Background

Traditionally cone crop prediction modeling has focused on calculating correlations between various predictors and the cone count observed for a stand of trees. Usually, predictor variables are functions of the average temperature in the years preceeding the observed cone crop during the growing season, since that is when the various biological processes involved in reproduction occur.

Kelly et al explored several such predictor variables and compared their correlations across a broad dataset involving multiple plant families and over a long period of time. The predictor variables considered were:

- T1 model: the mean summer temperature in the previous year (T_{n-1}).
- T2 model: the mean summer temperature 2 years previously (T_{n-2}).
- ΔT model: the change in mean summer temperature over the two preceding years ($T_{n-1} - T_{n-2}$).
- 2T model: both mean summer temperature in the previous year (T_{n-1}) and mean summer temperature 2 years previously (T_{n-2}).

Correlations were calculated using linear fits between the predictor variable and log of the annual seedfall c. Kelly et al compared the p-values calculated with the null hypothesis being that no correlation exists, and found that for most plant families the ΔT model produced the lowest p-values, and concluded that ΔT is an ideal cue for seed crop prediction.

Motivation

There are a number of reasons to suspect both alternative models and alternative mathematical approaches would be better motivated than the proposed approach of Kelly et al.

1. **p-values**: Although the p-values calculated are small, any analysis in which many p-values are calculated is bound to have statistically signficant p-values *somewhere*. No effort to control for the look-elsewhere effect was made. Furthermore the null hypothesis - that the models analyzed by Kelly have 0 correlation with seed crop - doesn't capture what is already known about plant reproduction; common sense tells us that temperatues *must* have an effect on reproductive processes because plants can't reproduce in temperatures inhospitable to life. It therefore shouldn't be surprising that there's a correlation between a predictor involving temperature and the seed crop.
2. **Linearity**: There's no reason *a priori* to think that the relationship between any of the predictors put forth by Kelly et al should be linear with $\log(c)$, except that any continuous and differential function can be approximated to first order as a line (Taylor series). No discussion of this choice is made, even though it may be true, although the implications of this implicit assumption mean that the calculations become easier.
3. **Homoskedasticity and normality**: In the supplemental material, Kelly et al argue that empirically the measured seedfall appears to be homoskedastic, i.e. that the variance doesn't change with the number of seeds observed. This is part of a larger implied (but not discussed) argument that the observed log-seedfall $\log(c)$ is a normally distributed random variable with a fixed variance σ^2 and a mean ΔT. In short, homoskedasticity is used as an argument that $\log(c) \sim \mathcal{N}(\Delta T, \sigma)$.

   I suspect that seed production is *not* a homoskedastic process; plants that have more resources available for seed production should have a greater *variation* in seed production. Furthermore the assumption of normality is not justified; the fact that the normal distribution has a nonzero probability on the domain $(-\infty, \infty)$, while $\log(c)$ is only defined (for real values) on the interval $(0, \infty)$.

4. ΔT **as a model**: Intuition tells us that plants that experience freezing conditions for multiple years will not reproduce as much as the same plants that experience ideal growing conditions for multiple years because they simply do not have the same resources available for reproduction. However, two subsequent years of ideal reproductive temperatures may have the same small value of ΔT as two subsequent years of terrible reproductive temperatures. Kelly et al discuss this as an interesting consequence of this model - that as a general rule plant reproduction will be insensitive to changes in global temperatures, as ΔT will remain unaffected by average changes in temperature.
5. **Data Manipulation**:

Mathematics

Likelihood

The number of cones c_i produced by a stand measured on a given day i is Poisson distributed:

$$P(c_i \mid \overline{c}) = \frac{\overline{c}^{c_i} e^{-\overline{c}}}{c_i!} \tag{1}$$

The number of number of cones $\overline{c}$ that we expect to see is given by the energy-conserving equation that we've discussed before,

$$\overline{c}_i = c_0 + \alpha \langle T \rangle_{i-l_0, w_0} + \beta \langle T \rangle_{i-l_1, w_1} - c_{i-l_2} \tag{2}$$

where e.g. $\langle T \rangle_{i-l_k, w_j}$ denotes the moving average of the temperature T over a window of size $2w_j + 1$ days surrounding the day $i - l_k$. Here, α and β are fit parameters which determine the relative importance of each year's sunlight contribution to the stand's energy reserves. c_0 is the initial energy reserves of the stand at the beginning of our observations.

The likelihood of observing the data $\{T_i, c_i\}$ from our dataset is just the product of the probabilities of each observation:

$$P(\{c_i, T_i\} \mid \overline{c}_i) = \prod_i \frac{\overline{c}_i^c e^{-\overline{c}_i}}{c!} \tag{3}$$

where $\overline{c}_i$ is the expected number of cones on day i, given by Equation 2. This is the **likelihood** distribution; it is the probability of observing our data given our model.

Priors

I chose some prior probability distributions based on what I know about cone production. These characterize the epistemic uncertainty about our system:

Parameter	Prior	Unit of measure	Comment
c_0	$\text{Uniform}(0, 1000)$	# of cones	Initial energy reserves (number of cones) at start of dataset; can be between 0-1000 cones
α	$\text{HalfNorm}(10)$	cones/°F	Weakly informative choice of half-normal distribution, since this is probably a small number
β	$\text{HalfNorm}(10)$	cones/°F	Weakly informative choice of half-normal distribution, since this is probably a small number
w_0	$\text{Uniform}(1, 100)$	days	Window size used to calculate the average temperature in the first year. Probably in the range of 1-100 days long
w_1	$\text{Uniform}(1, 100)$	days	Window size used to calculate the average temperature in the second year. Probably in the range of 1-100 days long
l_0	$\text{Uniform}(180, 545)$	days	Lag time of the moving average of the temperature in the first year; constrained to be 0.5 to 1.5 years before the measured crop
l_1	$\text{Uniform}(550, 910)$	days	Lag time of the moving average of the temperature in the second year; constrained to be 1.5 to 2.5 years before the measured crop
l_2	$\text{Uniform}(915, 1275)$	days	Lag time used to get the last cone crop, constrained to be 2.5 to 3.5 years before the measured crop

Posterior

Using the likelihood (Equation 3) and the priors (Table 1), we can construct the **posterior** distribution using Bayes' theorem:

$$P(\overline{c}_i \mid \{c_i, T_i\}) \propto P(\overline{c}_i) P(\{c_i, T_i\} \mid \overline{c}_i) \tag{4}$$

Using MCMC, we can sample from this distribution to get an idea of what it looks like.

MCMC

Next Steps

After some debugging it looks like the sampler is working reasonably well, but it clearly hasn't converged. The Markov chains for the lag and window size in the first year vary wildly, but we have to pay attention to the fact that the coefficient of the first year moving average term *did* converge to zero, which is why the lag and window size were able to vary so erratically - no matter their values, they had no impact on the cone count. In any case, we probably need to reparameterize in order for the model to converge.

If we can get a converged model post-reparameterization, the next thing to do will be to carry out some posterior predictive checks, i.e. generate fake data using these probability distributions to see if it looks like the data we measured. If they look similar, we'll know we've captured the important parts of the generating process that led to these datasets, and we'll actually be able to start connecting these parameter values with what we know about reproductive processes.

Modeling

Assume

$$c_{\text{obs}} \sim P(c_\mu) \tag{5}$$

Consider various models for c_μ:

n-Years Preceeding Model

$$c_{\mu,i} = c_0 + \sum_j \alpha_j \langle T \rangle_{\gamma, i-j} - \beta c_{i-k} \tag{6}$$

Where j runs over a few years preceeding the cone crop in year i. Here, $\langle T \rangle_{\gamma,i-j}$ means an average of the temperature for γ days starting on day $i-j$, and c_0, $\{\alpha_j\}$, β, and γ are fit parameters.

Generally these models are sort of unmotivated in the sense that the number of years included in the sum is arbitrarily chosen, although they are motivated by literature suggesting that the important reproductive processes leading up to cone production occur in either two or three years preceeding the cone crop - some species have a year of reproductive "dormancy", where immature cones remain on the tree for a period of time.

Resource-Accumulation Model (RAM)

$$c_{\mu,i} = c_0 + \underbrace{\alpha \int_0^{t_i} T(t)\,\mathrm{d}t}_{\substack{\text{Photosynthetically} \\ \text{Active Radiation}}} - \int_0^{t_i} c(t)\,\mathrm{d}t \tag{7}$$

Here the resources accumulated by the tree over time are considered: the Photosynthetically Active Radiation (PAR) received each day is approximated as being proportional to the temperature that day; a potentially dubious approximation. The available resources of the stand include all the PAR absorbed since the beginning of the dataset less any spent on cone production.

Other resource expenditure

Leaves, wood, and roots cost a lot of energy. One important nuisance parameter is the energy expenditure on wood/leaf/root growth. We can modify the RAM to include seasonal changes in non-cone resource expenditure:

$$c_{\mu,i} = c_0 + \underbrace{\alpha \int_0^{t_i} T(t)\,\mathrm{d}t}_{\substack{\text{Photosynthetically} \\ \text{Active Radiation}}} - \int_0^{t_i} c(t)\,\mathrm{d}t - \int_0^{t_i} R(t)\,\mathrm{d}t \tag{8}$$

The instantaneous resources available are thus

$$\alpha T(t) - c(t) - R(t) \tag{9}$$

and the change in expected cone crop from year i to year j is

$$\Delta c_{\mu,i\to i+1} = \alpha \int_{t_i}^{t_{i+1}} T(t)\,\mathrm{d}t - \int_{t_i}^{t_{i+1}} c(t)\,\mathrm{d}t - \int_{t_i}^{t_{i+1}} R(t)\,\mathrm{d}t \tag{10}$$

Transformations

Monte Carlo samplers are sensitive the data fed into them; generally they sample efficiently when data is distributed $\sim N(0,1)$.