Background

- T_0: Vegetative priming
- T_1: Floral initiation
- T_2: Pollination
- T_3: Seed maturation (*Pinus* only)

For 2 year reproductive cycles:

1. T0 occurs in early April; buds become dormant by November.
2. T1 occurs in early April; intial stages of sexual cone primordium formation, with differentiation

occuring through the spring of T1. These developing buds are sensitive to temperatures during development, and environmental conditions have an effect on their developmental path through May and June of T1. In unfavorable conditions, buds will abort. By mid-July of T1, vegetative and reproductive buds are differentiated; by October, they are morphologically identifiable. By December, buds have gone dormant.

1. Warm spring temperatures in T2 break dormancy and trigger pollen development. Fertilization and

embryonic development occurs from April through September. In autumn of T2, mature cones open and shed seeds until late November.

Symbols

$$x(t) y(t) \tau \widehat{R}$$

Modeling

So far we've talked a lot about ΔT as a quantity that is correlated with cone production, but as you've seen in our data, the degree to which it actually is a good predictor of cone crop varies with the intervals you choose to analyze; so making a principled choice about what intervals to analyze (start, offset, and duration) is difficult. You might be able to make some arguments about when vegetative priming is happening or when certain species are dormant, or when differentiation occurs, or something like that but *a priori* we don't know the intervals that are relevant here. An effective approach requires that you already know what these intervals should be, which is not very simple to do. And just choosing the intervals which yield the highest correlation opens up a huge can of worms due to look elsewhere effects that are pretty hard to control.

A different issue that is also something to be concerned with is that ΔT does not *explain* the reproductive processes we want to study. So I've wanted for a while to push forward on a cone production model that comes from first principles, and my progress so far is what I'll be telling you about today.

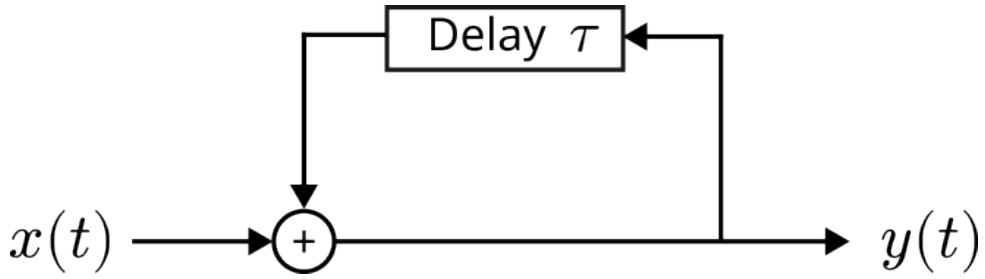
First principles modeling

Energy to produce cones comes from sunlight. It's hard to measure solar flux, but let's make the crude assumption that the temperature is proportional to the energy available from sunlight.

In a simple model that includes ΔT, the energy available to produce some number of cones in a stand is proportional to ΔT at a time in the past[1], but that energy budget is reduced by the amount of energy used to produce cones in the previous crop some time τ_1 ago:

$$n(t) = \alpha \Delta T(t - \tau_0) - \beta n(t - \tau_1)$$

Recurrence (or difference) equations very similar to this have been studied before in the context of signal processing. In that context, there closest analogy is called a comb filter:



Here, the output signal $y(t)$ is composed of an input signal $x(t)$ plus a part of the output signal from τ time ago $y(t - \tau)$:

$$y(t) = x(t) + \alpha y(t - \tau)$$

These systems are usually studied not in the time domain but rather in the frequency domain, and there's a good reason for that: the fact that you're adding a delayed version of a signal to itself means that there will be constructive and destructive interference. So we should naturally expect systems which behave this way to have interesting behavior in the frequency domain.

Let's return our cone crop model and apply the same mathematics used to analyze these circuits:

[1] During the differentiation, or possibly during vegetative priming or some other important time during the reproductive cycle. The fact that it's hard to talk about the meaning of ΔT *at a specific time* is probably an indication here that there's a more appropriate model

Numerical Analysis

General questions

- Why is Pearson's correlation coefficient insufficient?

In general, we don't expect there to be a simple linear relationship between n and T.

- Even the simplest models such as the crude one introduced here aren't linear
- We know that masting events happen. These clearly are nonlinear processes because the yield does not scale linearly with the tempearture or with ΔT.
- Why aren't you just using method X?

This bayesian approach is just one method. You could do X. You could do whatever. This is just one method. I think it has some nice features that make it useful, so that's why I chose it.

How do we know if our analysis is working?

1. Convergence diagnostics
   - Look at sample traces. The MCMC sampler produces a Markov Chain that samples from the posterior distribution. If the chain has converged, i.e. the mean has stabilized, the variance remains constant - then we know we've converged.
   - There are no divergences: a chain doesn't wander off into regions of low probability.
   - Bayesian fraction of missing information (BFMI). The latest gradient-based samplers like what you find in modern probabilistic programming packages use a family of solvers called Hamiltonian Monte Carlo methods to sample from the posterior distribution. These methods use Hamilton's equations of motion to do that - in essence, they're simulating how particles move around in a potential well, and they're generating samples from the posterior distribution from the position of the particles at various timesteps. And it turns out you can look at the energy at each step relative to the overall energy, you can estimate how well a sampler has been exploring the posterior distribution.
   - Similarly you can look at a plot of the energy transitions at each step for your chains, and compare that to the overall energy after fitting.
   - You can also look at something called the Potential Scale Reduction Factor (R-hat), which is basically an analysis of variance of different traces. Essentially you compare the variation in an individual chain to the variation between chains. And if they're the same, you've probably got a good fit. If your chains are in different spots, the between variance is wider than the variance within.
2. Goodness of fit
   - The best way to do this is to compare the output of your model to the data that was used to fit the model. Remember - we're writing down an actual PDF that we think represents the generative processes that we think produced the ΔT and n measurements. So if we draw a bunch of samples from that model, ourdata should look just like more samples from that model. Otherwise we can't really claim that we're modeling the process that generated our data.

     The way that you can do this is using something called Posterior Predictive Checks, which involves sampling from the Posterior Predictive Distribution. The Posterior Predictive Distribution gives the probability of obtaining some new data D_{new} given existing data D. So recall that for a fixed value of our model θ, our data D follows the likelihood distribution:

     $$\mathcal{L}(D|\theta)$$

     and our posterior distribution tells us the probability of our model given our data:

$$\mathcal{P}(\theta|D)$$

However, the true value of our model θ is uncertain, so we should average over the possible values of θ to get the distribution of D_{new} *given* the existing data D.

$$p(D_{\text{new}}|D) = \int \mathcal{L}(D_{\text{new}}|\theta)\mathcal{P}(\theta|D)d\theta$$

Here, D_{new} is hypothetical new data that would be expected, taking into account the posterior uncertainty in the model parameters. Here $\mathcal{P}(\theta|D)$ is the posterior distribution - the probability of the model given the data that we've calculated numerically, and $\mathcal{L}(D_{\text{new}}|\theta)$ is our likelihood - we're drawing values from our likelihood.

So what this does is include the residual uncertainty in our parameters in addition to the stochastic sampling uncertainty that comes from drawing from a distribution.

Modern probabilistic programming packages will give you this for free, so it can actually be really straightforward to carry out this kind of analysis.