West Virginia White

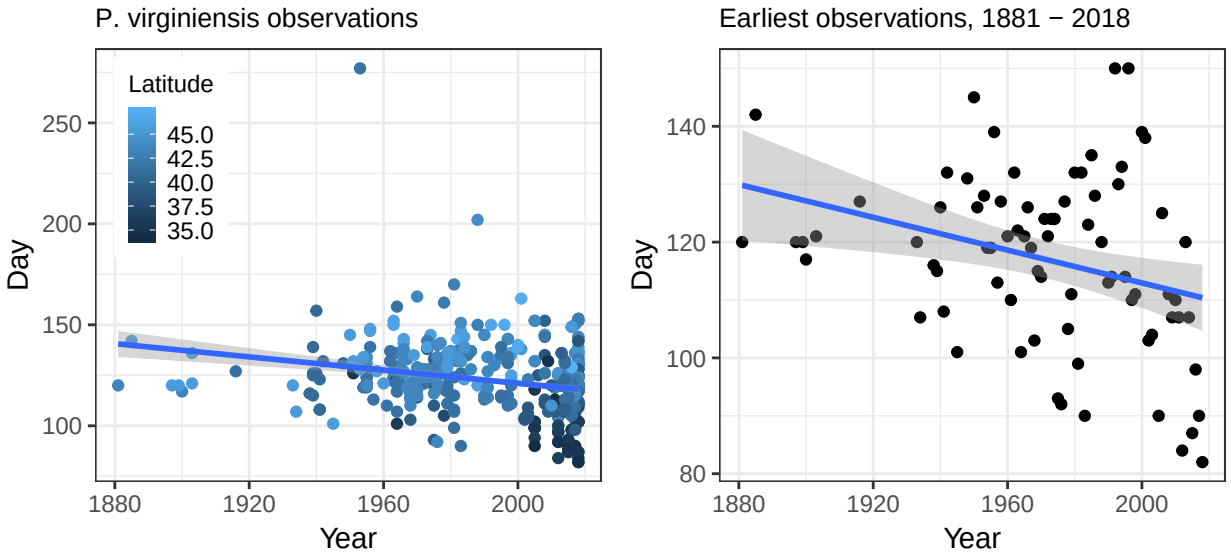
Jeff Oliver

22 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 and GBIF data downloaded on 17 May 2019.

After dropping some unrealistic GBIF observations (some from the Indian Ocean, some from January 1) and those from 2019, there are 487 observations. If we plot these by year and day of year, it looks like there is a shift to earlier emergence times through time.



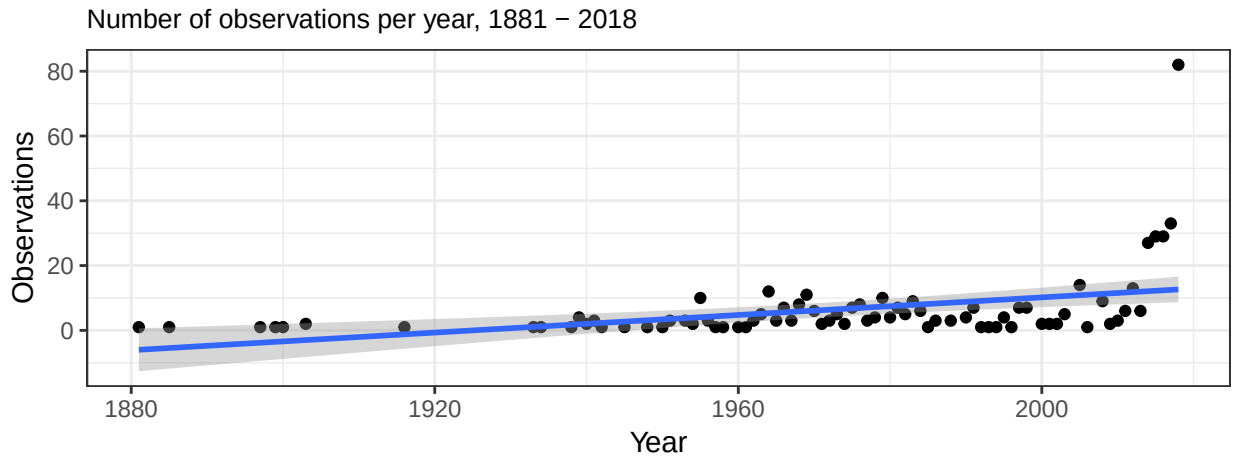
Note that observations from lower latitudes (darker points) are generally towards the bottom of the plot, and observations from more northern latitudes (lighter points) are nearer the top of the plot. No big surprise there. There *is* a trend towards earlier emergences in more recent years, so let's consider a very crude linear model:

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

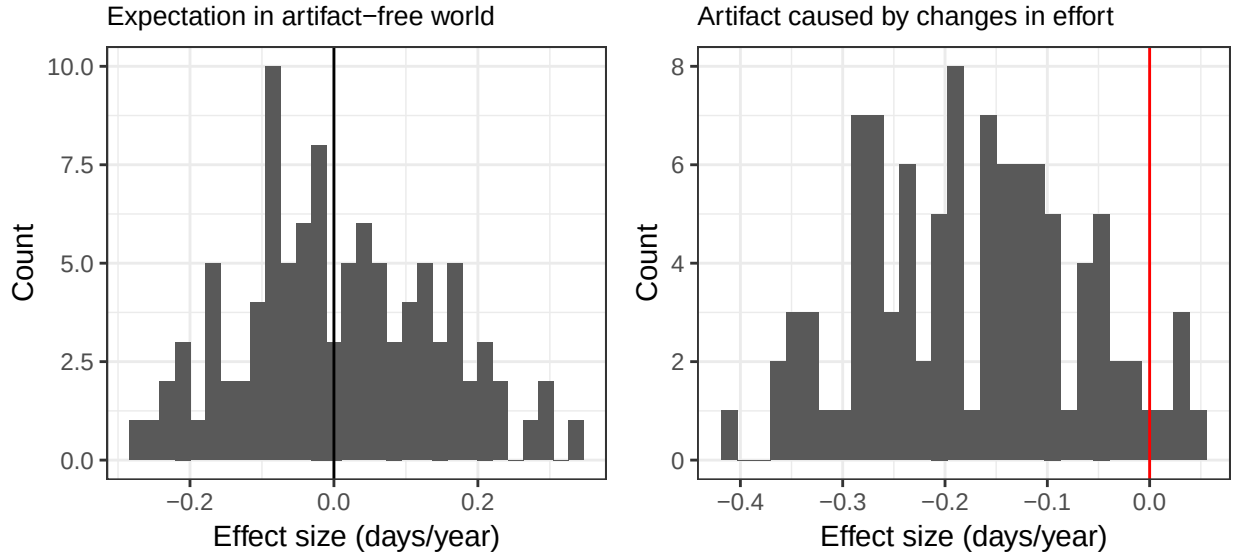
There is an effect of year on earliest observation date, with the first earliest observation occuring 0.14 days earlier each year ($p = 0.0059$). Cool!

But...

Given the increase in butterfly watching, this change in observations could be entirely due to sampling artifacts rather 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, we bootstrap 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. 1 in 1881, 2 in 1974, 82 in 2018, etc.), but instead of actual observations, sample *only* from the most recent complete 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).

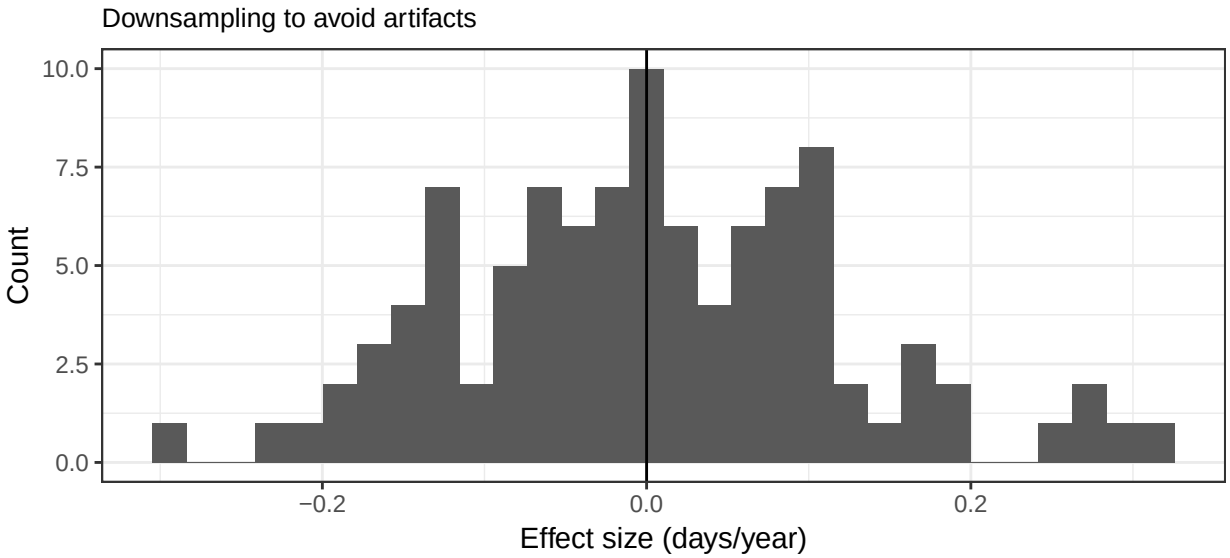


Repeating this process 100 times should result, on average, of an effect size of 0 (left panel). However, when we do the bootstrapping experiment, we see considerable potential for an artifact (right panel). The mean effect size from the right is -0.17, which we would take to mean that the first observation is getting earlier by 0.17 days per year. We know this is an artifact because all the data are based on 2018.

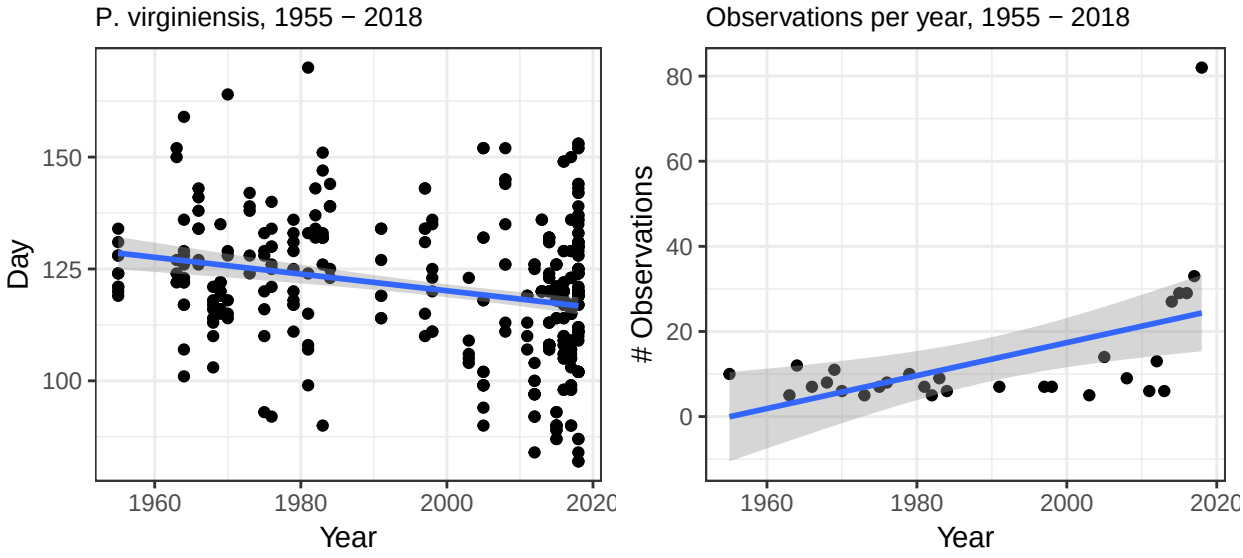
Back to the bootstrap

However, we can use bootstrapping and 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 require a minimum of 5 observations per year. Let's

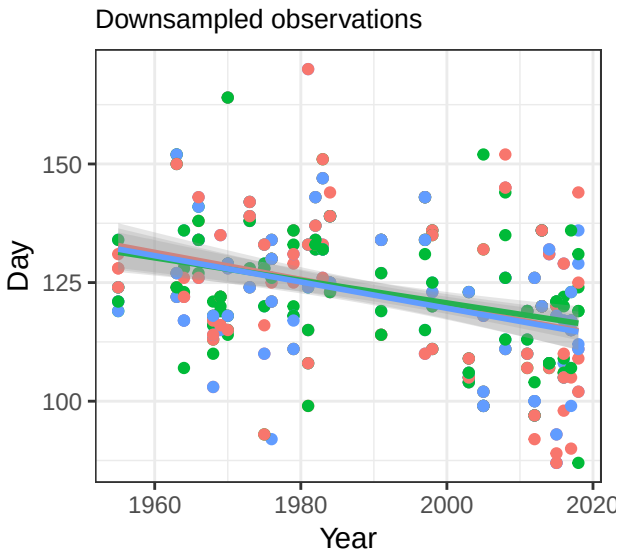
test this first by doing the same process we ran before, basing everying on data from 2018 alone, but now only drawing 5 samples for each year. Ideally, we should see no effect of year on earliest observation (i.e. an effect size of 0).

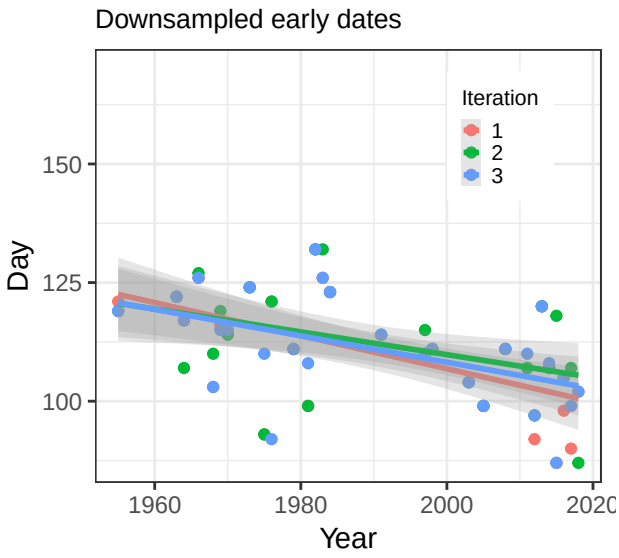


Woo-hoo! So now we have a way to avoid artifacts due to variation in effort. Let's try it for real, downsampling each years' data to only 5 per year. Before we try that, what effect does this restriction of 5 observations per year have on the size of our data set? We had 487 observations, but if we restrict it to only those years with at least 5 observations, we have 390 total observations, spanning 1955 through 2018. Taking a look at these data:



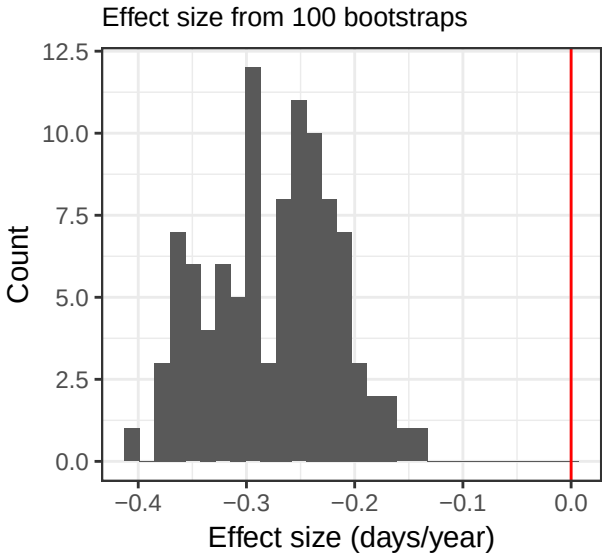
There is still an increase in number of observations per year (right panel), so we need apply the downsampling approach to avoid the artifact described above. Downsampling to include only 5 samples from each year is going to vary each time we do a bootstrapping event. To see this in action, the plots below show three iterations, with each iteration a different color.

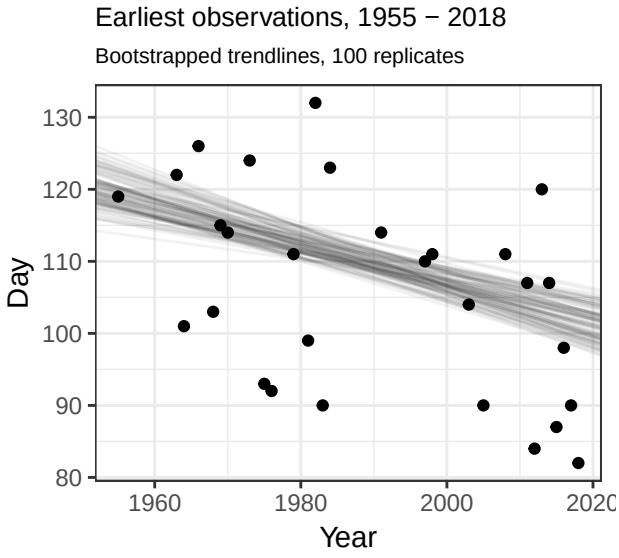




Moment of truth

Now we actually do it, running 100 bootstrap replicates, sampling only 5 observations in each year. We then pull out the minimum value for each year (the first observation of each year) and test for a change over time. If *P. virginiensis* is emerging earlier, we should see a negative trend over time.

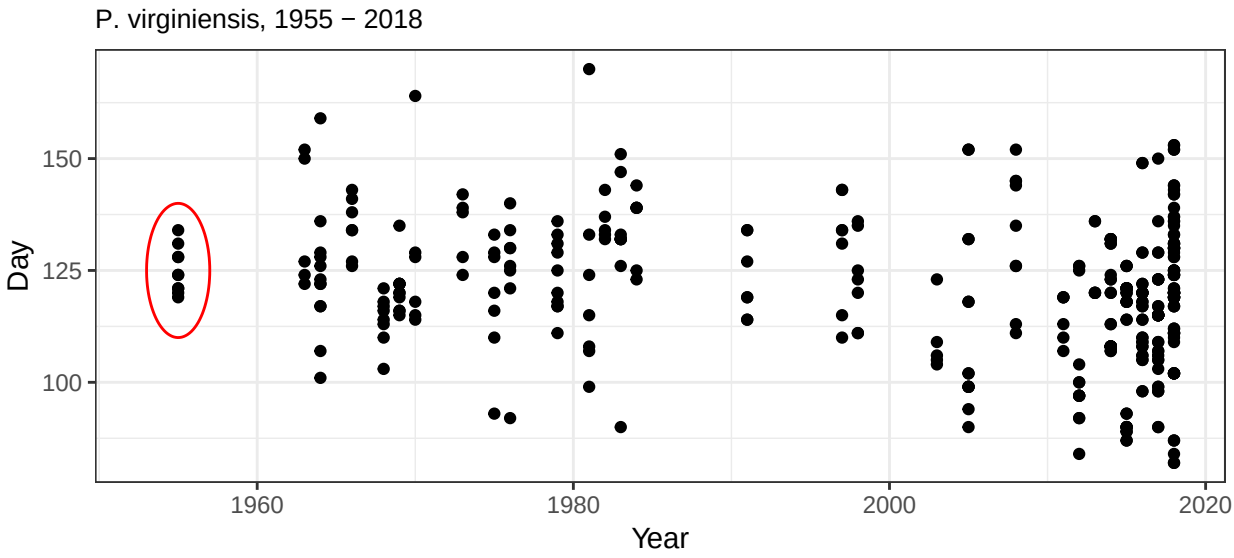




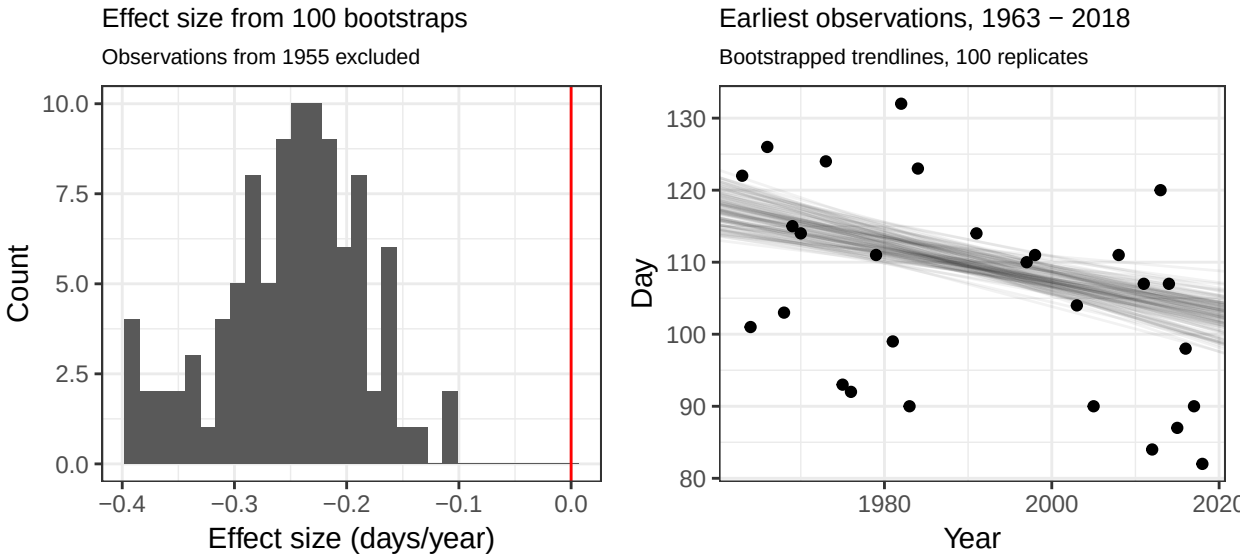
Which is indeed what we see!

From the bootstrap replicates, we see emergences are getting earlier by 0.27 days per year ($p < 0.01$).

But I'm still somewhat skeptical because of the observations in 1955.



My apprehension with 1955 observations is that they are "anchoring" the regressions due to the fairly low variance observed (the observations for 1955 only span 15 days). So I went ahead and dropped all observations for 1955 and ran the bootstrap experiment again.



So even when controlling for my paranoia, *P. virginiensis* emergence dates are getting earlier through time. Excluding the 1955 data, emergences are still getting earlier by 0.25 days per year ($p < 0.01$).

Next steps

So, *P. virginiensis* appears to be coming out earlier. What about *P.oleracea*? Don't know. And *P. virginiensis*'s hosts? Don't know that either. I'm especially keen to see if the slopes of *P. virginiensis* and its host plants are the same. I think that will be especially relevant to address those questions raised by the the heat-addled Vermonter.