How to estimate effects of population versus species

In July 2020 during a ranges meeting discussion we discussed how we assume that the cues do not vary too much across a species' range. We realized we could test (a little) this assumption with the data we have on populations. We tested this somewhat in the budburst ms's latitude model (thank you Cat, who led work on this!).

This is what the latitude model looked like:

$$y_i = \alpha_{sp[i]} + \beta_{forcing_{sp[i]}} + \beta_{photoperiod_{sp[i]}} + \beta_{chilling_{sp[i]}} + \beta_{latitude_{sp[i]}} + \beta_{photoperiod:latitude_{sp[i]}} + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$

The α and each of the five β coefficients were modeled at the species level, as follows:

$$\begin{aligned}
\alpha_{sp} &\sim N(\mu_\alpha, \sigma_\alpha) \\
\beta_{forcing_{sp}} &\sim N(\mu_{forcing}, \sigma_{forcing}) \\
\beta_{photoperiod_{sp}} &\sim N(\mu_{photoperiod}, \sigma_{photoperiod}) \\
\beta_{chilling_{sp}} &\sim N(\mu_{chilling}, \sigma_{chilling}) \\
\beta_{latitude_{sp}} &\sim N(\mu_{latitude}, \sigma_{latitude}) \\
\beta_{photoperiod:latitude_{sp}} &\sim N(\mu_{photoperiod:latitude}, \sigma_{photoperiod:latitude})
\end{aligned}$$

In the latitude model, we had one aspect of the range that we were focued on it—latitude—and included it as a continuous predictor that we could test for an interaction with the photoperiod effect. For the ranges work, we're more interested in the effect of population as a grouping factor (not as a continuous predictor). What we need to decide then is how to represent population in the model.

Our main choices (I think) are to treat population as a *crossed* with species or *nested* within species. Some of our studies have examined multiple species from (the same) multiple locations (i.e., they examine species A and B at locations Z, Y, and X). This is what is classically considered a *crossed design* because you have information on each species at each location. In contrast, if each location was unique to each species (i.e., species A is examined at location X and and Y, and species B is examined at Z and W) and there are enough populations to estimate both a species and population effect, then population would be nested within species (note that if generally you observe just one population of each species, then you don't have enough information to separate out species versus population effects).

Let's start thinking about this from the intercept perspective: if you have a crossed design you would model it this way (shown just for forcing for simplicity):

$$y_i = \alpha + \alpha_{sp[i]} + \alpha_{pop[i]} + \beta_{forcing_{sp[i]}} + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$
$$\alpha_{sp} \sim N(0, \sigma_{\alpha sp})$$
$$\alpha_{pop} \sim N(0, \sigma_{\alpha pop})$$
$$\beta_{forcing_{sp}} \sim N(\mu_{forcing}, \sigma_{forcing})$$

There's a grand mean (α) and added onto that are effects of species (some species leafout inherently early or late across all populations) and a population effect (if this is each unique location, then it means all species at the same location leaf out earlier or later compared to another location—this makes sense as some places are colder or warmer).

In contrast for a nested model:

$$y_i = \alpha_{sp[pop[[i]]]} + \beta_{forcing_{sp[i]}} + \epsilon_i,$$

$$\epsilon_i \sim N(0, \sigma_y^2)$$
$$\alpha_{sp[pop]} \sim N(\alpha_{sp}, \sigma_{\alpha - sp})$$
$$\alpha_{sp} \sim N(\mu_\alpha, \sigma_\alpha)$$
$$\beta_{forcing_{sp}} \sim N(\mu_{forcing}, \sigma_{forcing})$$

Here, $\alpha_{sp[pop]}$ is a vector giving effects of each population nested within each species: so each population effect is drawn from a distribution that defines each species effect (the distribution of α_{sp}). Thus—if you had the same population (location) for two species—this model would allow different effects for the same location based on species. For simplicity, all population variances are the same ($\sigma_{\alpha - sp}$ is one number, not a vector, though it could be a vector). I skipped the grand mean here, but you can have it mathematically if you switch things around.

When you think just about the intercept, it seems easy to think that you should have crossed effects since some of the studies are designed that way and we do think different locations might be inherently early or late, but remember that these are cuttings that then go into experimental treatments and that we're interested in differences on the slopes especially (how does the cue vary by population?). Then, I think, it gets messy.

Do we expect the **same** shift to a forcing cue (for example) for all species sampled at the same location (in which case it's crossed) or do we expect the shift may depend on the species (in which case it's nested)? The problem in many ways relates to terminology. I specifically referred to populations here—which I think are unique to species (so, nested)—but if I had referred to locations or sites, it would seem more obviously crossed (and if I wrote, 'provenance'). I think also, you would code it differently... in Stan I would code each location x species as a unique ID (that is location X could be ID=1 for species A and ID=2 for species B), whereas location has to be consistent across species for crossed (location X=1 for species A and B).

Here's the model for nested effects, again considering just forcing:

$$y_i = \alpha_{sp[pop[[i]]]} + \beta_{forcing_{sp[pop[[i]]]}} + \epsilon_i,$$

$$\begin{aligned}
\epsilon_i &\sim N(0, \sigma_y^2) \\
\alpha_{sp[pop]} &\sim N(\alpha_{sp}, \sigma_{\alpha - sp}) \\
\alpha_{sp} &\sim N(\mu_\alpha, \sigma_\alpha) \\
\beta_{forcing_{sp[pop]}} &\sim N(\mu_{forcing[sp]}, \sigma_{forcing[sp]}) \\
\beta_{forcing_{sp}} &\sim N(\mu_{forcing}, \sigma_{forcing})
\end{aligned}$$

There's also a logistical issue in that I am not sure how to code a crossed slope! My guess is this, but I am not sure and could not find any notes online about it (I should crack open a book):

$$y_i = \alpha + \alpha_{sp[i]} + \alpha_{pop[i]} + \beta_{forcing} + \beta_{forcing_{sp[i]}} + \beta_{forcing_{pop[i]}} + \epsilon_i,$$

$$\begin{aligned}
\epsilon_i &\sim N(0, \sigma_y^2) \\
\alpha_{sp} &\sim N(0, \sigma_{\alpha sp}) \\
\alpha_{pop} &\sim N(0, \sigma_{\alpha pop}) \\
\beta_{forcing_{sp}} &\sim N(0, \sigma_{forcing - sp}) \\
\beta_{forcing_{pop}} &\sim N(0, \sigma_{forcing - pop})
\end{aligned}$$

This model is a bigger ask I think because you are trying to seperate out population from species effects. In a nested model, you assume population effects are dependent on species—you regularize more.

There's also the issue that traditionally if you want to partition variance, then you run a nested model and you get the direct comparison of how much variance there is in population versus species. But we're smart Bayesian folk, we could come up with other ways to get the comparison we want, so I have tried to lay this out to let us think more about what model makes the most sense for our interpretation of the world.