Cherry blossomphenology competition narrative

Andrew McDougall

28/02/2022

In preparing an entry for this competition, I decided to use the demonstration analysis as a starting point, with the intention of making as few modifications necessary to produce a predictive analysis of peak flowing day that best used the data available to me in a biologically sensible manner. Apart from some data joining and the final submission code, very little of the original remains. I decided to separate the analysis for the predictions for 2022 from the remainder of the decadal predictions. I first made some visualizations which show that the trend for flowering timing has not advanced linearly equally at all sites over time. I used a similar method to the demonstration analysis in extrapolating from the existing data, however, I extrapolated from a nonlinear starting point. I feel that the practice of extrapolating based on past observation is highly problematic, regardless of the linearity, and that my predictions (apart from 2022) should be considered with the caveat that the passing of time itself is probably not a driver of variation in flowering date. In order to estimate for Vancouver, where no historical phenological record is available to us, I read "siteclimatedata.csv"(A) into R, and created a GAM for bloom_doy in relation to the maximum temperature for January (tmax1), and February (tmax2). From this model, I predicted some recent dates based on the same variables for Vancouver; "vancouver.csv" (B). These predictions were then extrapolated to produce the decadal predictions for Vancouver. This method has a considerable limitation in that the decadal predictions are based upon a modelled trend, which is itself extremely short: there is a circularity in the inference and projection is biased by recent variation. Perhaps artefact of this, the predicted trend for Vancouver is a delay in flowering timing, in contrast to all other sites.

For estimating the predictions for 2022, I read in a file composed of daily observations from NOAA, coupled with predictions from BBC weather and 'topped up' with missing data from Accuweather: "Climate2022.csv" (C). The models were created from "siteclimatedatawv.csv" (D), 'wv = with Vancouver', meaning that this file has historic Vancouver climate data coupled with the bloom_doy estimates from the previous modelling, added into file (A). While this does continue the problem of circularity, it also allows location to be used as a factor, which allowed me to test the use of individual smooths at each location, and to also test the use of 'lat, long' variables as tensor, i.e., the data arrangement in the model must be the same as the new data.

Model selection and choice of variable rationale

The climate in the month of flowering itself has the highest correlation, but is not useful in forecasting, both outside this competition (unless used at a daily level, not the monthly mean of observations as I used), since some flowering activity may occur in March, or for this competition, since the close date is the 28th of February. I expected the variables for the month closest to flowering (February) to be the most use in prediction based on the climate driven variability in the sequence of events in the flowering process: after induction of the floral meristem and cell division, which are themselves temperature limited processes, the further cell divisions and cell expansion can be accelerated or delayed by variations in temperature. Therefore, it seems reasonable that later variables should be more important for forecasting purposes compared to earlier ones. By modelling the GAMs in the mgcv package and visualizing with the gratia package, I was able to evaluate how variables improved or did not improve correlation when added to the model. To allow the February maximum temperature to have more influence over the prediction than January, I constrained the January smooth (k = 3), but I did not constrain the February smooth. I verified that there were many observations in the model at the same temperature of the data used for prediction. The year variable is not used in these models, and each bloom_doy observation has the most appropriate observation data available to me.

Data have spatial temporal clustering

The observations are considered as independent chronologically, in that the flowering time in one year may not be deterministic of the flowering time in the next year (although I wave no simple way of knowing whether this is true). However, since these are multiple time series data, the data have high spatial clustering, so are not truly spatially independent. This fact, coupled with the desire to model regional effects led me to test the use of use location as a factor.I then selected from the models the one with the lowest AIC score using the 'MuMin' package and estimated for 2022. The model selected for prediction was:

gam(bloom_doy ~ s(tamx2, by = location_ +s(tamx1, k = 3), data = W1data)

Cherry tree flowering phenology 'peaks'.

I modelled all observations together, i.e., under the genus Prunus, regardless of species. Ideally, I would not have done this, but I did not feel confident in interpreting the raw phenological data for USA. The peaks used in the competition vary. Because Switzerland uses 25% flowers open, USA and Canada use 70% flowers open, and Japan uses 80% flowers open, I decided to estimate the number of days between the peaks used. Based on a time-lapse youtube video (1) with visible date and time, I estimated that the number of days between 25% and 80% open is 5 days, and the time between 70% and 80% open is 1 day, based on counts in the frames at 4 pm each day in the video. I then added five days to every Switzerland bloom_doy observation, and one day to every U.S.A. and Canada bloom_doy observation used in regression. The decision to index toward the Japan data was made because I found the most spatially appropriate data for observations with interpretable phenological data form Japan (i.e., no calculations were needed to determine the Japan peak since this had already been done). I did this because I considered that all bloom_doy observations in the regression should be considered as like entities. This method has a limitation in that one indexing value is used for all observations at one site, and yet this value in reality probably varies enormously with temperature, so it introduces an undesirable layer of linearity into the analysis. In addition to this forward indexing, the prediction estimates are also indexed back through the same amount, respective to their sites after regression.

Data: explicit but spatially inappropriate vs. gridded

I used observation data rather than a gridded climate data set. This is mainly because a gridded climate time-series data set was not available to me. For most sites, this was not a problem, but I did not locate appropriate historic data for Liestal. Instead, I used data from Mulhouse, France, which is 54km away, and additionally was informed by predictions from Basel for February 2022, which is 20km away. These were checked against predictions for Liestal in February 2022, which were available. On the 25th of February, the data for four sites for prediction taken for the month of February 2022, to date available at the time (19th of February), using NOAA data. Because temperature can vary, and the underlying trend at this time of year for the northern hemisphere is expected to be a constant increase, a simple mean to date, i.e. the total temperature of 19 available day's data divided by 19, would not be representative of the true mean. Therefore, the additional 11 days' data were added from the forecasts provided by local authorities. Data between the release of NOAA (there seems to be a 2—3 day delay) were added from accuweather.

GDD

GDD methodology provides excellent explanations for phenological observations, but there is currently a caveat that I considered when choosing predictive analyses: although this is science based on observations that can be shown to be an excellent explanation for a phenological result, once the result is known, and in industrial practice used for predictions, this methodology cannot predict forward exactly since weather predictions upon which this method is based are also limited in their accuracy in the same time frame of phenological forward uncertainty. Therefore, I did not use this method.

Interpretation

Underneath the use of non-linear models lies the possibility that flowering time cannot move within infinitely within the year because the year is not infinite, nor is it, or the events within it biologically linear. At some point, flowering may either advance into the previous year, or more likely, fail altogether. Flowering is possible because of plants' accumulation of the products of photosynthesis, but photosynthesis itself is dependent on

a narrow temperature window, and the near entirety of life on earth is dependent on photosynthesis either directly or indirectly. The decision to use observations as close as possible to flowering time, coupled with forecasts of the last days in order to 'top up' the data, is based upon our understanding of the biological nature of flowering phenology, similar to the growth of leaves and roots. The initiation of floral meristems occurs before the expansion of these cells, and further division and increase in cells often takes place during times of ideal, or enzymatically optimal temperatures. Within certain temperature ranges, these activities can also slow due to lower temperatures. From a predictive perspective, the knowledge of the timing of a spell of consecutive days in the optimal range would be expected to allow extremely accurate flowering date forecasting. It is my hypothesis that the relationship with the February temperature with flowering date is indirect, and the true relationship is with these data and the forementioned ideal weather windows in March—April, depending upon site, and that the more accurate predictions will be for sites that typically flower earlier. I base these statements on the fact that flowering does not occur synchronously at all sites, and that the February temperature data would be expected to form closer relationships with March weather data compered to April or May weather data.

Data notes

Note: "liestal1.csv" and washintondc1.csv" have their bloom_doy indexed as described above.

(A) "siteclimatedata.csv": Spatially and chronologically explicit/appropriate climate data assembled against bloom_doy data from the "japan.csv", "kyoto.csv", "meteoswiss.csv", "liestal.csv", and "washingtondc.csv". Climate data were taken from NOAA website and Japan Meteorological Agency (JMA) website. Manipulations, such as removal of redundant header rows were performed within excel, files were read into R and joined to phenological data and written to the final CSV file. NOAA: https://www.ncdc.noaa.gov/cdo-web/confirmation JMA: https://www.data.jma.go.jp/obd/stats/etrn/view/monthly_s3_en.php?block_no=47759&view=1

(B) "Vancouver.csv": Spatially and chronologically explicit/appropriate climate data assembled against latitude and longitude for the Vancouver site, with data from NOAA from the nearest site, Vancouver INT.

(C) "climate2022.csv': spatially appropriate climate data for January and February 2022 used for predicting the 2022, compiled from NOAA, and 'topped up'/ informed by recent observations from accuweather and BBC weather, MeteoSwiss, and meteoblue forecasts.

The climate data for the following was missing and 'topped up' from accuweather: accessed on 25/02/2022 Kyoto, 6th of February https://www.accuweather.com/en/jp/kyoto-shi/224436/february-weather/224436

Mulhouse 2, 6, 8, 12, 13, 16, 17, 18, 23, 25, 27 January, and 2, 3, 6, 11, 16 February https://www.accuweather.com/en/fr/mulhouse/131835/january-weather/131835?year=2022

Basel: MeteoSwiss: https://www.meteoswiss.admin.ch/home.html?tab=overview Liestal: meteoblue (free option, accept all) https://www.meteoblue.com/en/weather/week/liestal_switzerland_2659891 Mulhouse, Kyoto, Washington DC, and Vancouver INT: bbc.com https://www.bbc.com/weather/

(D) "siteclimatedatawv.csv": File (A) with modelled bloom_doy for Vancouver, together with the climate data from NOAA used to create the bloom_doy estimates.

(1) Youtube video of cherry flowering timelapse (https://www.youtube.com/watch?v=uL6XLdZJ35o)

Acknowledgement: I discussed some aspects of GAM modelling with Dr. Steve Delean of the University of Adelaide and thank him for his advice. However, I consulted near the competition close date and he did not have time to comment on the appropriateness of my statistical models. Any mistakes or overreach are all mine.