Predicting Cherry Blossom Bloom Dates

Joshua Wright and Taehoon Song

2/28/2022

Introduction

Cherry tree blossoms are celebrated across multiple cultures. Festivals and sight-seeing tourism often revolve around this natural event. Planning events around this event is difficult seeing as the date is not the same from year to year. As such, it is of interest to try and predict the date of the blossoming. In this research, we demonstrate our proposed model for predicting the peak bloom dates in the coming decade for Kyoto, Liestal, Washington D.C., and Vancouver.

For this report, we will go through each location individually as a different model is used for each. In each location's portion, we describe the logistic growth model's specification, the data used for fitting the model, and the predicted day of year for full blossoming in the years 2022 through 2031.

Washington DC

In the field of phenology, plant growth is often modeled by using growing degree-day (GDD) as a predictor. Since the National Park Service (NPS) publishes dates of various growth stages for 2004 through 2021, we use this to build a logistic growth model (also known as the Verhulst model) that helps us estimate the threshold GDD for which we can expect the stage for the flowers to bloom. The authors use 4 degrees Celsius as the base for calculating GDD and fit a logistic function to the phenological stages from side green as a function of accumulated GDD.

$$y = \frac{k}{1 + \left[\frac{k-n_0}{n_0} \cdot \exp(-r \cdot \text{GDD})\right]}$$

where y is the phenological stage and k and r are empirical factors related to limited growth factors and growth rate, respectively. This type of model has been shown to work for predicting phenological stages given GDD for sour cherry trees (see Phenological Models of Flower Bud Stages and Fruit Growth of 'Montmorency' Sour Cherry Based on Growing Degree-day Accumulation by C. Zavalloni, J. Adresen, and J. Flore). y in our model is an integer value which ranges from 3 (green budding stage) and 8 (peak blossoming stage). Stages 4 through 7 are those defined in the NPS data, but are only used for fitting this model. We use this model to fit the growth stages published by the NPS as a function of cumulative GDD. In our data from the NPS, we only have data starting at green budding (stage 3), as such our n_0 will be 3.

Obtaining temperature and phenological data for DC

In order to determine at what day of year we start calculating GDD, we need to determine when our n_0 stage is estimated to begin. To do this, we will look at the range of dates that green budding is reported and take the minimum, which is found to be at 50 days into the year. Thus, we only use the average temperatures after the 49th day of the year to ensure that we start calculating GDD around when Stage 3 occurs (the reason we calculate the GDD at this point is because the logistic growth model assumes the GDD is zero at stage n_0). For accessing historical temperature data, we use DayMet for years 1980 through 2020 and the NOAA database for 2021.

Building the growth model

As mentioned above, we calculate the GDD using a base of 4 degrees Celsius and starting from 49 days from the beginning of the year. Then, we fit the known phenological stages for the cherry blossoms in DC to the calculated GDDs. Fitting the model in *R*, we get the following output:

```
##
## Formula: stagenum ~ k/(1 + (k - n0)/n0 * exp(-r * GDD))
##
## Parameters:
##    Estimate Std. Error t value Pr(>|t|)
## k 10.223870   1.362473   7.504 4.72e-11 ***
## r  0.011973   0.001598   7.495 4.93e-11 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.9374 on 88 degrees of freedom
##
## Number of iterations to convergence: 8
## Achieved convergence tolerance: 7.789e-06
```

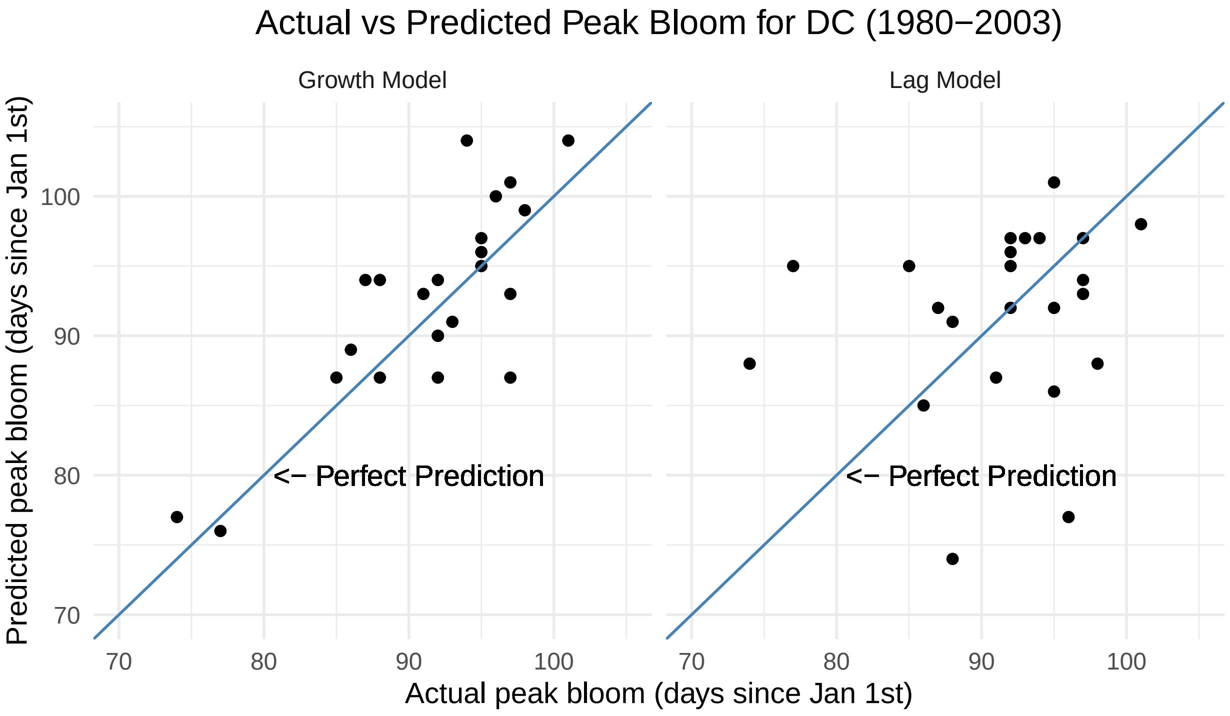
The model results show that both k and r are significant being estimated to be 10.22 and 0.012, respectively. What we are interested in is at what day of year do we expect the cherry tree to reach stage 8 (peak blossoming). With this model, we calculate the expected stage from a given GDD. Once we reach the first day that is estimated to be stage 8, that is our prediction for bloom date of the year. We can test the prediction performance of this model by calculating predicted peak bloom dates in the past.

Checking performance on historical data

Now that we have the growth model, we can test its prediction performance out on historical data that we did not use. Using temperature data from 1980 to 2003, we compare the logistic growth model to the lag model to see which is a better fit. We define the lag model to be where the predicted bloom date is just the previous year’s date.

Logistic Growth Model MAE	Lag Model MAE
3.347826	6.304348

Our growth model has a much lower mean absolute error (MAE) compared to simply using the previous year’s bloom date as a predictor. This is good because this is the bare minimum a useful model should be able to do. This is more apparent when we plot the two models against the actual peak bloom days since January 1st.



Predicting future GDD and future peak bloom dates

For us to predict future peak bloom dates for Washington D.C., we must first predict future GDDs. We could predict future temperature and calculate GDD from that; however, that would be unnecessary because we always start calculating GDD at 50 days and use the base of 4 degrees Celsius. We can directly predict GDD to avoid that step. For GDD prediction, we will fit a model to the GDD using a time series seasonal ARIMA model.

With a time series model fit, we can now predict the peak bloom day for DC using the predicted GDD as an input for the logistic growth model. Feeding our predicted GDD values into our model outputs the following predictions:

Year	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031
Day of Year	86	86	86	85	85	85	84	83	83	83

Vancouver

Vancouver doesn't provide historical peak bloom dates. However, we use the growth model from DC to predict future peak bloom dates because the tree species are identical (or close to identical) and the peak bloom definition is the same. We will use temperature data to once again calculate the historical GDD values and fit another time series model to predict future bloom dates.

Similar to DC, we obtained the historical weather data for Vancouver via DayMet. However, this dataset only goes from January 1980 to December 2019, so we use the `rnoaa` *R* package provided by NOAA to obtain the remaining data up to Jan 2022. We undergo the same process as we did for Washington D.C. in order to fit the time series model and use the exact same logistic growth model for bloom day predictions.

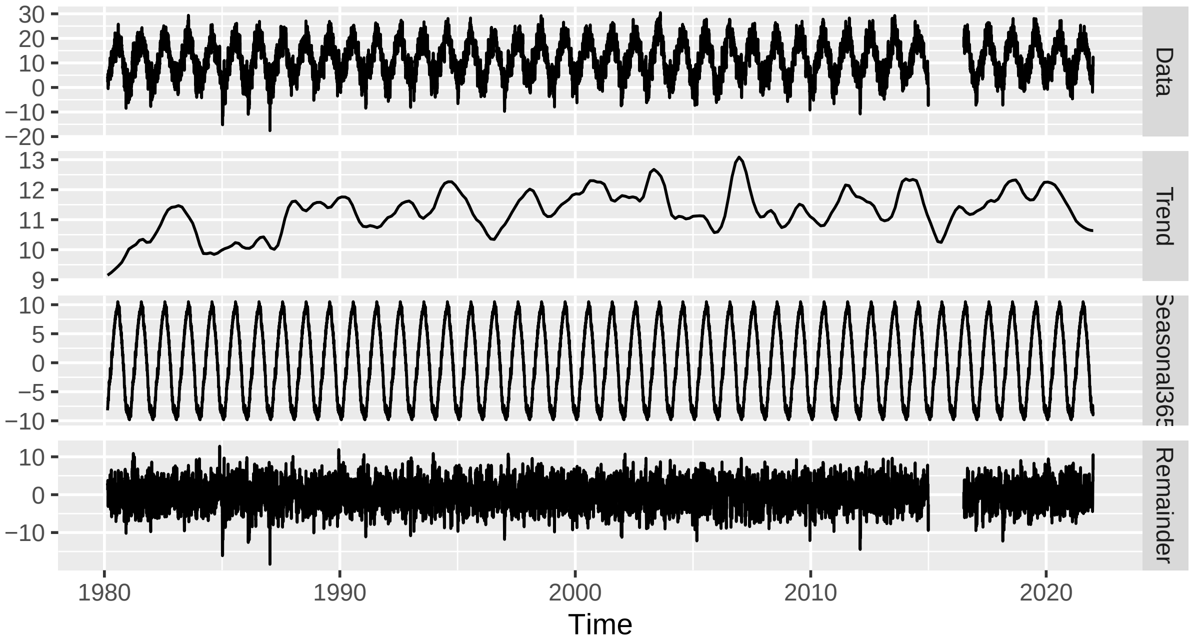
Year	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031
Day of Year	109	108	108	107	106	106	106	105	104	104

What we have are later predictions later than what we saw in Washington D.C., but that is to be expected given the lower average temperatures in Vancouver.

Liestal

The cherry tree species in Liestal is a different species from the ones in Washington D.C. and Vancouver. Additionally, the percentage of blooming for the tree to be declared as at the stage of peak bloom is less as well. Ideally, we would want to fit a new growth model using historical phenological stages. Unfortunately, such a dataset is not readily available. Thusly, we use the same growth model parameters k and r from original to calculate the mean growth stage for historical peak bloom dates. This is to say that we will use the cumulative GDD from each of the historical bloom dates to see what y value our model would predict at that GDD. We will once again start calculating GDD at the 50th day of the year. We use the average growth stage for the past 5 years as the cutoff for our prediction, which is found to be approximately 6.8.

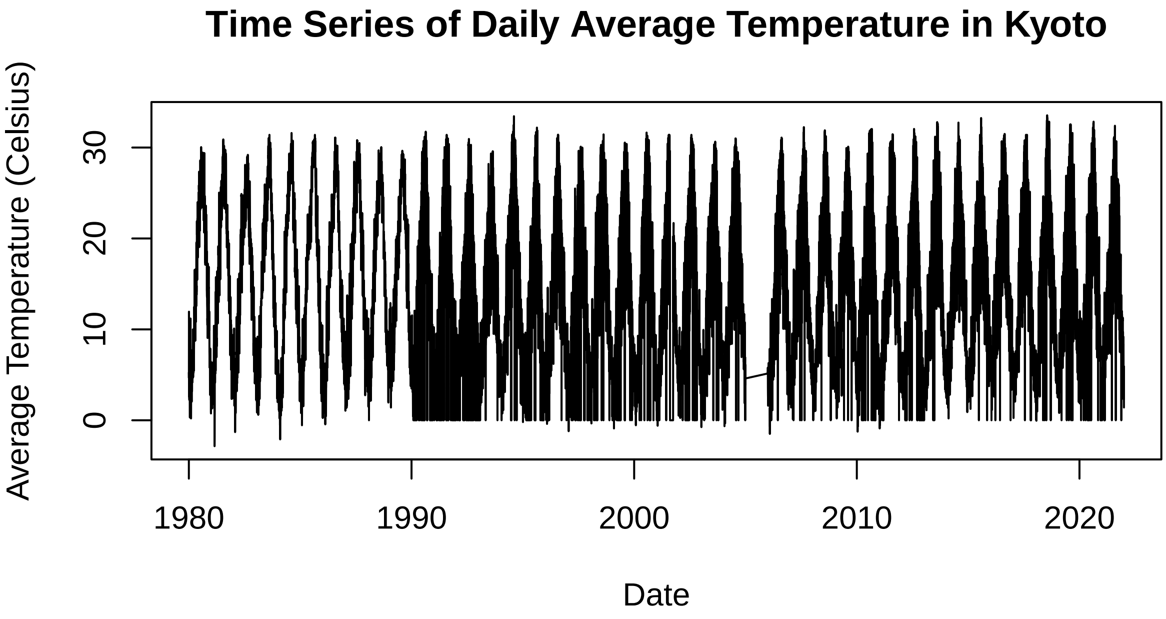
We can approximate the peak bloom for the cherry trees in Liestal (as defined by MeteoSwiss) using the growth model for the trees in Washington D.C. if we use 6.8 as the predicted growth stage for full blossoming. Another change we use here is directly predicting the temperature to then find the GDD as opposed to just fitting a time series model to the GDD. This is done due to the new circumstances making GDD time series model fits less reliable than just predicting temperature. After predicting the temperature, we then calculate the GDD and find which day of year is the first to reach stage 6.8. Our predictions are listed below.



Year	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031
Day of Year	88	88	88	87	87	87	87	86	86	86

Kyoto

Similar to predicting peak bloom dates for Liestal, we would like to predict the peak bloom dates for Kyoto (as defined by a local newspaper in Arashiyama) using the growth model for the Washington D.C. cherry trees. Alongside the issues of having a different species of cherry tree without phenological stage data and a differing definition of peak bloom, we also have the issue of unreliable historical temperature data provided by NOAA. We look at the time series and see the extent to which it is unsuitable for modeling below.



Notice that the data abruptly cuts off at 0 commonly starting at the year 1990. This is likely representing missing values due to these freezing temperatures being abruptly different than previous temperatures and sometimes even in the middle of summer. Due to the unreliable nature of weather data for Kyoto, we reluctantly use a simple linear model to predict future peak bloom dates using year as the predictor.

Year	2022	2023	2024	2025	2026	2027	2028	2029	2030	2031
Day of Year	93	93	92	92	92	92	92	92	91	91

Summary of Predictions

We have described the methodology used for bloom day predictions for each of the locations. The summary table below shows the final predictions for the four locations from 2022 to 2031.

Year	Kyoto	Liestal	Washington D.C.	Vancouver
2022	93	88	86	109
2023	93	88	86	108
2024	92	88	86	108
2025	92	87	85	107
2026	92	87	85	106
2027	92	87	85	106
2028	92	87	84	106
2029	92	86	83	105
2030	91	86	83	104
2031	91	86	83	104

Closing thoughts

If historical phenological stages were recorded for all locations in this competition, we would have been able to fit a growth model for each species and location. This would have allowed us to create more accurate models specific to that cherry tree species and environmental factors. Additionally, weather data was extremely limited or unreliable for locations outside the United States, particularly Japan. If historical weather data was freely and readily available to the public, we would have been able to create a better time series model for GDD.