error structure thinking

mark sorel

10/30/2020

The background

Model the production of each emigrant life history type as a function of eggs and use correlated errors to account for dependency among life history types.

$$J_{h,y,a} = f(S_{y,a}|\boldsymbol{\theta}_{h,a})e^{\epsilon_{h,y,a}}$$

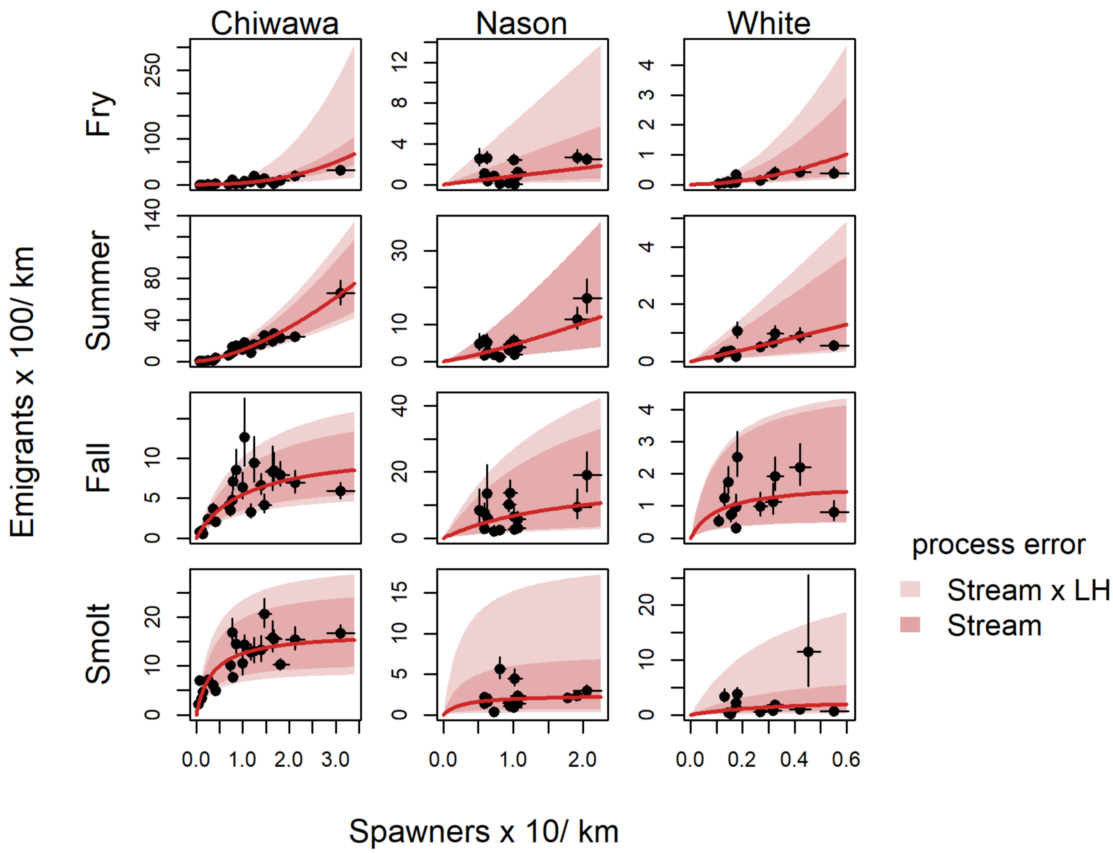
where f is the funcitonal form of the relationship between spawners $S_{y,a}$ in year y in stream a and juveniles $J_{h,y,a}$ expressing life history h with parameter vector $\boldsymbol{\theta}_{h,a}$ and the residual errors $\epsilon_{h,y,a}$. The residual errors $\epsilon_{h,y,a}$ are modeled as,

$$\epsilon_{h,y,a} = \mathbf{x}'_{h,y,a}\boldsymbol{\beta}_{h,a} + \zeta_{y,a} + \eta_{h,y,a}$$

with the effects $\boldsymbol{\beta}_{h,a}$ of covariates $\mathbf{x}_{h,y,a}$, and additional synchronous components $\zeta_{y,a} \sim N(0, \sigma^2_{\zeta_a})$ and asynchronous components $\eta_{h,y,a} \sim N(0, \sigma^2_{\eta_{h,a}})$.

Note I tried including error that was common among all streams and life histories each year, but it shrank to zero. Down below where I talk about error you can see why.

One example of what this model might look like is this:

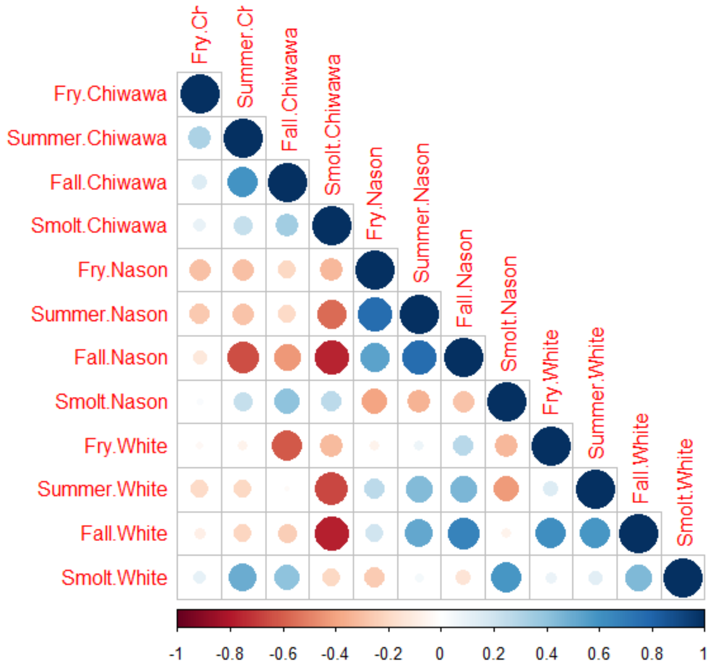


The question

Does it make more sense to account for "time ordering" of the life histories when modeling their shared error within streams. In other words, should I allow for fry and summer parr to have more error in common than fry and smolts. I'm imagining I would accomplish this with a multivariate normal error distribution for each stream, where there are correlations between consecutive life histories. Similarly, perhaps it makes sense to explicitly model the correlations between individual streams.

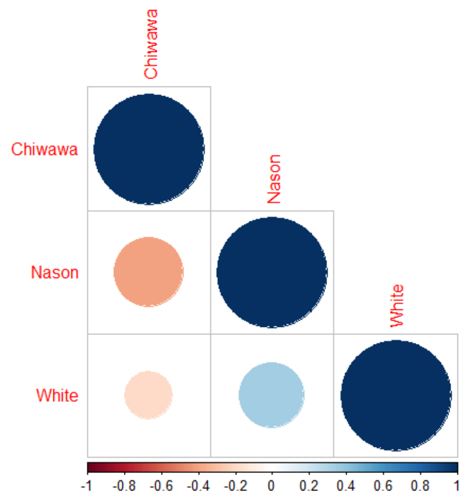
My first thought was to look at the errors for the model I describe above and see if this might be supported.

First, I'll show the correlation among asynchronous errors $\eta_{h,y,a} \sim N(0, \sigma^2_{\eta_{h,a}})$, but from a model where **all** the error (not explained by covariates) is asynchronous (i.e., $\zeta_{y,a}$ was fixed at 0). To be clear, these correlations are *post hoc* from simply looking at the correlations of the empirical Bayes (can be thought of as posterior mean) estimates of the random effects from the above models. What I'm wondering is if I should try to make *some* of this correlation more explicit in the model, knowing that an ultimate goal here is prospective simulation.

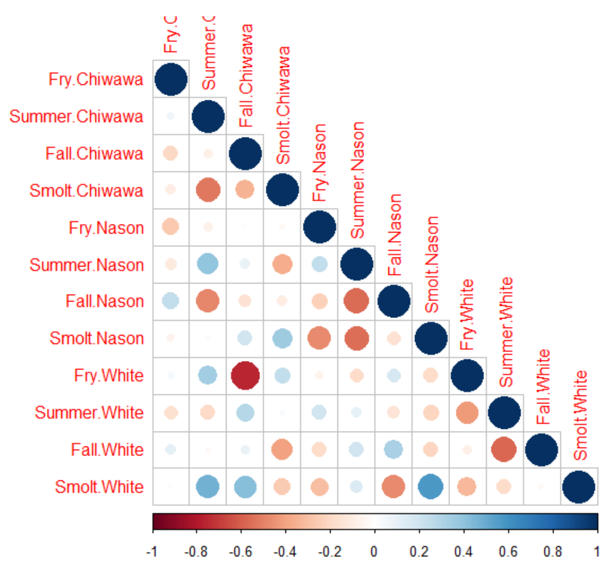


You can see some evidence that consecutive life histories (e.g., summer and fall, fall and smolt) might be more correlated within streams than nonconsecutive life histories (e.g., fry and smolts). Looking at the values adjacent to the diagonal.

Now, for the model with both synchronous and asynchronous errors. First the across lifehistory stream-specific errors. You can see that the relationship between streams differs. Maybe this is why the across-stream errors shrank to zero.



Now for the residual asynchronous error.



As expected, there is less correlation than when I didn't account for the stream-specific synchronous error (the first correlation plot). So that is good.

Overall, I think that my current approach is fine for looking at ICC and all that (so fine for this stand alone chapter), but I'm wondering if others think I need a more structured approach to modeling the process error if I want to use this for prospective simulation at some point down the road.