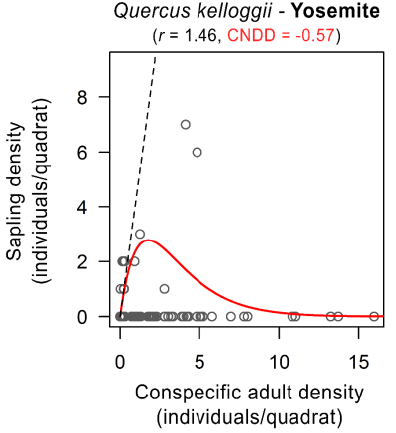
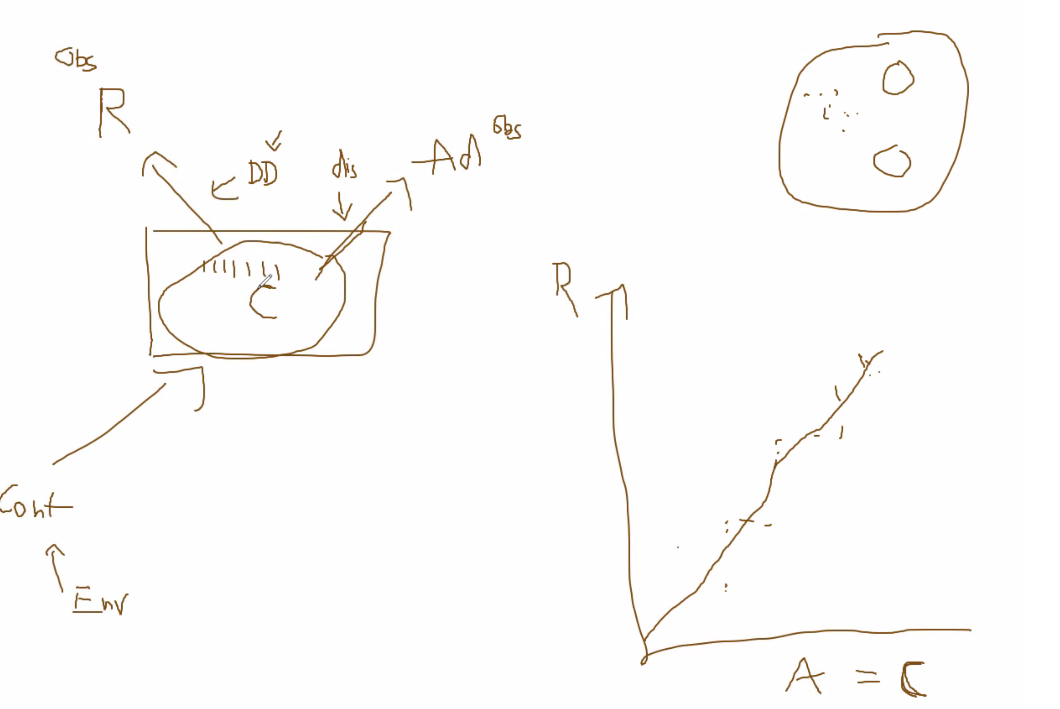
**Bayesian-thinking workshop/group assignment: Density dependence in forests**

***Focus of the project:*** We agreed that it would be interesting to focus on the problem of measurement error in the density of tree adults in the types of observational data typically used in studies of density dependence in forest tree communities. Previous approaches have been based on regression approaches that did not consider this measurement error (Fig. 1). Specifically, we thought it would be interesting to explore whether we can construct adult density as a latent variable, not only using the plot-scale observations, but also regional information on species composition and (continental-scale) variation in habitat and climatic conditions. We also agreed that to assess whether specfic approach(es) work well or not, it would be good to start with simulated data for which true adult densities and strengths of density dependence are known.



*Figure 1: Density dependence estimation approaches used by LaManna et al. (2017, left) and Johnson et al. (2012, right). While the two estimation methods differed slightly, neither considered measurement error in the independent variable, i.e. adult density.*

***Thoughts on the statistical modelling framework:*** Florian explained how the following framework could be provide the basis for the analyses (Fig. 2):



**C**ommunity/ecosystem composition

Observed **A**dult density in the plot

Observed recruitdensity in the (sub)plot

Continental-scale species presence/abundance

Use of proportion of the species in **C**, rather than observed **A**

The strength of density-dependence

Abiotic/environmental variables

*Figure 2: Conceptual framework discussed*

***Questions:*** Maybe We all agreed with Yannek that the following questions would provide a good starting point:

Does a latent variable eliminate errors in observed adult densities?

Is a model with latent variable suitable to detect density dependence (or stabilizing) effects on tree recruitment?

How does sampling design (i.e. spatial arrangement of plots, plot size and plot shape) affect parameter estimates?

What is the relationship between recruitment strategies (tree species traits) and estimates of the latent variable?

***Starting with the latent variable approach:***

I thought it might be useful with trying to set up a simulation of a simplified situation where the latent variable approach should work very well. Then, we can add complexity in steps to see whether the approach keeps doing well. Specifically, I started with a simulation that considered three levels of spatial scale:

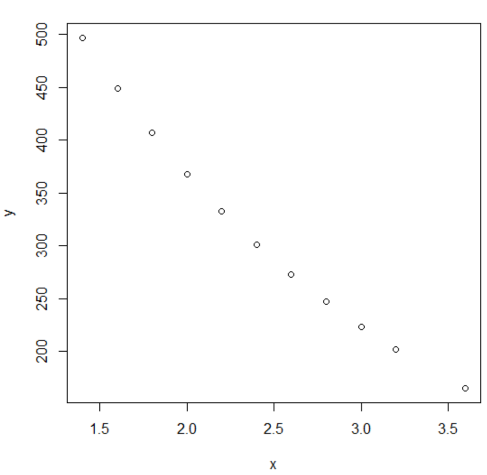
1. **The (continental-scale) species pool: all species that are in the system, their densities supposedly relatively homogeneously distributed (created species abundances using the function of Lisa and Florian)**
2. **The community-scale species pool: this is the actual composition of species that determines the density-dependent effects on seedlings.(For now, I only considered conspecific effects).**
3. **The plot-scale, observed species: these are the observed trees in the plot, but it is assumed that this is only a subset of the trees that are exerting density dependent-effects within the plot.**

*Continental species pool*

*Community pool*

*Observed pool*

1000 seedling – conspecific density effects

To get a clean, true signal of density dependence, I started with the major assumption that in the absence of density-dependent effects, there would be 1000 seedlings in the plot. Hence, any reduction of this number is due to conspecific density dependent effects, exerted by adult species that are present in the community pool. Specifically, there is a perfect negative exponential relationship between number of conspecific trees in the community and seedling number:

Where x is the number of adult conspecific trees

(not whole number as a correction is made to convert to

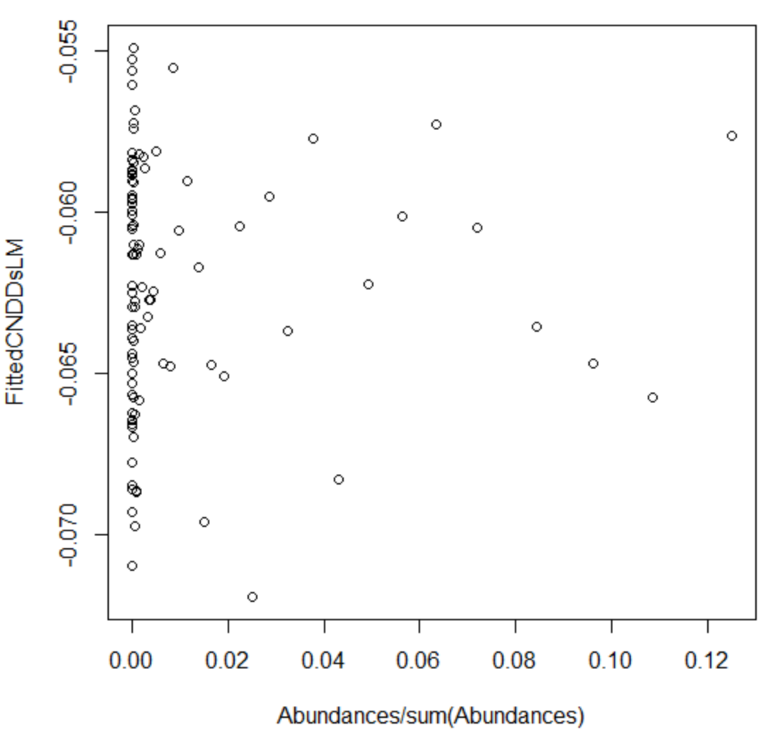
plot density equivalents), and y is the number of surviving

seedlings.

Although each community has the same number of trees, differences in composition emerge because this number is limited, and sampled from the continental species pool. Observed species pools are also sampled, from the community pool. However, we sample a number of replicate plots from the same community, so we can use this variation between replicates to quantify the error in adult densities, and create the latent variable.

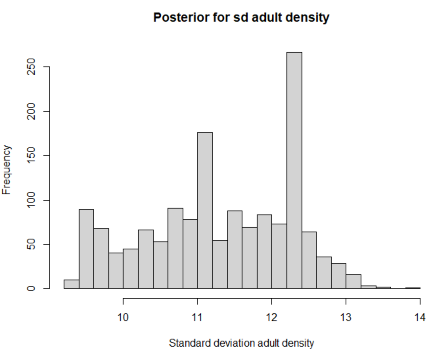
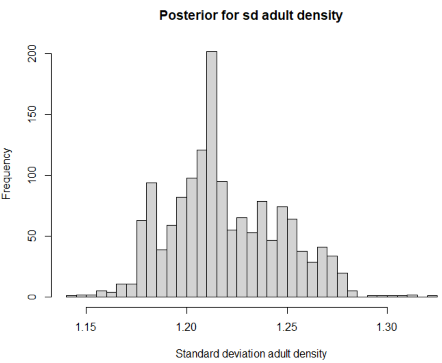
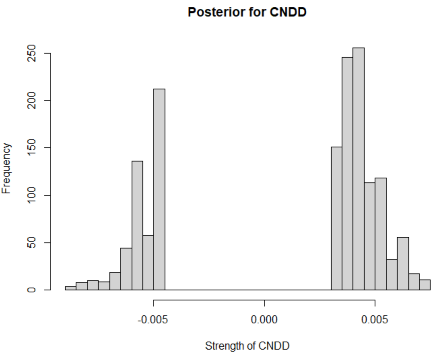
For all the community- and plot-level data, the dataframes are created as matrices where the different species on the rows, and different communities/plots are presented in the columns.

We can then compare the naïve approach (just using the plot-level data as predictors) with the latent variable approach. For the latent variable approach, I used the script from the blogpost that Florian sent, with one adjustment; it seemed in the blogpost that of the replicate measurements, 1 observation was considered the ‘true mean’, from which the likelihood of the other replicates were calculated. As we cannot assign one true observation among the replicates in our dataset, it seems more reasonable to calculated the expected mean from all the replicates (created an extra variable xmeans for the stan file).

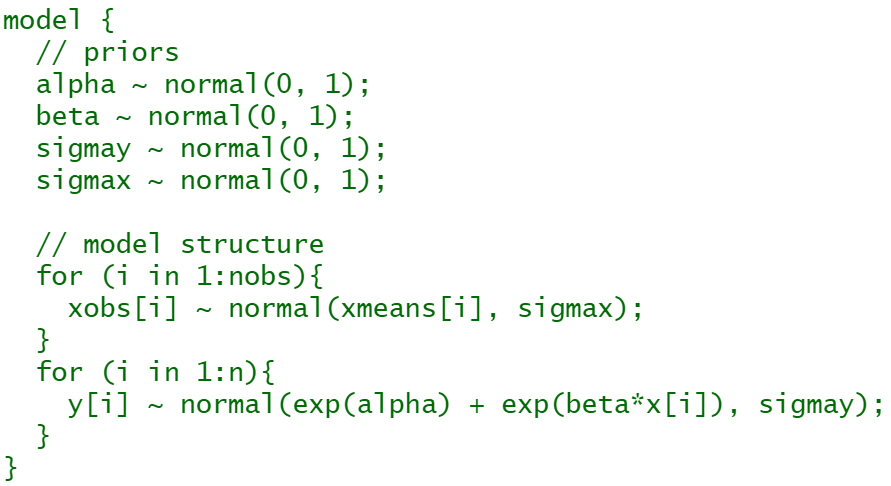
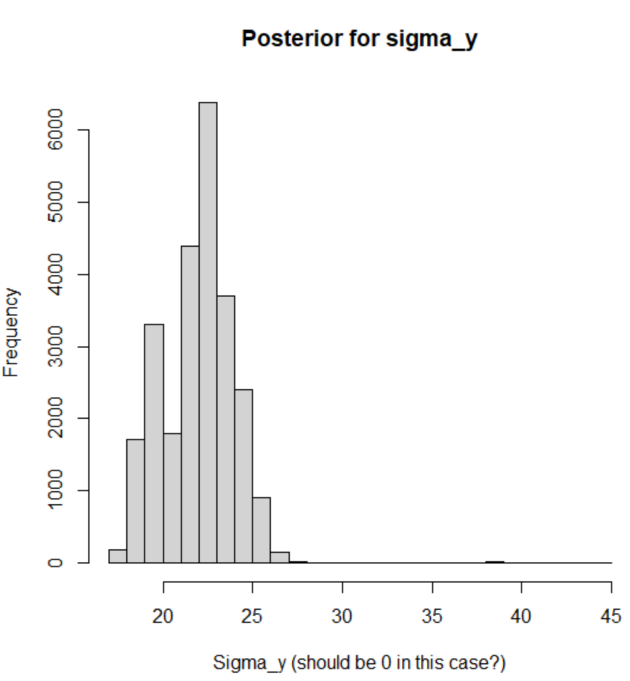
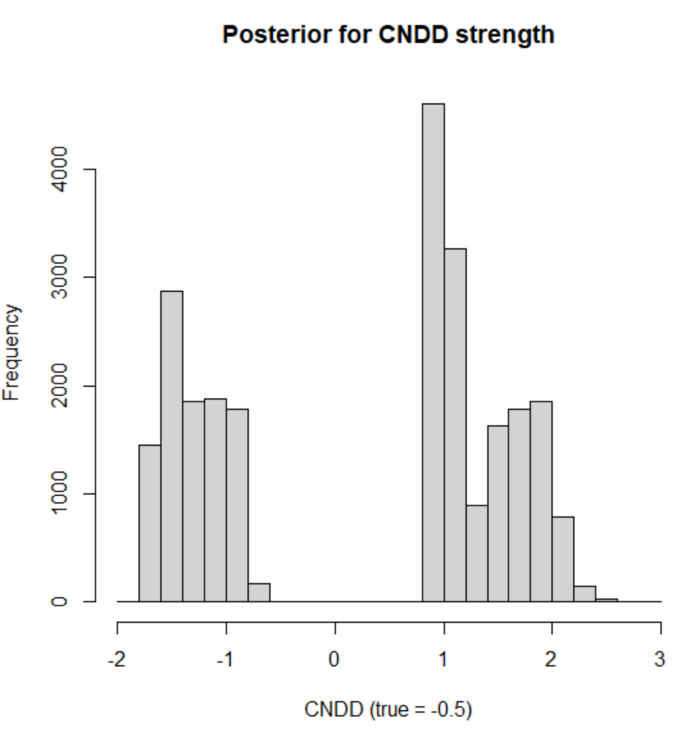
The naïve approach seems to yields CNDD strengths that become more variable for rarer species. Strengths also seem to be underestimating the actual strength (which is -0.5 for all the species shown in the plot on the right), even when a correction is made for the smaller number of trees within the plot as compared to the community (which may or may not be appropriate to do, up for discussion).

Until now, however, I have not been able to get better results with the latent variable approach. The chains do not converge well, and estimates of CNDD seem to diverge into positive and negative extremes. (Note that these results are only for the 1 species (a dominant one).

At least one of the problems seems to be the fitting of the model, as the posterior distribution of the error here (sigma\_y) shows very large values. This is even the case if we use the idealized case with the true predictor data (see below), so something is not going right. Specifically, I think the equivalent to the glm() using earlier in the script would be to use poisson\_log\_glm in the script circled below, but so far I have not been able to get this to work, so I am missing something here; specifically, I do not seem to be able to index y[i] and x[i] in the poisson formulation, so perhaps my idea for the model formulation itself is not correct; I am a bit confused that in the script from the blog post x[] is a parameter, rather than data.



We see the same issues even when using the true predictors as input, which should yield perfect fits, so basically sigma\_y = 0.



Some thoughts on further things to do, also to address research questions 3 and 4:

* The number of plot replicates, the number of species, and the number of adult trees in communities and plots are parameters that can be varied to study specific sampling designs.
* Starting with a 1000 seedlings isolates the effect of density-dependent effects, but seedlings are resulting from seed deposition by adults. This can be included by using the Ricker- and/or Power-offset models, and create a seed input that is dependent on the number of adult trees present.
* Seedlings themselves may be subject to detection error as well, and the number of trees in communities and plots may vary (and be taken from a distribution as well).
* Most of the above characteristics are already incorporated in the functions written by Lisa and Florian, and could be included/switched-on in a stepwise manner to see how robust approaches are to increasing complexity in the simulated study system.