

CoringTreespotters Model notes

Christophe

May 9, 2025

15 April 2025

Quick notes after meeting with Victor where we talked about preliminary steps of building a model to answer the question of how juvenile trees respond to climate change. I would like to approach this as a recruitment capacity of tree saplings and how their growth varies across years and how it's impacted by temperature and growing season length (or maybe not length... more on that below)

Below I will be stating the obvious because I want to make sure I understand the notations and stuff. The steps will be notes and thoughts that I have when I code my model in R.

1 Come up with a model

What are my predictors?

We discussed which of the following would be the most relevant variable that relates to growth:

1. **Growing Degree Days (GDD)** which would be calculated between the budburst and budset date of each individual.
2. **Growing Season Length (GSL)** which could be calculated by:
 - (a) Subtracting budset DOY by budburst DOY or
 - (b) By taking the number of days of each year when the mean (or max/min) was above 5°C (or maybe the nb of consecutive days when the temperature was above 5 or something like that).

Preliminary model

$$\log(w_{i,t}) \sim \text{normal}(X, \sigma) \quad (1)$$

$$\alpha + \alpha_{sp[i]} + \beta_{sp[i]} X_{i,t} + \dots \quad (2)$$

What kind of pooling? Partial pooling since they are all within the same family?

2 Simulate data

I arbitrarily decided to sum up th GDD between 100 and 250 days for now in order to simplify data simulation.

yhat: combine the effects that I generated from the grand mean, the year, species and individual effect deviation effect from the mean, + error returns my yhat. That returns the observed outcome which is composed of multiple nested effects.

3 Set your priors

4 Run model on empirical data

5 Perform retrodictive checks using the model fit to your empirical data

6 Mis

Common Bayesian Notation and other definitions

- θ — Parameters of the model (can be a vector of parameters)
- Standard — How much the data are spread out. It quantifies the variation within a set of measurements
- Standard error — How much the measurements are spread around their means. It is the standard deviations of the means of the different treatments/individuals/whatever. It quantifies the variation in the means from multiple sets of measurements.
- n_{eff} — N effective: Rhat denotes the split Rhat convergence diagnostic, which compares the results from different chains, reaching the value 1 at convergence and being higher than 1 for chains that have not fully mixed. n_{eff} denotes the effective sample size of the Hamiltonian Monte Carlo run, which due to its iterative nature can be less or more than the efficiency of independent draws. Usually $n_{eff} \geq 400$ is sufficient for reliable diagnostics and accurate parameter estimates.
- $mcse$ — $mcse$ stands for Monte Carlo standard error, which is the additional uncertainty due to using stochastic algorithm (not to be confused with the uncertainty presented by the posterior distribution). Monte Carlo standard error is negligible in all the examples in this book.
- y — Observed data
- x — Predictor variables or covariates
- $p(\theta)$ — Prior distribution over parameters
- $p(y | \theta)$ — Likelihood: the probability of the data given parameters
- $p(\theta | y)$ — Posterior distribution: updated beliefs about parameters after observing data
- $p(y)$ — Marginal likelihood or evidence: probability of the data under the model
- $p(y | x, \theta)$ — Likelihood conditional on predictors
- $p(\theta, y)$ — Joint distribution of parameters and data
- μ — Mean (typically of a normal distribution or prior)
- σ — Standard deviation (scale of a distribution)

- τ — Precision (inverse of variance: $\tau = 1/\sigma^2$)
- $\mathcal{N}(\mu, \sigma^2)$ — Normal (Gaussian) distribution with mean μ and variance σ^2
- $\text{Beta}(\alpha, \beta)$ — Beta distribution (common prior for probabilities)
- $\text{Gamma}(\alpha, \beta)$ — Gamma distribution (common prior for positive variables)
- $\text{InvGamma}(\alpha, \beta)$ — Inverse Gamma distribution (common prior for variances)
- $\text{Bernoulli}(p)$ — Bernoulli distribution for binary data
- $\text{Binomial}(n, p)$ — Binomial distribution
- $\text{Poisson}(\lambda)$ — Poisson distribution for count data
- \sim — "Is distributed as" (e.g., $y \sim \mathcal{N}(\mu, \sigma^2)$)
- \propto — Proportional to (used in unnormalized posteriors: $p(\theta | y) \propto p(y | \theta)p(\theta)$)