Model-based ordination for phenological studies: from controlling sampling bias to inferring temporal associations

Hao Ran Lai[1,2,*]

[1]School of Biological Sciences, University of Canterbury, Christchurch, Aotearoa New Zealand

[2]South East Asian Rainforest Research Partnership (SEARRP), Kota Kinabalu, Malaysia

[*]Corresponding email: hrlai.ecology@gmail.com

Abstract

1. Willig et al. (*Methods in Ecology and Evolution*, 15, 868–885, 2024) cautioned that unequal sampling effort and pseudoreplication can bias the characterisation of species phenology using circular statistics. Borrowing concepts from rarefaction, they proposed bootstrapping to control for time-varying marginal totals that arise from unequal sampling effort over time.

2. This study extends their cautionary notes to regressions of phenological time series, where bootstrapping can be replaced by various built-in functionalities of generalised linear mixed-effect models. I further take this opportunity to borrow a key innovation in model-based ordination and joint species distribution modelling — generalised linear latent variable models (GLLVM) — to illustrate its ability in extracting more information out of multispecies phenological data beyond circular statistics.

3. **Synthesis:** With sampling-bias adjustment, GLLVMs, or regressions in general, are robust predictive and inferential tools that enrich our phenological understandings in conjunction with circular statistics for hypothesis testing.

Introduction

A recent paper by Willig et al. (2024) highlighted how statistical analyses of phenological data could be biased by unequal sampling effort and pseudoreplication. When sampling events differ systematically in marginal totals (i.e., total number of observations per time point), they showed that circular statistics used to characterise periodic time series can lead to misleading conclusions. For unbiased circular statistics of uniformity (i.e., whether phenology is spread across or concentrated within certain periods), they proposed bootstrapping to fix marginal totals across time — a concept similar to rarefaction. After controlling for marginal totals, the resulting *proportional* quantities of phenology may display an opposite circular pattern compared to the raw counts. This cautionary advice is timely as ecologists are increasingly reliant on heterogeneous observations, such as herbarium specimens, to address climate-induced phenological changes in data-poor regions (Davis et al., 2022).

However, bootstrapping is not necessary for time-series regressions. Instead, regressions using generalised linear mixed-effect modelling (GLMM) have build-in functionalities to account for sampling effort and pseudoreplication in multiple ways: likelihood ("distribution family"), offset, covariate and/or random-effect structure (Bolker et al., 2009; Zuur et al., 2009). While circular statistics aim to test whether phenology is uniform or modal (Landler et al., 2020), regressions aim to predict phenological quantities at a given time (Fidino & Magle, 2017). Both goals are complementary. My aim is therefore to extend Willig et al. (2024)'s message to regressions of phenology (or time series in general) for completeness. I further take this opportunity to leverage a recently development in GLMM — generalised linear latent variable models (GLLVMs; Hui et al., 2015; Niku et al., 2019) — to characterise species' temporal niche (Zurell et al., 2024) using an accessible `R` package, `glmmTMB` (Brooks et al., 2017).

A regression recipe for phenology

Consider a measured quantity of phenology Y_t observed over discrete time step t. In ecology, Y_t usually do not have Normal error distributions and can either be skewed continuous (e.g., leaf-litter biomass) or discrete (e.g., number of reproductive individuals). Willig et al. (2024) added that phenological data often contain pseudoreplication (e.g., repeated sampling from the same individuals), sampling bias (e.g., towards more abundant species) and unequal sampling effort (e.g., different total

number of individuals sampled at each t). Fortunately, both non-Gaussian responses and sampling issues can be flexibly accommodated by GLMM.

Using the same data from Willig et al. (2024), I re-analysed the reproductive phenologies of five Amazonian bat species in a single GLMM to demonstrate its ability to account for unequal sampling effort and provide novel insights. Let Y_{jt} be the total number of pregnant female individuals in bat species j at time t, our basic GLMM can be:

$$\begin{aligned} Y_{jt} &\sim \text{Binomial}\left(N_{jt},\ p_{jt}\right) \\ \text{logit}\left(p_{jt}\right) = \eta_{jt} &= f(\text{Month-specific variables},\ \text{Species-specific coefficients}), \end{aligned} \tag{1}$$

where N_{jt} is the total sample size of species j that could vary across time t. A binomial GLMM therefore accounts for marginal totals as Willig et al. (2024) have alluded to. For dealing with unequal sampling effort that is continuous (e.g., plot size or observation duration), including an offset term could standardise the linear predictor to amount per area or per time (e.g., leaf-litter mass per area or number of pollinator per hour; Warton et al., 2015). Alternatively, including a proxy of sampling bias (e.g., distance from road) as a predictor will also allow us to "zero out" the bias by setting the predictor to zero when making predictions (Warton et al., 2013). Here I will only focus on accounting for sampling effort by explicitly stating trial sizes via a binomial GLMM.

With sampling effort being controlled, the next step is to model the proportion of pregnant females of species j at time t, p_{jt}. With the canonical logit link function, our linear predictor η_{jt} includes month-specific variables and species-specific coefficients to predict species phenology by month (Equation 1). There are various options to formulate the model, including autoregressive model (Hyndman & Athanasopoulos, 2021) and Fourier-based cosinor rhythmometry (Fidino & Magle, 2017; Lai et al., 2025). Instead of reiterating these approaches, here I introduce a formulation with GLLVM — a class of GLMM that has become increasingly popular in joint species distribution modelling (JSDM) due to its ability to infer multiple species' spatial niches while accounting for their non-independence (Niku et al., 2021), but remains underused in phenological studies. Applying GLLVM to multivariate time series represents the characterisation of species' temporal niche from their joint phenologies. Our linear predictor is thus:

$$\eta_{jt} = \alpha_0 + \alpha_t + \beta_{0j} + \sum_{k}^{K} X_{tk}\beta_{jk} + \sum_{m}^{M} Z_{tm}\theta_{jm}\,, \tag{2}$$

where α_0, α_t and β_{0j} are the overall fixed intercept, month-specific random intercepts and species-specific random intercepts, respectively; they capture the community-, month- and species-average reproduction. The predictors X_{tk} represent measured monthly variables, such as precipitation and temperature. When these abiotic variables are available, we could explain monthly phenology with species' environmental responses, β_{jk}. Lastly, the latent component contains month-specific latent variable Z_{tm} that accounts for unmeasured or missing predictors, while θ_{jm} captures species' responses to these latent month variables. Similar to JSDM, Z_{tm} and θ_{jm} are interpretable as month factors and species loadings in an ordination, allowing us to infer species' temporal niche as their affinities to particular months (Zurell et al., 2024).

When monthly environments are not available (as in this study), Equation 2 reduces to a pure latent variable model (Hui, 2016):

$$\eta_{jt} = \alpha_0 + \alpha_t + \beta_{0j} + \sum_{m}^{M} Z_{tm}\theta_{jm}\,, \tag{3}$$

which resembles unconstrained ordination of species in a latent temporal space based on their phenology. The next decision is how many latent dimensions to use. I fitted Equation 3 to Willig et al. (2024)'s dataset using two latent dimensions ($M = 2$) for three reasons: (i) two is the minimum number of axes to visualise ordination in a conventional sense, (ii) a larger number would quickly move us away from parsimony since there are only five species and twelve months, and (iii) the two leading latent dimensions are relatable to Fourier decomposition of time series with a single annual periodicity (see Appendix S1 in Supporting Information).

The GLLVM was fitted with the `glmmTMB` v1.1.10 package (Brooks et al., 2017) in `R` v4.3.3 (R Core Team, 2024). The model's formula syntax was

```
cbind(Pregnant, Not_pregnant) ~
    1 + (1 | month) + (1 | species) + rr(species + 0 | month, d = 2)
```

which maps directly to Equation 3 (see also Table S1). Note that the same could also be achieved with the `gllvm` package (Niku et al., 2019). Using the latent species loadings θ_{jm}, I then calculated two types of species–species associations (direct-and-indirect vs. direct-only) across months following Hui (2016) and Popovic et al. (2019) to demonstrate the additional insight about resource partitioning that GLLVM provides.

Results and Discussions

Controlling for unequal sampling effort across months, the binomial GLLVM predictions (Fig. 1a) compared favourably to the observed proportions and modalities in Willig et al. (2024, see their Fig. 4). More interestingly, the model-based ordination (Fig. 1b) clustered species by their phenology similarly to the original conclusion in Willig and Presley (2023). The first latent dimension distinguishes the dry season (July–September) from other months, separating two species (*Artibeus lituratus* and *Glossophaga soricina*) with peak pregnancy during drier months from the rest. The second latent dimension somewhat distinguishes earlier months from later months, suggesting a *plant* phenological gradient from floral nectar to fruit availability; this is evident in the separation of nectarivore *Glossophaga soricina* from the remaining fruigivores.

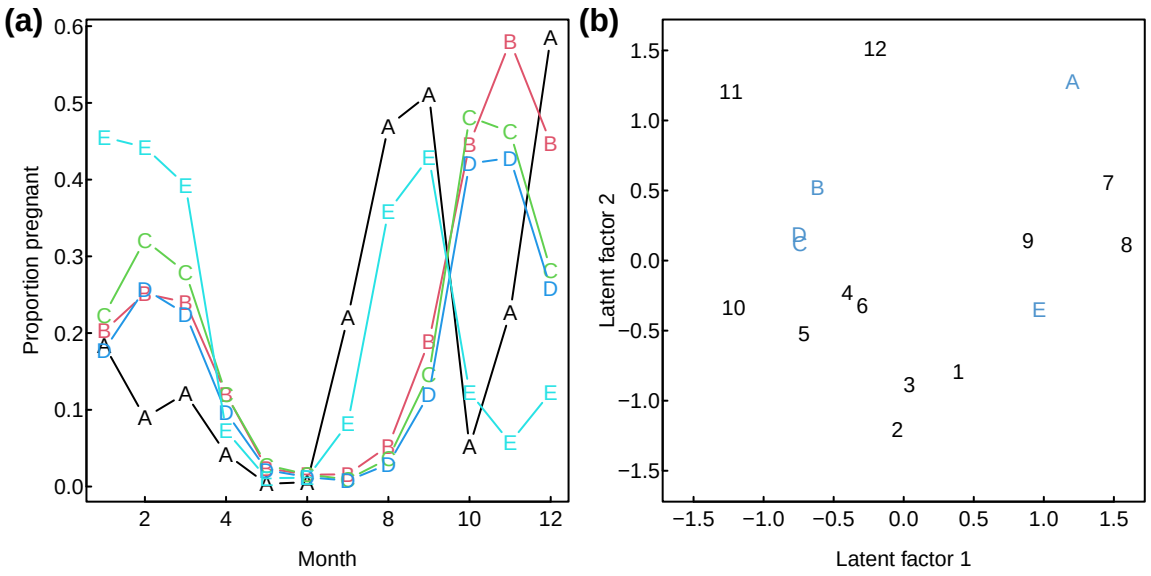

Figure 1: **(a)** Predicted proportion of pregnant females, p, across months. Lines with different letters and colours denote individual species. **(b)** Model-based ordination of species and months in two-dimensional latent space. Letters denote species and corresponding to panel a, while numbers denote month. Species key: A = *Artibeus lituratus*, B = *A. planirostris*, C = *Carollia brevicauda*, D = *C. perspicillata*, E = *Glossophaga soricina*.

Regressions also provide two additional insights unavailable from circular statistics. After controlling for marginal totals, the random intercepts (α_t and β_{0j}) in GLLVM further standardise the phenological ordination by month and species average reproduction (Hui et al., 2015). This is important to ensure that species–species associations only reflect their joint temporal fluctuations, rather than a mixture of both temporal fluctuations and overall abundance or fecundity (though one may drop the random intercepts if the goal is to capture both as a life-history whole). Another nuance lies in the antagonistic associations among three frugivorous bats (Fig. 2b) despite their positive correlations

in phenology (Fig. 2a). This suggests that species that cooccur temporally due to shared resource preferences (positive correlations in Fig. 2a) may in fact be competing directly (negative precisions in Fig. 2b). I will be brief about direct vs. indirect associations here and refer to Popovic et al. (2019) for further discussions. A related question is whether independent bootstrapping per species (as in Willig et al., 2024) is valid when species phenologies are correlated.

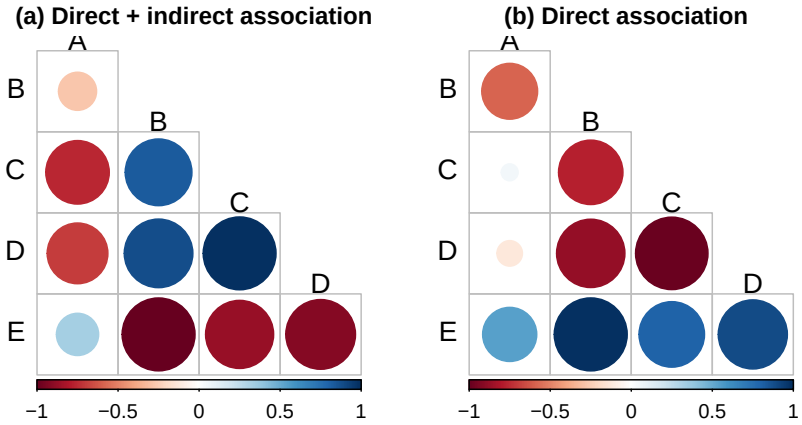

Figure 2: Opposite pairwise species associations revealed by direct-and-indirect associations **(a)** versus direct-only associations **(b)**, which were calculated as correlations and precisions respectively (Popovic et al., 2019). See Fig. 1 for species key.

Here I have only touched the minimum capabilities of GLLVM for phenology. The basic regression recipe could include more ingredients, including environmental or anthropogenic predictors, additional random effects to account for other sources of pseudoreplication (e.g., spatial autocorrelation), an offset term to standardise phenology by area, or proxy covariates to adjust for preferential sampling (see Table S1 for details). Furthermore, GLLVM can combine predictor components from other regression tools, such as autoregression and cosinor rhythmometry (Hyndman & Athanasopoulos, 2021; Lai et al., 2025), provided that informed decisions are made to avoid overfitting. To better connect regressions to circular statistics, it is worth exploring the von Mises distribution (Godoy et al., 2009; Graves et al., 2024) available in some GLMM packages in `R` (e.g., `brms`; Bürkner, 2021). These are accessible solutions to most if not all issues listed in Willig et al. (2024), at least for predictive and inferential purposes, and will enable ecologists to properly accelerate phenological studies using unconventional data from herbaria (Davis et al., 2022) and citizen science (Binley & Bennett, 2023).

References

Binley, A. D., & Bennett, J. R. (2023). The data double standard. *Methods in Ecology and Evolution*, *14*(6), 1389–1397. https://doi.org/10.1111/2041-210X.14110

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 and Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*(2), 378–400. https://doi.org/10.32614/RJ-2017-066

Bürkner, P. C. (2021). Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software*, *100*(5). https://doi.org/10.18637/JSS.V100.I05

Davis, C. C., Lyra, G. M., Park, D. S., Asprino, R., Maruyama, R., Torquato, D., Cook, B. I., & Ellison, A. M. (2022). New directions in tropical phenology. *Trends in Ecology and Evolution*, *37*(8), 683–693. https://doi.org/10.1016/j.tree.2022.05.001

Fidino, M., & Magle, S. B. (2017). Using Fourier series to estimate periodic patterns in dynamic occupancy models. *Ecosphere*, *8*(9). https://doi.org/10.1002/ecs2.1944

Godoy, O., Richardson, D. M., Valladares, F., & Castro-Díez, P. (2009). Flowering phenology of invasive alien plant species compared with native species in three Mediterranean-type ecosystems. *Annals of Botany*, *103*(3), 485–494. https://doi.org/10.1093/aob/mcn232

Graves, S., Spitz, G., & Manzitto-Tripp, E. (2024). Observing Shifts In Global Tropical Flowering Phenology. *Research Square*, 1–11. https://doi.org/10.21203/rs.3.rs-4469241/v1

Hui, F. K. C. (2016). BORAL – Bayesian Ordination and Regression Analysis of Multivariate Abundance Data in R. *Methods in Ecology and Evolution*, *7*(6), 744–750. https://doi.org/10.1111/2041-210X.12514

Hui, F. K., Taskinen, S., Pledger, S., Foster, S. D., & Warton, D. I. (2015). Model-based approaches to unconstrained ordination. *Methods in Ecology and Evolution*, *6*(4), 399–411. https://doi.org/10.1111/2041-210X.12236

Hyndman, R., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (3rd editio). OTexts.

Lai, H. R., Hill, T., Stivanello, S., & Chapman, H. M. (2025). Changes in quantity and timing of foliar and reproductive phenology of tropical dry-forest trees under a warming and drying climate. *Journal of Ecology*. https://doi.org/10.1101/2024.03.24.585819

Landler, L., Ruxton, G. D., & Malkemper, E. P. (2020). Grouped circular data in biology: advice for effectively implementing statistical procedures. *Behavioral Ecology and Sociobiology*, *74*(8). https://doi.org/10.1007/s00265-020-02881-6

Niku, J., Hui, F. K., Taskinen, S., & Warton, D. I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, *10*(12), 2173–2182. https://doi.org/10.1111/2041-210X.13303

Niku, J., Hui, F. K., Taskinen, S., & Warton, D. I. (2021). Analyzing environmental-trait interactions in ecological communities with fourth-corner latent variable models. *Environmetrics*, *32*(6), 1–17. https://doi.org/10.1002/env.2683

Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K., & Moles, A. T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. *Methods in Ecology and Evolution*, *10*(9), 1571–1583. https://doi.org/10.1111/2041-210X.13247

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Warton, D. I., Foster, S. D., De'ath, G., Stoklosa, J., & Dunstan, P. K. (2015). Model-based thinking for community ecology. *Plant Ecology*, *216*(5), 669–682. https://doi.org/10.1007/s11258-014-0366-3

Warton, D. I., Renner, I. W., & Ramp, D. (2013). Model-based control of observer bias for the analysis of presence-only data in ecology. *PLoS ONE*, *8*(11). https://doi.org/10.1371/journal.pone.0079168

Willig, M. R., & Presley, S. J. (2023). Reproductive phenologies of phyllostomid bat populations and ensembles from lowland Amazonia. *Journal of Mammalogy*, *104*(4), 752–769. https://doi.org/10.1093/jmammal/gyad032

Willig, M. R., Rojas-Sandoval, J., & Presley, S. J. (2024). Phenological patterns in ecology: Problems using circular statistics and solutions based on simulations. *Methods in Ecology and Evolution*, *15*(5), 868–885. https://doi.org/10.1111/2041-210X.14316

Zurell, D., Zimmermann, N. E., & Brun, P. (2024). The niche through time: Considering phenology and demographic stages in plant distribution models. *Journal of Ecology*, *112*(9), 1926–1939. https://doi.org/10.1111/1365-2745.14361

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