Will Pearse and Davies method work for my data?

Susannah Tysor

September 22, 2017

What am I even doing

I want to predict when and for how long lodgepole pine will release pollen and be receptive to pollen. So I need to build a model for pollen shed and a model for cone receptivity. The models will be similar to one another (and identical in form). For today, it doesn't matter which one I'm talking about. I'll just use "phenologically active" or "flowering" to refer to either of those phenological events. For more background on why I'm doing this and some early modeling attempts, see my last lab meeting presentation (**click here**).

I have 2 main sources of data - event data and survey data. The data was collected in seed orchards (representing different provenances) at 7 different sites in BC. It's sort of a bad common garden design.

- Event data: The start date and the end date for a clone's pollen shed and cone receptivity in an orchard.
- Survey data: "Active", "not active", or "not recorded" pollen shed and cone receptivity data for individual trees in an orchard recorded about every other weekday over a period of about 2 weeks. At least one survey period is very short (one day) and others are three weeks long. The survey data has *end censoring* and *interval censoring.*
    - end censoring: The survey period may not contain the start and end dates of an individual
    - interval censoring: The start or end date may occur *between* survey dates.

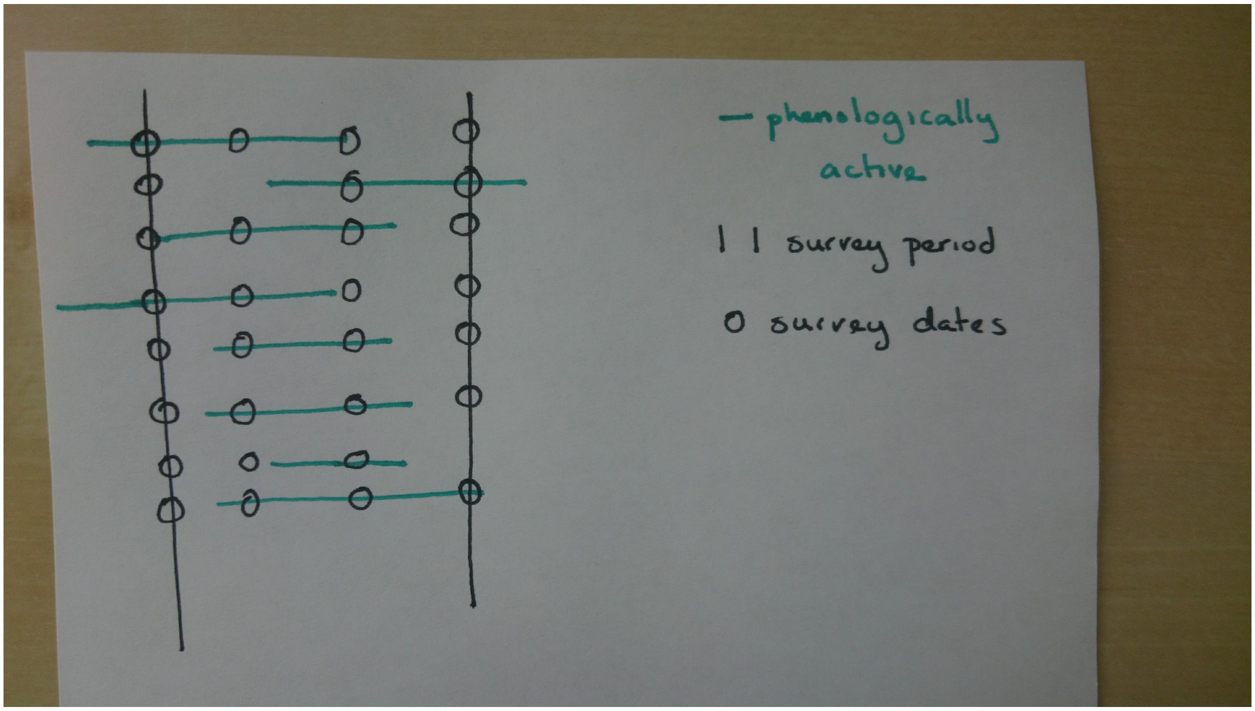


Figure 1: Survey data illustration

I need to use both kinds of data in my model and be able to compare start and end dates between orchards that used different data collection methods. I was thinking about converting the survey data to event data or vice versa and trying to figure out how to do either.

Sally recommended I look into converting the survey data to event data, estimating the start and end dates for the survey data based on a talk she heard Jonathan Davies give in Victoria this summer. Davies and his collaborators have adapted the use of a method for estimating the extinction date of a species to estimating phenological event dates from herbarium specimens.

[1] described in Roberts & Solow (which is from a method described in a 1980 Biometrika paper by Cooke)

Davies and his collaborator Will Pearse generously shared their in-prep manuscript and the part of their code where they'd implemented the method. Here's the crucial bit from from Roberts & Solow:

We applied an optimal linear estimation method based on the sighting record of the dodo to determine when the bird finally became extinct. Let T1 > T2 > . . . > Tk be the k most recent sighting times of a species, ordered from most recent to least recent..In this context, optimal linear estimation is based on the remarkable result that the joint distribution of the k most recent sighting times has (at least roughly) the same 'Weibull form', regardless of the parent distribution of the complete sighting record.

To be honest, I don't entirely understand this "remarkable result" and I couldn't immediately determine how end or interval censored data would affect this method. Colin recommended simulations. So. . . .

I made some very simple fake data ("Active" dates over a period of 10 days, interval or end censored) and used the Cooke/Roberts&Solow/Davies *et al* method to estimate a "start date."

It turns out this method is totally inappropriate for my data because the estimate is a function of the interval between observations and it cannot accept negative observations, only positive.

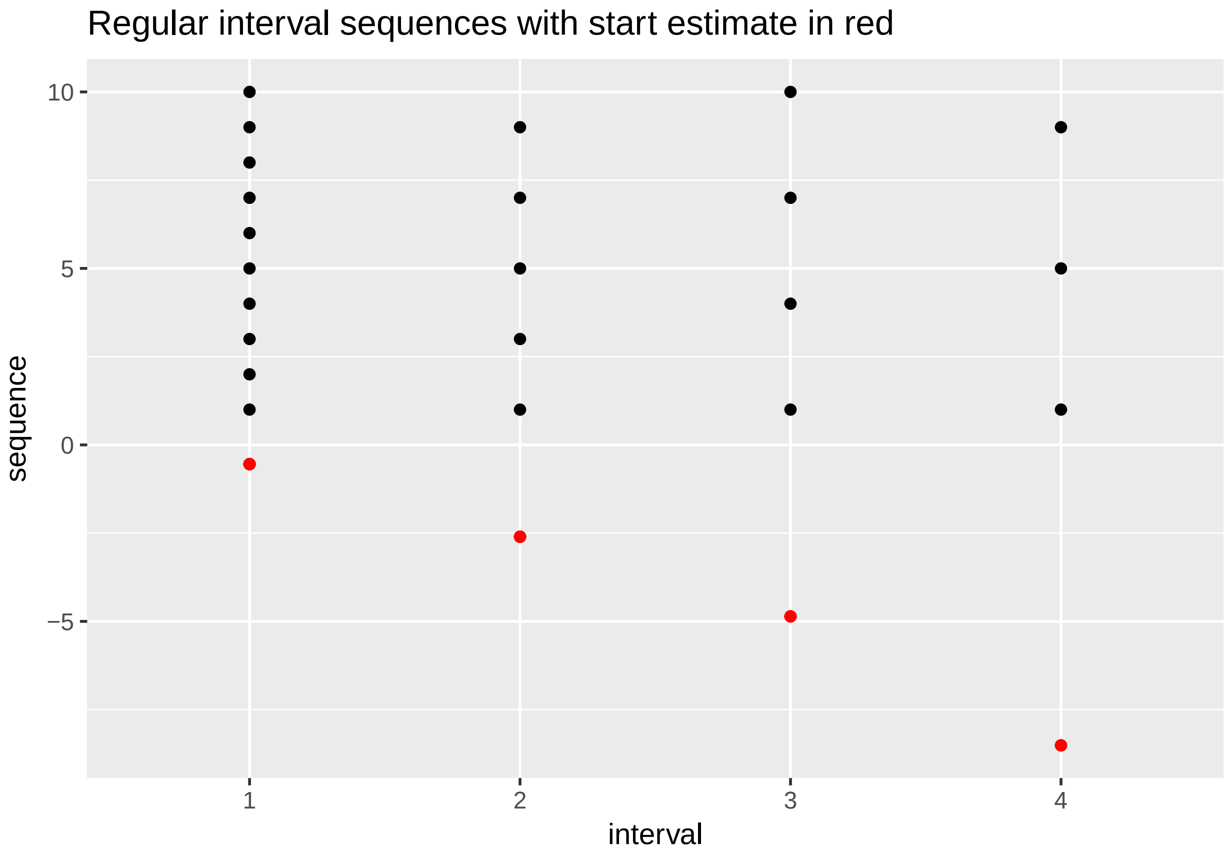
The interval of data collection affects the estimate a lot, but end censoring doesn't. I need it to respond to end censoring but not interval censoring.

Interval censoring

How sensitive is the method described by Pearse and Davies to interval censoring?

I created 4 sequences representing the same flowering period for the same individual. The only difference is that the data is "collected"" at 1, 2, 3, or 4 day intervals. All contain the same initial day and are sampled over the same 10 days. The first sequence (interval 1) represents the full phenological period.

The start day is always estimated to be earlier than the first measurement by more than the length of the sampling interval. **The start day estimate is highly sensitive to interval censoring.** In fact, the estimate *depends on* the interval between observations.



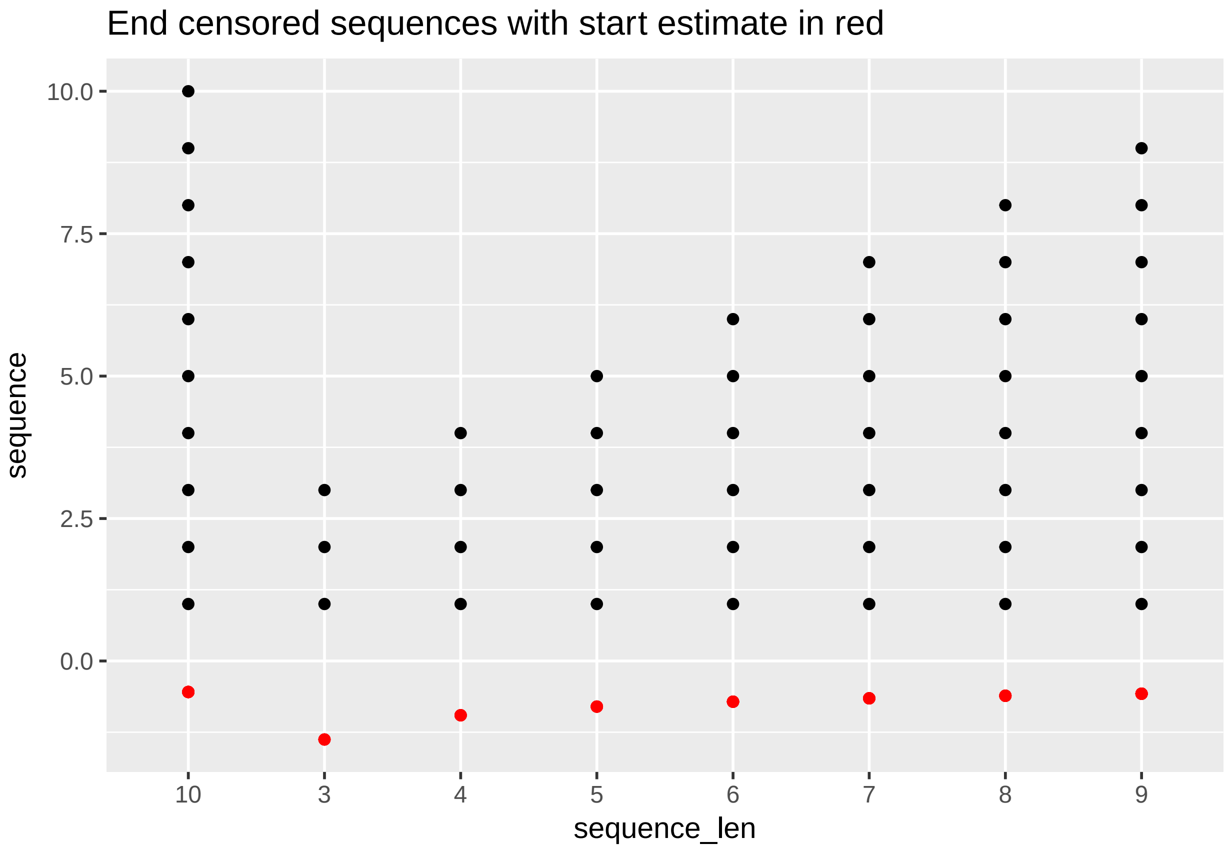
End censoring

How sensitive is this method to end censoring?

I created 8 sequences of observations for 8 different individuals. All 8 individuals have a phenological period that is the same length, but starts on a different day. We do not observe the first 3 to 7 days of 7 of the individuals.

The estimation method produces nearly identical start day estimates for all sequences.

The method is insensitive to end censoring.



In retrospect, this result is trivially obvious. ><

The end

In my actual data, I'd expect all the individuals in the interval censored example to have similar start dates and those in the end censored graph to have really different start dates. This method doesn't work for my data.