Background

Traditionally cone crop prediction modeling has focused on calculating correlations between various predictors and the annual 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

Disadvantages of correlations

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 temperatures *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 valid, although the implications of this 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 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 $\log(c) \sim \mathcal{N}(\Delta T, \sigma)$.

   But is seed production really a homoskedastic process? Naively 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 by the fact that the normal distribution has a nonzero probability on the support $(-\infty, \infty)$, while $\log(c)$ is only defined (for real values) on the interval $(0, \infty)$, meaning that *any* normally-distributed predictor can't possibly explain the range of seed crops observed in nature. Most importantly, log-transforming the seedfall is certainly unjustified for the simple fact that this approach cannot be used to correlate with years in which 0 seeds are observed ($\log(0) = -\infty$). These data points are arguably the *most* important observations in species that exhibit masting because in non-mast years, few to no seeds are often observed.

4. **Data Preprocessing**: As a result of the log-transformation that is carried out on the observed seedfall, years in which no seeds are observed cannot be used in the correlation without utterly dominating the linear fit used to produce the correlation. Kelly et al directly modify these observations before the log transformation, replacing them with values that are half the smallest nonzero value. Similar procedures are carried out elsewhere in the literature, but in all cases this approach is mathematically unjustified; any valid model of cone production must be able to predict 0-seedfall years.

5. 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 because ΔT is unaffected by average changes in temperature.

The Bayesian Approach

Instead of carrying out linear correlations, let's focus on a different approach. Besides the reasons I've previously discussed for trying something different than Kelly et al, there are also other reasons why one might want to take a Bayesian approach. Here's a good discussion about this if you want to learn more.

Here's the general idea:

1. We start by writing down a probability distribution which describes the probability of observing an individual data point c_i given some model. I propose that the probability is given by a Poisson distribution:

   $$P(c_i \mid c_{\mu,i}) = \frac{c_{\mu,i}^{c_i} e^{-c_{\mu,i}}}{c_i!} \tag{1}$$

   where c_i is the number of cones observed for the stand on day i, and $c_{\mu,i}$ is the Poisson rate parameter - the expected number of cones on day i.

   The Poisson distribution is a *counting distribution*, a distribution that gives the probability of a number of events happening in a given amount of time. It has some nice properties: it's discrete (i.e. it is only defined for integer cone counts) and is defined on the support $[0, \infty)$, as is required by our data.

2. From the probability of observing an individual day's cone crop c_i, we can write down the probability of observing the cone crop for all N days (the entire dataset) by taking the product of all the individual probabilities:

   $$P(\{c_i\} \mid \{c_{\mu,i}\}) = \prod_{i=0}^{N} \frac{c_{\mu,i}^{c_i} e^{-c_{\mu,i}}}{c_i!} \tag{2}$$

   This is the *likelihood*; in the literature it is usually written as

   $$\mathcal{L}(D \mid \theta) \tag{3}$$

   where for us the data $D = \{c_i\}$ and the model $\theta = \{c_{\mu,i}\}$.

3. From here, we can write down an expression for the *posterior* probability distribution $\mathcal{P}$, which is a distribution over our model parameters conditioned on our data. We make use of Bayes's theorem, and up to a normalization constant

   $$\mathcal{P}(\theta \mid D) \propto \mathcal{L}(D \mid \theta) P(\theta) \tag{4}$$

   where $P(\theta)$ is a prior distribution; this distribution characterizes the epistemic uncertainty in our model *prior* to observing any data.

In the Bayesian approach, getting the posterior probability distribution is the entire goal. We could spend a long time talking about how this differs from frequentist approaches which usually focus on metrics related to the likelihood (for example, least squares fitting is equivalent to maximum likelihood estimation with the implicit and extremely restrictive assumption that there is no uncertainty in the independent variable, and normally distributed uncertainty in the observed dependent variable), but for now I'm going to just point to Jake VanderPlas's blog and this talk by Chris Fonnesbeck.

As you'll see with models discussed below, usually the posterior can only be written down in a nice closed form for certain special cases where the likelihood and priors are *conjugate*. In that case you can actually calculate the posterior by hand, but in most cases you really need to turn to computers to sample from the posterior distribution numerically, and that's what we'll do here.

Modeling

This part focuses on coming up with reasonable expressions of the rate parameter $c_{\mu,i}$. Fundamentally I think that the expected number of cones should be determined entirely by what resources are available to produce them.

Three years preceeding model

This model assumes that the energy available to produce cones is mostly gathered from average temperatures in specific windows in the three years preceeding the cone crop, with no contribution from years before that. As a proxy for Photosynthetically Active Radiation (PAR) we instead use temperature (since that's the data we have), and make the assumption that the two are proportional. Under that assumption we can write down the expected number of cones:

$$c_{\mu,i} = c_0 + \alpha_0 \langle T \rangle_{i-l_0,w_0} + \alpha_1 \langle T \rangle_{i-l_1,w_1} + \alpha_2 \langle T \rangle_{i-l_2,w_2} - c_{i-l_3} \tag{5}$$

where $\langle T \rangle_{i-l_k,w_k}$ denotes the moving average of the temperature T over a window of size $2w_k + 1$ days surrounding the day $i - l_k$. Here, the α_k 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, c_{i-l_3} is the lagged cone count from l_3 days in the past.

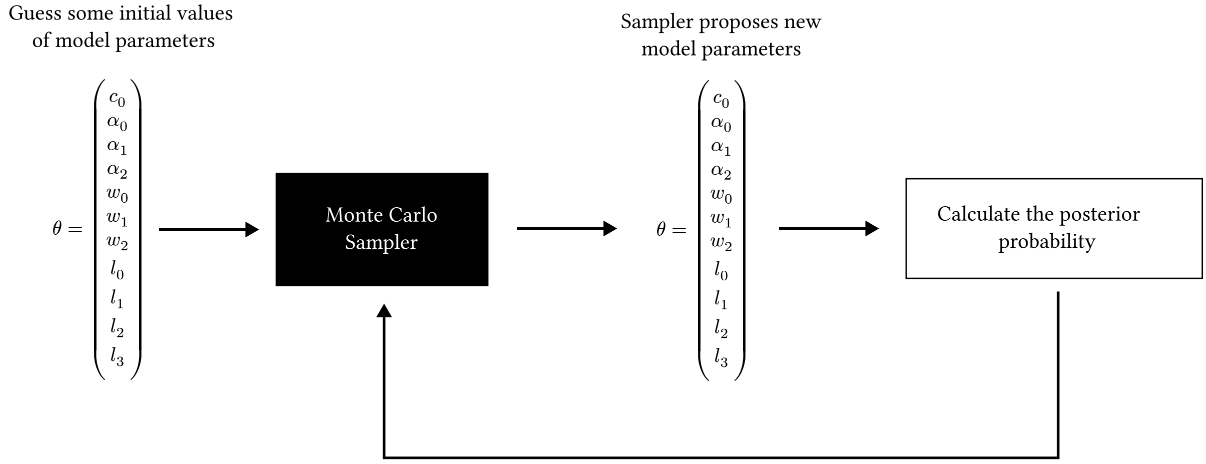
Priors

I chose some prior probability distributions based on what I know about cone production. These characterize the epistemic uncertainty about our model parameters prior to observing any data:

Parameter	Prior	Unit of measure	Comment
c_0	$\text{HalfNorm}(10)$	# of cones	Initial energy reserves at start of dataset
α_0	$\text{HalfNorm}(10)$	cones/°K	Weakly informative choice of half-normal distribution. This is probably a small number
α_1	$\text{HalfNorm}(10)$	cones/°K	Weakly informative choice of half-normal distribution. This is probably a small number
α_2	$\text{HalfNorm}(10)$	cones/°K	Weakly informative choice of half-normal distribution. 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
w_2	$\text{Uniform}(1, 100)$	days	Window size used to calculate the average temperature in the second year. Probably in the range of 1-100 days long
l_0	$\text{Uniform}(185, 545)$	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	$\text{Uniform}(545, 910)$	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	$\text{Uniform}(910, 1275)$	days	Lag time of the moving average of the temperature in the second year; constrained to be 2.5 to 3.5 years before the measured crop
l_3	$\text{Uniform}(910, 1275)$	days	Lag time used to get the last cone crop, constrained to be 2.5 to 3.5 years before the measured crop

Posterior

I used Markov Chain Monte Carlo to sample from the posterior distribution using a software package called emcee. You can think of this like a black box which proposes values of the model parameters $\theta = (c_0, \alpha_0, \alpha_1, \alpha_2, w_0, w_1, w_2, l_0, l_1, l_2, l_3)$; I wrote code that takes those proposed values and uses them to compute the prior and likelihood, which are then multiplied to get a posterior probability, which I then pass back to the black box. The `emcee` sampler then uses this posterior probability to generate new proposed values of θ, and the process repeats many times. After an initial burn-in period, if the probabilities are defined correctly and the sampler is able to efficiently explore the parameter space, the sampler will eventually converge to certain values of θ, and due to some clever underlying math we can guarantee that the samples being generated will match the posterior distribution.



Guess some initial values of model parameters
$\theta = \begin{pmatrix} c_0 \\ \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ w_0 \\ w_1 \\ w_2 \\ l_0 \\ l_1 \\ l_2 \\ l_3 \end{pmatrix}$
Monte Carlo Sampler
Sampler proposes new model parameters
$\theta = \begin{pmatrix} c_0 \\ \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ w_0 \\ w_1 \\ w_2 \\ l_0 \\ l_1 \\ l_2 \\ l_3 \end{pmatrix}$
Calculate the posterior probability

Figure 1: The MCMC sampler takes in an initial θ, and proposes new values of θ. For each proposal, we calculate the posterior probability of the proposed value, and pass it back to the sampler; it uses this value to generate new proposed values of θ. After some point, the new proposed values will converge to posterior probability distribution.

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

$$c_{\text{obs}} \sim P(c_\mu) \tag{6}$$

Consider various models for c_μ:

n-Years Preceeding Model

$$c_{\mu,i} = c_0 + \sum_j \alpha_j \langle T \rangle_{\gamma,i-j} - \beta c_{i-k} \tag{7}$$

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)

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

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:

$$c_{\mu,i} = c_0 + \underbrace{\alpha \int_0^{t_i} T(t)\,\mathrm{d}t}_{\substack{\text{Photosynthetically} \\ \text{Active Radiation}}} - \int_0^{t_i} c(t)\,\mathrm{d}t - \int_0^{t_i} R(t)\,\mathrm{d}t \tag{9}$$

The instantaneous resources available are thus

$$\alpha T(t) - c(t) - R(t) \tag{10}$$

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

$$\Delta c_{\mu,i\to i+1} = \alpha \int_{t_i}^{t_{i+1}} T(t)\,\mathrm{d}t - \int_{t_i}^{t_{i+1}} c(t)\,\mathrm{d}t - \int_{t_i}^{t_{i+1}} R(t)\,\mathrm{d}t \tag{11}$$

Transformations

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

Symbols

$$\theta = \begin{pmatrix} c_0 \\ \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ w_0 \\ w_1 \\ w_2 \\ l_0 \\ l_1 \\ l_2 \\ l_3 \end{pmatrix} \tag{12}$$