Cherry Blossom Analysis Narrative

Cherry Blossom Prediction Competition

Matt Harding

2/27/2022

My fiancé and I recently moved to Washington D.C. this past August (2021) and not too long after her best friend moved here as well. I vividly remember them talking about how they needed to have all of their friends come into town for the cherry blossoms in the spring. Neither of them are from near Washington D.C. and I grew up about 45 minutes away. I told them it would be tough to tell people when to come because its hard to know exactly when the trees will bloom. We spent the next half hour trying to find predictions for when the trees would bloom. The best predictions we could find gave us ranges no smaller than a few weeks.

This is a microcosm of a much larger issue in Washington D.C. and other cities that boast large cherry blossom festivals such as Kyoto and Vancouver. The festivals are both a giant tourism opportunity for the cities and a massive attraction for tourists. However, it is difficult for cities to advertise the festivals when they don't know when exactly they are going to occur. And its significantly more difficult for tourists more than a few hours of driving away to take time off work, book flights and lodging, and make a trip to a city where they are only hoping to see the cherry blossoms blooming.

My approach to tackling the issue of predicting the peak bloom dates of cherry blossom was first to understand what makes tress bloom. I found through a few different sources including an article from the Washington Post[1] that the flowers on the trees blossom after a series of "warm enough" days. With that information I decided I would try to develop a model that attempted to find and account for an accumulation of temperature rather than using averages or other summary statistics.

I gathered weather data as far back as I could from Meteostat[2] for Washington D.C., Kyoto, and Vancouver. Unfortunately, there was not data available from that source for Liestal so I gather that from rNOAA's database. I wrangled the data to get values of minimum, maximum, and average daily temperatures. From there I summarized each years weather data into different variables in order to try to capture certain weather patterns. From both what I read and testing different variables in models I found that the best way to summarize the weather was to capture trends around the time that the cherry blossoms peak. With this in mind, most of the variables created were cumulative temperatures over different spans of time in March and April.

Using these variables, I trained a net elastic regression model to predict the days of peak blossom in each location. I used 10-fold cross validation to tune the parameters of the model. I ended up choosing this type of model over others because I was using a large amount of variables for the sample size and the net elastic regression allowed me to reduce the bias of the model. Additionally, each of the variables created followed a mostly linear trend, as can be seen in Figure 1. Because of this I did not have to consider modeling with polynomials or a spline and a linear net elastic regression model works well.

To train the model, I used weather data in combination with peak bloom data from Washington D.C., Kyoto, and Liestal. There was not available data from Vancouver for peak blooms so I could not use it to train the model. I split the data into training and testing sets with 75% of data being used to train the model. Using 10-fold cross validation to to tune the parameters of the net elastic regression model, I found the optimal alpha and lambda parameters for the model to be 0.1 and 2.55 respectively. After training the model using the training data, I used the test data to test the predictive accuracy of the model which had a root mean squared error of 5.92.

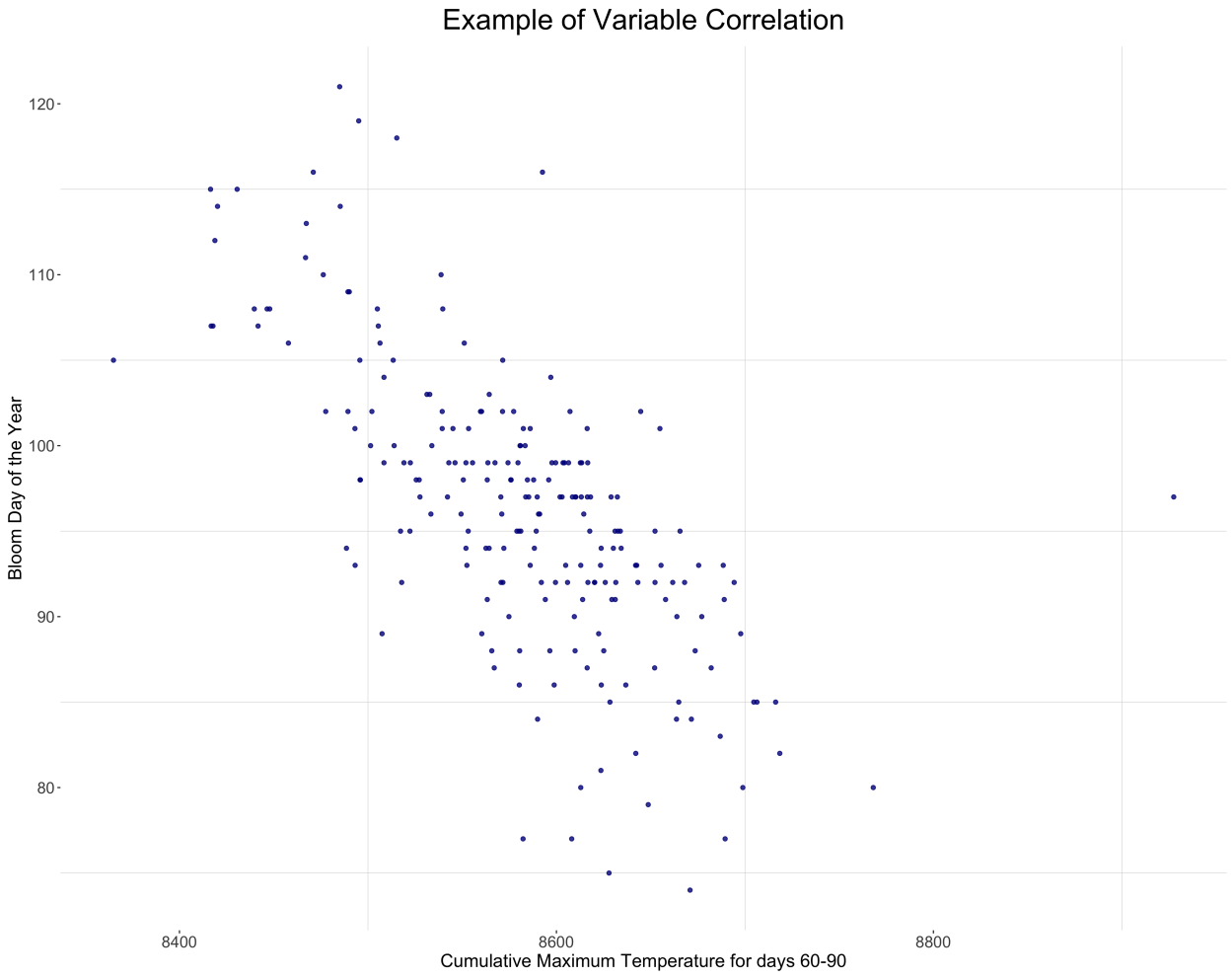

Figure 1: Figure 1

After developing the model to predict the peak blossom days, I needed a model that predicted temperature because weather forecasts do not predict more than a few months in advance at most. It was difficult to capture the randomness that accompanies weather patterns but I tried my best to capture the trends using a generalized additive model. I chose this type of model for the weather data because some of the data followed non-linear trends. I used a spline regression to capture how temperature changed over the years and a quadratic regression for how temperature changed over the course of a year. I also used latitude in order to capture weather differences in the different locations. This model did not capture the randomness of weather patterns but instead the overall trend. I did attempt to capture the randomness of the weather however ultimately did not figure out a way to do so.

Using both models, I was then able to predict the peak bloom days for each city as seen below in Table 1. I was able to include real temperature data from 2022 up until the end of February for a couple of the locations where data was available for that and I feel as though that should help with the prediction for this year. Ultimately, the real peak bloom date will likely not be as consistent as shown in Table 1. However, I do believe that using the model for predicting the peak blossoms, people could get a more accurate time frame for the peak blossoms as weather predictions become more accurate a few months out from blossom season. At the very least, I hope that the ideology of utilizing the accumulation of temperature over the time frame right before and around peak blossom times can be used to make more accurate predictions with better weather data and models and ulitmately allow people to make more concrete plans around the peak of the blooms.

Cherry Blossom Predictions

Table 1

Year	Kyoto	Liestal	Washington D.C.	Vancouver
2022	89	96	91	97
2023	89	96	91	97
2024	89	96	91	97
2025	89	96	91	97
2026	89	96	91	97
2027	89	96	91	97
2028	88	96	91	97
2029	88	96	91	96
2030	88	95	90	96
2031	88	95	90	96

[1] https://www.washingtonpost.com/national/what-makes-the-cherry-trees-bloom-when-they-do/2011/03/29/AFvRbRfC_story.html

[2] https://dev.meteostat.net/bulk/daily.html#endpoints