West Virginia White

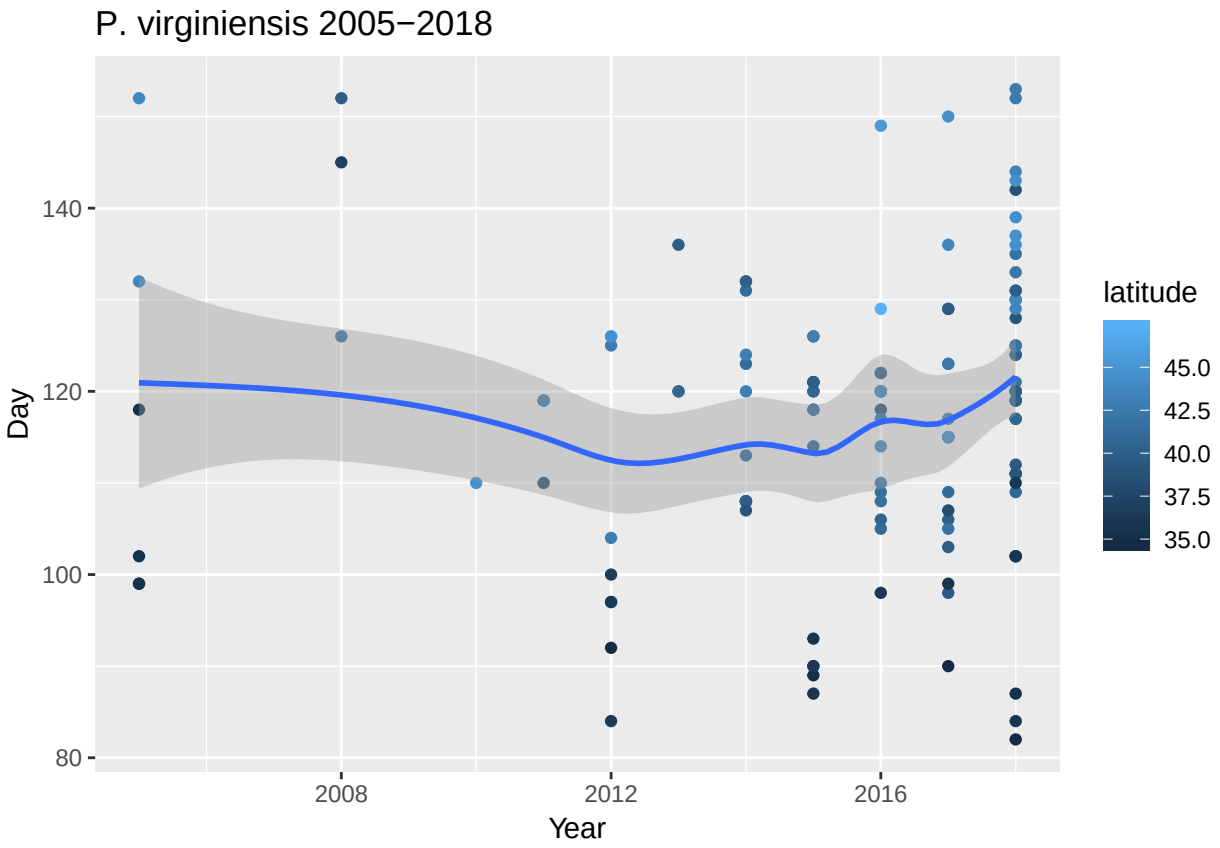
Jeff Oliver

14 May, 2019

Preliminary analyses

"Preliminary"" does not do it justice. This is very, very back of the envelope. Data are from iNaturalist observations downloaded on 2 May 2019.

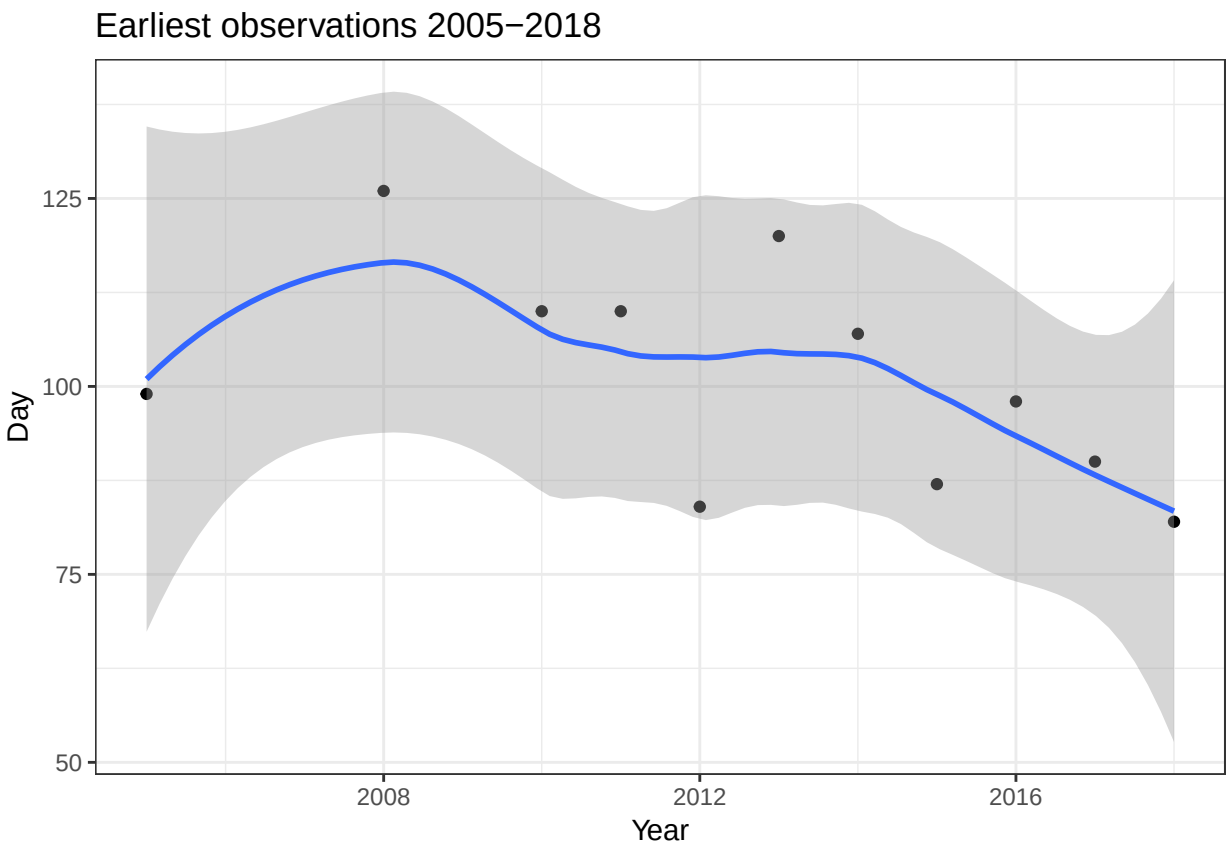
Exluding a single observation from 1997 and observations from 2019, there are 140 observations. If we plot these by year and day of year, it is a bit tough to see much of a trend:



Note that observations from lower latitudes (darker points) are generally towards the bottom of the plot, and observations from more northern latitudes (lighter points) are nearer the top of the plot. No big surprise there.

Earliest emergences

If we only consider the earliest emergences,



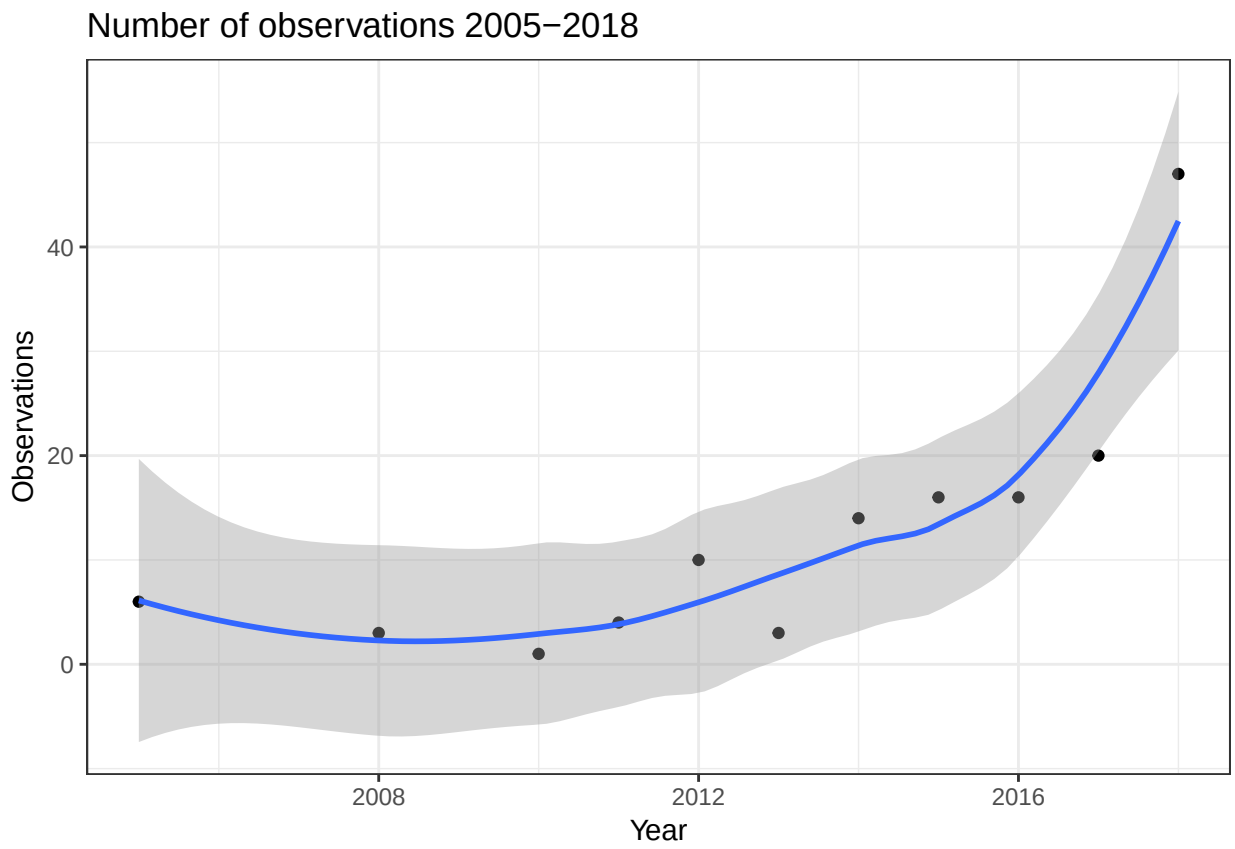
There *is* a trend towards earlier emergences in more recent years, so let's consider a very crude linear model:

$$Earliest\ \ observation\ \ day_i = Year_i + \epsilon_i$$

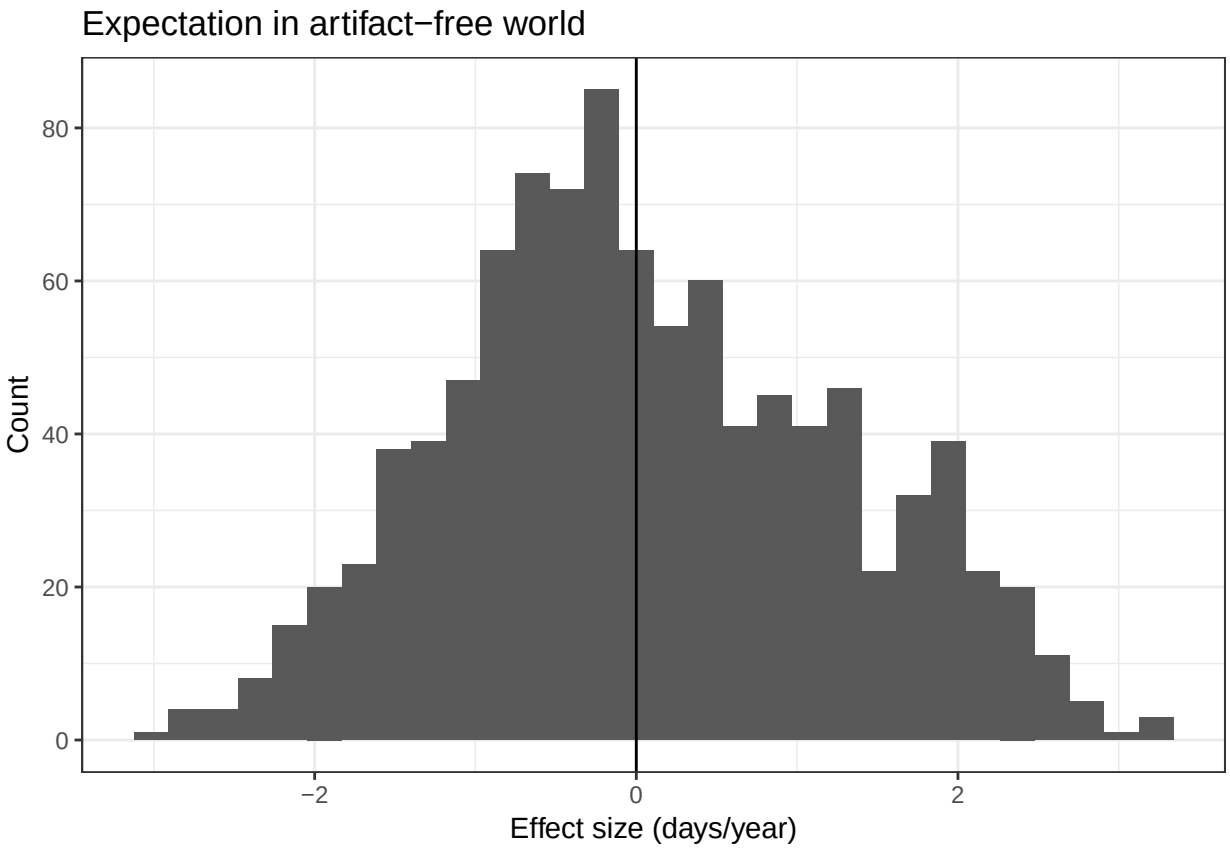
There is a very slight effect of year on earliest observation date, with the first earliest observation occuring 1.96 days earlier each year. However, this effect is only marginally significant ($p = 0.0963$).

But...

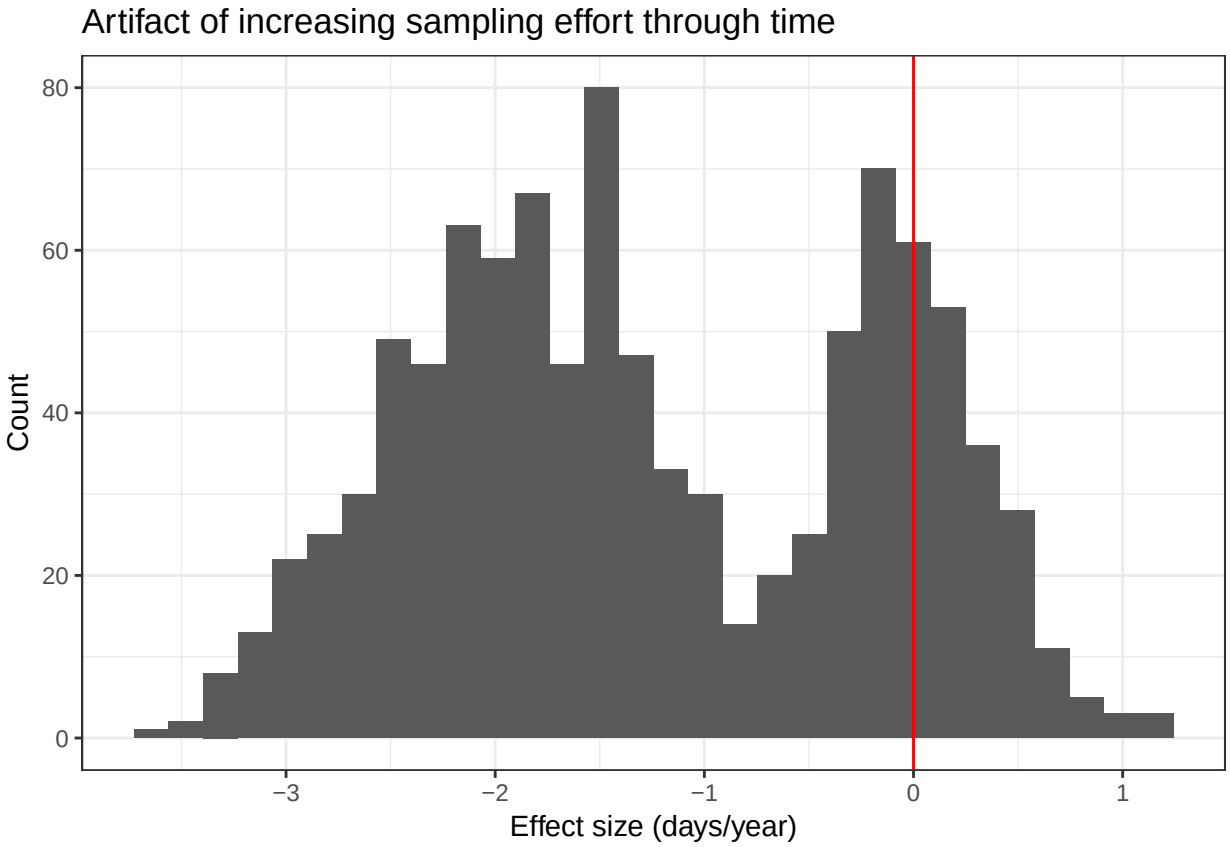
Given the increase in butterfly watching, this change in observations could be entirely due to sampling artifacts than biological reality. Consider the number of observations of *P. virginiensis* through time:



Pretty clearly increasing. So this means that by chance, recent years are more likely to "catch" earlier observations, just because there are more opportunities. To see this in action, consider a thought experiment where we make up data. Well, bootstrapping data, but it's nearly the same thing. If we create a data set that mimics the observation efforts for the observed data (i.e. 6 observations for 2005, 47 observations for 2018, etc.), but instead of actual observations, sample only from the most recent year of observations (2018). We then use those data to run the linear regression again and see if there is an effect. Ideally, if there is *no* artifact of sampling, we should see, on average, no effect of year on earliest observation (this is because, for these data, *all* days of observation are being drawn from the "real" data for 2018 alone). Repeating this process 1000 times should result, on average, of an effect size of 0:



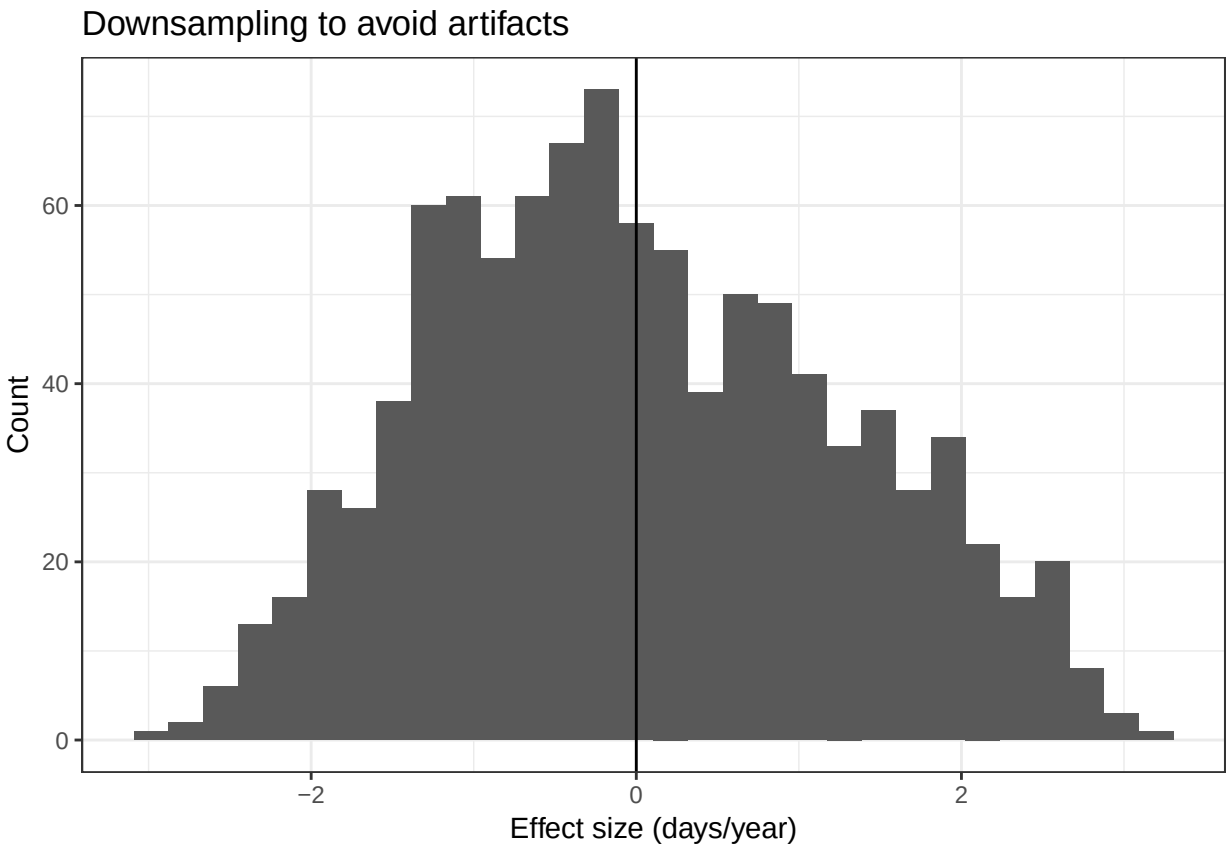
However, when we do the bootstrapping experiment, it looks like there is considerable potential for an artifact:



And the mean effect size is observations are getting earlier by 1.23 days per year. We know this is an artifact because all the data are based on 2018.

Back to the bootstrap

However, we can use bootstrapping to down-sample observations to make effort across years consistent. That is, for each year, we randomly sample a subset of observations so we only have a certain number of observations per year. For that "certain number", we'll use the number of observations from 2005 (n = 6), which is the fewest observations in a single year for these data. Let's test this first by doing the same process we ran before, basing everying on data from 2018 alone, but now only drawing 6 samples for each year. Ideally, we should see no effect of year on earliest observation (i.e. an effect size of 0).



Woo-hoo! So now we have a way to avoid artifacts due to variation in effort. Let's try it for real, downsampling each years' data to only 6 per year.