limitados.

Durante meu estágio sanduíche na Alemanha ([Universidade de Regensburg](https://www.uni-regensburg.de/biologie-vorklinische-medizin/theoretische-oekologie/startseite/index.html)), engajei dois alunos de doutorado do laboratório para realizarmos um estudo motivado justamente por estas perguntas (Oberpriller et al., 2022). Nós utilizamos simulações baseadas em um contexto ecológico simples de amostragem aninhada, como se quiséssemos estudar o efeito da temperatura na altura de plantas, estabelecendo transectos altitudinais em diferentes montanhas (ver Box 1 da publicação original). A princípio, a variável montanha deveria ser modelada como efeito aleatório (*broad-sense inference*), dado que plantas de uma mesma montanha não são amostras independentes (pseudoreplicação). Nós analisamos as consequências de tratar a variável montanha com 2 a 8 níveis como efeito fixo ou aleatório em modelos corretamente especificados (ou seja, mesmo modelo usado na simulação dos dados) e modelos alternativos (sub- ou sobreparametrizado). Com poucos níveis, as estimativas de variância dos efeitos aleatórios podem ser imprecisas e enviesadas e isso pode também influenciar as estimativas dos efeitos fixos de interesse (efeito de temperatura na altura da planta). Calculamos as taxas de erro tipo I[[3]](#footnote-3) e o poder estatístico destes modelos e encontramos que inferir a estrutura correta dos efeitos aleatórios é de grande importância para obter taxas de erro tipo I corretas. Elaboramos recomendações práticas para ecólogos na escolha das melhores estruturas do modelo dado as hipóteses e o desenho amostral do estudo. Uma destas recomendações é sempre começar com um modelo mistos com os efeitos aleatórios especificados independentemente do número de categorias e mudar para um modelo de efeitos fixos se houver problemas em estimar as variâncias dos termos aleatórios (“*singular fit*” ou variância estimada em zero). Com isso, acreditamos em escolhas mais informativas para análises baseados no desenho do estudo e natureza dos dados, e fazer inferências mais robusta para o caso de poucas categorias. Realizar este estudo foi muito gratificante! É um trabalho do qual me orgulho e espero que possa contribuir para o uso e desenvolvimento de modelos mistos na ecologia.

## Particionar a variabilidade: o “quanto” ao invés de “qual”

O uso de modelos cada vez mais complexos na ecologia é uma tendência que está ligada não só com a oferta das ferramentas, mas principalmente com uma mudança de mentalidade analítica, e possivelmente com o amadurecimento de teorias (Low-Décarie et al., 2014, p.). Há um bom tempo, Quinn & Dunham (1983) afirmaram ser enganosos modelar as variações observadas nos padrões e processos de comunidades atribuindo-as a apenas uma única causa, e propuseram uma visão não mutualmente exclusiva das causas de um fenômeno em termos da contribuição relativa de cada alternativa. Bolker (2007) chega à mesma conclusão ao dizer que, em vez de conceber um único experimento que determina qual explicação está certa, somos motivados a fornecer respostas quantitativas que descrevem que fração de um padrão em particular é explicada por diferentes mecanismos: o "quanto" em vez de "qual". Essa mudança de perspectiva nos leva, por exemplo, a repensar os processos ecológicos em ecologia de comunidades e como cada um contribui relativamente para a formação do padrão de abundâncias de espécies na comunidade. E as perguntas se tornam, por exemplo: *Qual a contribuição relativa de diferentes processos de montagem de comunidades após perda de habitat (Leite et al., 2022)? Qual a importância relativa de mecanismos de nicho na variabilidade de taxas vitais em árvores (Leite et al., 2024)?*

Para responder às perguntas de “quanto” em vez de “qual”, entram em cena os componentes de variação e as formas de quantificação da contribuição relativa destes na variabilidade total dos dados. Os primeiros métodos a se propor a partição de variâncias em ecologia de comunidades foram os métodos multivariado de RDA (Borcard et al. 1992), cujas ideias foram trazidas dos modelos lineares (p.ex. ANOVA). Em um modelo linear, a forma mais simples para quantificar o quanto de variabilidade dos dados (variável resposta) é explicado por variáveis de interesse (variáveis explicativas) é o uso do coeficiente de determinação (R2) que é expresso como uma proporção da variância total na resposta. Esta é uma forma crua de se assumir que existe algum tipo de variação decorrente de uma combinação de fatores desconhecidos (B. Bolker, 2007) que geralmente não são de interesse, os famigerados resíduos. Entretanto, R2 é um conceito intuitivo (Nakagawa & Schielzeth, 2013) e vai além dos testes de significância ao colocar os efeitos estudados em perspectiva da variação total nos dados (Stoffel et al., 2021), e que pode significar uma quantidade biológica relevante se esta variação é representativa da população total (de Villemereuil et al., 2018).

Modelos que fazem premissas explícitas sobre as causas subjacentes de variabilidade podem tanto prover mais informação sobre os processos ecológicos quanto obter mais dos dados (B. Bolker, 2007). Desta forma, podemos ir um pouco mais além da abordagem de partição de variâncias em modelos lineares ao especificar diferentes fontes de variabilidade e particionar a variância do que antes era tido como “não explicada” em diversos componentes. Tais componentes, por sua vez, podem refletir os diferentes processos ou grupos de processos ecológicos em questão. E aqui entram os modelos mistos e seus efeitos aleatórios, nos quais é possível atribuir uma parte da variação a componentes que não são efeitos em si, apenas fonte de variação, mudando assim a forma de entender o que “não é explicado”.

O histórico de uso de R2 em modelos mistos começa com o termo Coeficiente de Partição de Variâncias (em inglês, Variance Partition Coefficient, VPC) para descrever a porcentagem de variação que é atribuído a um nível ou classificação particular nos dados (Browne et al., 2005). Porém, dada a complexidade de modelos mistos e as relações multiníveis entre termos de efeitos fixos e aleatórios, calcular o VPC dos componentes ou mesmo o R2 total destes modelos não é tarefa fácil. Felizmente, muitos estudos recentes têm desenvolvido métodos para a partição de variâncias em modelos mistos, especialmente para modelos com distribuição não normal[[4]](#footnote-4) (Browne et al., 2005; Goldstein et al., 2002; Schulz et al., 2025) e muitos na área da ecologia (Ives, 2019; Nakagawa et al., 2017; Nakagawa & Schielzeth, 2013; Stoffel et al., 2021).

A flexibilidade de modelagem e as técnicas de partição de variâncias (R2 ou VPC) em modelos mistos me permitiu fazer uso criativo de variáveis categóricas que podem explicar parte da variação das comunidades de forma bem diferente do senso comum elencado nas explicações acima. Por exemplo, o modelo utilizado em Leite et al. (2022) é um *Joint Species Distribution Model* (JSDMs), uma classe de modelos mistos usados para modelar abundância e ocorrência de espécies em comunidades, podendo levar em consideração variáveis ambientais, atributos das espécies, autocorrelação espacial e filogenética (Miller et al., 2018; Ovaskainen et al., 2017; Pichler & Hartig, 2021; ter Braak, 2019; Warton et al., 2015). Eu criei um modelo baseado nas características do meu estudo e dos meus objetivos, interpretando então os R2 de cada componente de efeito aleatório e dos efeitos fixos de acordo com o que seria esperado encontrar mais variação nos dados de acordo com os processos dominantes de montagem. Por exemplo, escolhemos variáveis preditoras associadas aos processos de filtro ambiental, mas com a relação entre os atributos funcionais das aves e a o gradiente de cobertura florestal. Incorporamos também efeitos aleatórios para capturar a variação estocástica da ocorrência das espécies em manchas pequenas e isoladas de habitats remanescentes.

Já em Leite et al. (2024), o modelo proposto é um modelo de efeitos aleatórios apenas, ou seja, não existem variáveis de efeito fixo, pois o objetivo principal é a formulação de um modelo base (*baseline*) para partição de variâncias entre grupos de componentes que sabemos criar variabilidade nos dados (*organizing principles*) de forma a elucidar padrões macroecológicos e indicar foco em testes de mecanismos e hipóteses sobre variabilidade em taxas vitais de árvores.

Em suma, o uso de modelos mistos, ou apenas modelos de efeito aleatório, como base para construção de modelos mais complexos tem surgido timidamente na literatura (Camac et al., 2018; Schulz et al., 2025), e eu acredito ser uma vantagem no entendimento primordial das fontes de variabilidade nos dados antes mesmo de testar hipóteses e mecanismos. Concordo com Schulz et al., (2025) ao dizer que partição de variâncias baseada em modelos:

*“É uma ferramenta geral para resumir e interpretar modelos lineares (ou de regressão) para produzir ideias, insights adicionais em estudos ecológicos comparado com o que a tradicional inferência sobre parâmetros estimados pode revelar dos dados.”* (Schulz et al., 2025).

Em suas próximas análises, avalie se é possível e vale a pena aplicar a partição de variâncias em um model linear (misto) e veja o quanto mais você pode aprender dos seus dados e modelo!

## 

## Referências

Arnqvist, G. (2020). Mixed Models Offer No Freedom from Degrees of Freedom. *Trends in Ecology & Evolution*, *35*(4), 329–335. https://doi.org/10.1016/j.tree.2019.12.004

Boisgontier, M. P., & Cheval, B. (2016). The anova to mixed model transition. *Neuroscience & Biobehavioral Reviews*, *68*, 1004–1005. https://doi.org/10.1016/j.neubiorev.2016.05.034

Bolker, B. (2007). *Ecological Models and Data in R*. Princeton University Press.

Bolker, B. (2009). Learning hierarchical models: Advice for the rest of us. *Ecological Applications*, *19*(3), 588–592. https://doi.org/10.1890/08-0639.1

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Browne, W. J., Subramanian S. V., Jones K., & Goldstein H. (2005). Variance partitioning in multilevel logistic models that exhibit overdispersion. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(3), 599–613. https://doi.org/10.1111/j.1467-985X.2004.00365.x

Camac, J. S., Condit, R., FitzJohn, R. G., McCalman, L., Steinberg, D., Westoby, M., Wright, S. J., & Falster, D. S. (2018). Partitioning mortality into growth-dependent and growth-independent hazards across 203 tropical tree species. *Proceedings of the National Academy of Sciences*, *115*(49), 12459–12464. https://doi.org/10.1073/pnas.1721040115

Cressie, N., Calder, C. a., Clark, J. S., Ver Hoef, J. M., & Wikle, C. K. (2009). Accounting for uncertainty in ecological analysis: The strengths and limitations of hierarchical statistical modeling. *Ecological Applications*, *19*(3), 553–570. https://doi.org/10.1890/07-0744.1

de Villemereuil, P., Morrissey, M. B., Nakagawa, S., & Schielzeth, H. (2018). Fixed-effect variance and the estimation of repeatabilities and heritabilities: Issues and solutions. *Journal of Evolutionary Biology*, *31*(4), 621–632. https://doi.org/10.1111/jeb.13232

Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning Variation in Multilevel Models. *Understanding Statistics*, *1*(4), 223–231. https://doi.org/10.1207/S15328031US0104\_02

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6*, e4794. https://doi.org/10.7717/peerj.4794

Ives, A. R. (2019). R2s for Correlated Data: Phylogenetic Models, LMMs, and GLMMs. *Systematic Biology*, *68*(2), 234–251. https://doi.org/10.1093/sysbio/syy060

Leite, M. de S. (2022). *Incorporando a variabilidade: O uso de modelos de efeitos mistos em ecologia* [Text, Universidade de São Paulo]. https://doi.org/10.11606/T.41.2022.tde-22032023-120555

Leite, M. de S., Boesing, A. L., Metzger, J. P., & Prado, P. I. (2022). Matrix quality determines the strength of habitat loss filtering on bird communities at the landscape scale. *Journal of Applied Ecology*, *59*(11), 2790–2802. https://doi.org/10.1111/1365-2664.14275

Leite, M. de S., McMahon, S. M., Prado, P. I., Davies, S. J., Oliveira, A. A. de, De Deurwaerder, H. P., Aguilar, S., Anderson-Teixeira, K. J., Aqilah, N., Bourg, N. A., Brockelman, W. Y., Castaño, N., Chang-Yang, C.-H., Chen, Y.-Y., Chuyong, G., Clay, K., Duque, Á., Ediriweera, S., Ewango, C. E. N., … Hülsmann, L. (2024). Major axes of variation in tree demography across global forests. *Ecography*, *2024*(6), e07187. https://doi.org/10.1111/ecog.07187

Leite, M. de S., Queiroz, T. L., Galliez, M., de Mendonça, P. P., & Fernandez, F. a. S. (2013). Activity patterns of the water opossum Chironectes minimus in Atlantic Forest rivers of south-eastern Brazil. *Journal of Tropical Ecology*, *29*(03), 261–264. https://doi.org/10.1017/S0266467413000187

Low-Décarie, E., Chivers, C., & Granados, M. (2014). Rising complexity and falling explanatory power in ecology. *Frontiers in Ecology and the Environment*, 140821075159009. https://doi.org/10.1890/130230

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan. 2nd edition* (2nd ed.).

McLean, R. A., Sanders, W. L., & Stroup, W. W. (1991). A Unified Approach to Mixed Linear Models. *The American Statistician*, *45*(1), 54–64. https://doi.org/10.1080/00031305.1991.10475767

McMahon, S. M., & Diez, J. M. (2007). Scales of association: Hierarchical linear models and the measurement of ecological systems. *Ecology Letters*, *10*(6), 437–452. https://doi.org/10.1111/j.1461-0248.2007.01036.x

Miller, J. E. D., Damschen, E. I., & Ives, A. R. (2018). Functional traits and community composition: A comparison among community-weighted means, weighted correlations, and multilevel models. *Methods in Ecology and Evolution*, *10*(3), 415–425. https://doi.org/10.1111/2041-210X.13119

Nakagawa, S., Johnson, P. C. D., & 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. https://doi.org/10.1098/rsif.2017.0213

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*(2), 133–142. https://doi.org/10.1111/j.2041-210x.2012.00261.x

Oberpriller, J., Leite, M. de S., & Pichler, M. (2022). Fixed or random? On the reliability of mixed-effects models for a small number of levels in grouping variables. *Ecology and Evolution*, *12*(7), e9062. https://doi.org/10.1002/ece3.9062

Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. *Ecology Letters*, *20*(5), 561–576. https://doi.org/10.1111/ele.12757

Pichler, M., & Hartig, F. (2021). A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods in Ecology and Evolution*, *12*(11), 2159–2173. https://doi.org/10.1111/2041-210X.13687

Quinn, J. F., & Dunham, A. E. (1983). On Hypothesis Testing in Ecology and Evolution. *The American Naturalist*, *122*(5), 602–617.

Schulz, T., Saastamoinen, M., & Vanhatalo, J. (2025). Model-based variance partitioning for statistical ecology. *Ecological Monographs*, *95*(1), e1646. https://doi.org/10.1002/ecm.1646

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance Components*.

Stoffel, M. A., Nakagawa, S., & Schielzeth, H. (2021). partR2: Partitioning R2 in generalized linear mixed models. *PeerJ*, *9*, e11414. https://doi.org/10.7717/peerj.11414

ter Braak, C. J. F. (2019). New robust weighted averaging- and model-based methods for assessing trait–environment relationships. *Methods in Ecology and Evolution*, *10*(11), 1962–1971. https://doi.org/10.1111/2041-210X.13278

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. *Developmental Psychobiology*, *49*(7), 664–674. https://doi.org/10.1002/dev.20245

Warton, D. I., Blanchet, F. G., O’Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. C. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution*, *30*(12), 766–779. https://doi.org/10.1016/j.tree.2015.09.007

Zuur, A., Ieno, E. N., Walker, N., Saveiliev, A. A., & Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R*. Springer.

1. Ou modelos de efeito misto, ou modelos hierárquicos ou multiníveis. [↑](#footnote-ref-1)
2. Este fato me lembrou do meu artigo do mestrado (Leite et al., 2013), no qual também usei noite**,** além do indivíduo, como efeito aleatório para modelar os padrões de atividade da cuíca d’água, *Chironectes minimus*. Este foi o primeiro modelo misto que construí e publiquei! [↑](#footnote-ref-2)
3. Erro tipo I é quando não existe efeito da variável preditora sobre a resposta, neste caso, não existe efeito de temperatura nas plantas, mas o teste aponta que sim. [↑](#footnote-ref-3)
4. Particionar variância em modelos lineares de distribuição não normal, como Poisson ou Binomial, é um pouco mais complicado, porque a variância residual não pode ser facilmente definida nestes modelos. Assim, geralmente chama-se este tipo de R2 de pseudo-R2, pois, apesar de ter mesmas propriedades quanto à interpretação, são calculados de formas diferentes. [↑](#footnote-ref-4)