Cherry Blossom Prediction Narrative

Nick Pullen

28/02/2022

Introduction

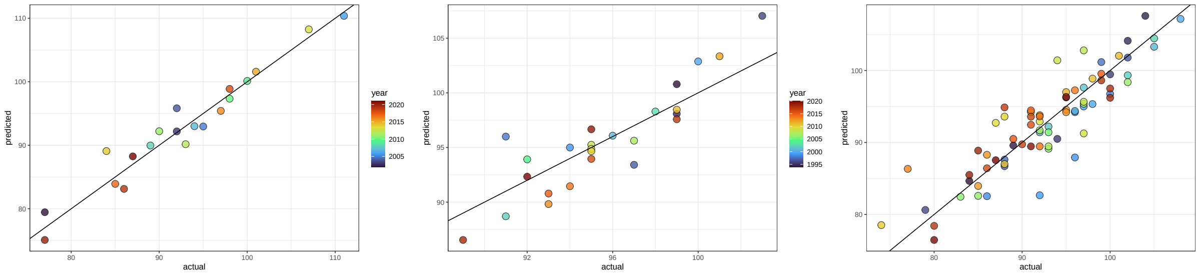
My personal interest for this competition was three-fold. With some colleagues we had just read the end of chapter 4 of Richard McElreath's Statistical Rethinking, where he gives an example of modelling Japanese Cherry blossoms with splines. I am currently living in Switzerland and in December I read on the MétéoSuisse blog an article about the long phenological record of cherry trees in Liestal near Basel. Finally, when I started my postdoc the group I joined had just published a nice paper on linked phenology models for understanding the life history of *Arabidosis thaliana*. I subsequently spent some time looking at the topic and phenology models in general, so this competition was able to rekindle some long-lost memories.

I decided to stand on the shoulders of phenology modellers in my approach. Many temperate fruit trees (in which I include flowering cherries), require a chilling period before spring warmth can be conducive to flowering. I considered two widely-used models which require hourly temperature data for these processes. The so-called Dynamic Model attempts to model chill accumulation with some relation to the molecular biology of plants. The growing degree hours model accumulates a bud-breaking factor in a temperature-dependent manner. A recent development is the PhenoFlex model by Luedeling et al. which links these two with a sigmoid function that transitions between chill and heat. This, and its implementation in the `chillR` package made my life easier. I should also acknowledge Eike Luedeling's online book *Tree phenology analysis with R*, which reminded me of the many things I've forgotten since my postdoc days and gave good examples on how to use the package.

Challenges

The out of sample elements in this competition lead to many challenges. There is the unknown future weather, not just for this year, but until 2031. Further there is no standardised data for Vancouver peak bloom dates, and indeed no definition of peak bloom there.

In an ideal world, experiments would be conducted so that estimates of chilling/heat requirements could be made. Such data of course not being available means other approaches are necessary. I decided to use as much hourly data as I could find (which was unfortunately rather limited for the sites with the longest phenological records i.e. Kyoto and Liestal) with the accompanying bloom records, to optimise parameters of the PhenoFlex model using simulated annealing. This was conveniently available out-of-the-box in the `chillR` package. Using an optimisation algorithm such as this, and with limited data, causes quite the risk of overfitting. The figures below show the actual vs predicted Julian days of peak bloom for Liestal, Kyoto and Washington DC.



By restricting myself to a more mechanistic model I hope to at least capture something of the biology even with this overfitting risk. Due to lack of data I also did not try any forms of cross-validation or similar techniques. I think it could be argued that compared with a deep learning model (my other consideration at the start), fitting 12 parameters for this fixed model, might allow some relative parsimony. I am very interested in seeing the results from other contestants who may have approached this in a more "data-science" way.

Workflow

I optimised the parameters of the PhenoFlex model for each location. One of the parameters defines a threshold for heat accumulation, at which point peak blooming is reached. For Liestal near Basel, this value is around 196 arbitrary units. Figure 1 shows the heat accumulation for the 20 years with sufficient data until this threshold is reached.

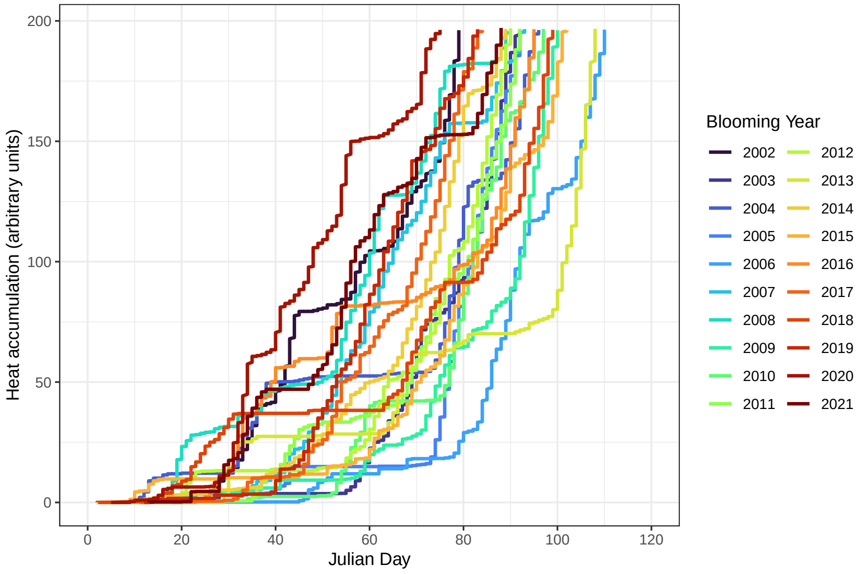


Figure 1: Liestal heat unit accumulation

Following this, we can also look at what has happened for this season since Nov 2021 (Figure 2). This requires the use of the optimised parameters, but now with unseen weather data. Hourly data were available up until the 25th February for the closest station to Liestal, which is Basel-Binningen. Currently we are a little ahead of most years.

To make my 2022 estimates I considered a counterfactual approach. I set the climate data from November to the end of February for every past year to that in 2021/22. Then from this point the actual data for the different years is used. This leads to a range of peak blooming dates but all starting from a common origin (Figure 3). We can answer the question, what would happen if 2022 continues like any of the other 20 years?

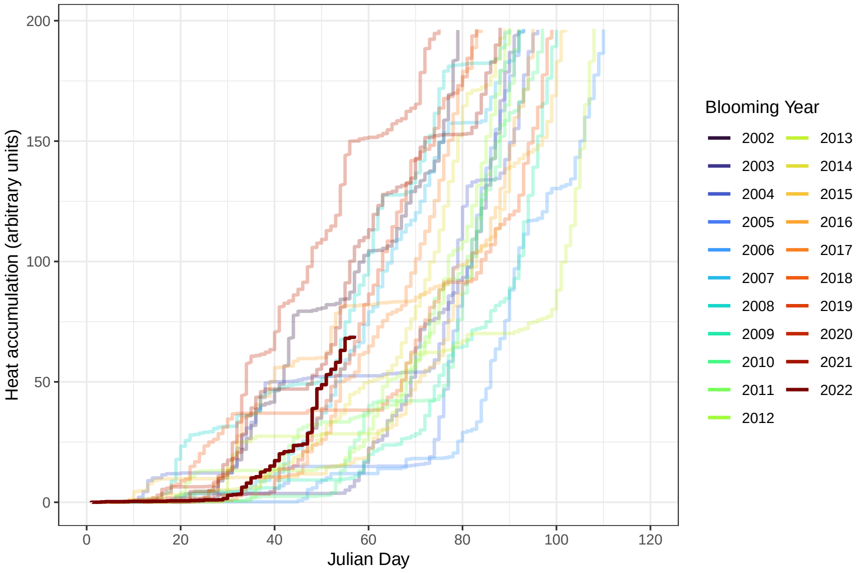


Figure 2: Liestal heat unit accumulation in 2022

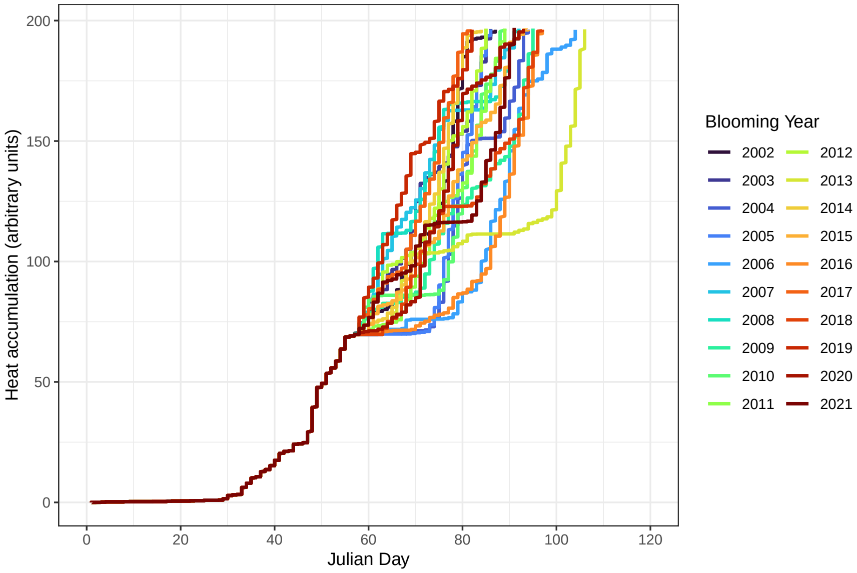


Figure 3: Counterfactual Liestal heat unit accumulation

I use the mean of the predicted Julian dates of peak blooming as my point estimate, but note that it's simple to also provide a range which is slightly more condensed compared with original fits from the optimisation (Figure 1).

For the years up until 2031, although the general trends appear to be for climate warming, accurately predicting the hourly weather is impossible. I thus left my predictions as for 2022 although I did consider randomly sampling within my range of counterfactual predictions for each year.

Extensions

The use of more hourly weather data should help the optimiser seach parameter space. It is possible to generate hourly temps from daily Tmax/Tmin values, see e.g. https://cran.r-project.org/web/packages/chillR/vignettes/hourly_temperatures.html. This would hopefully allow many more blooming records to be used in Kyoto and Liestal. However I think it's worth mentioning that it is the more recent data that has hourly temperatures, and it is also likely to be the most relevant data for estimating parameters that are relevant in a warming environment.

There could also be scope for optimising the start date of accumulation, rather than the generally-used 1st November. I hope that the community of phenology modellers will be able to find new conceptual advances as the main models used here are over 30 years old. Both chilling and forcing could benefit from more biologically-inspired foundations. Nevertheless for the locations in this competition, it may be that chilling requirements may always be overcome anyway even with predicted planetary warming. This suggests that the heat side of the dormancy conundrum could be a lower-hanging fruit. If prediction rather that mechanistic understanding is the goal, I think machine-learning methods will be able to power this by creating or finding new features that have never been used before. The future is ripe for both biological and data-driven approaches.

Finally, to come back to my former postdoc life, in work I did on cauliflower, by mistake I found that it was a harder challenge to predict the date of first flowering, rather than a later date more equivalent to the definition of peak bloom used in this competition. It seems likely that once the first flower has opened, a plant is committed to flowering, and this will generally take place for a set period of time. Being able to resolve what in my eyes is an even more complicated problem would be the cherry on top.