George Mason University

Peak Bloom Prediction

Daria Kearney

Jeremy Fischer

Justin Sachse

STAT 634: Case Studies in Data Analysis

Dr. David Kepplinger

March 22, 2022

Abstract

Blossoming cherry trees have long been an anticipated event around the world, which has continued to this day. As such, when cherry trees reach peak bloom is a question of scientific interest and cultural importance. This paper focuses on analyzing publicly available data to predict peak bloom dates for cherry trees located in Kyoto, Japan; Vancouver, Canada; Liestal, Switzerland; and Washington DC, USA between the years 2022-2031, respectively. Historical climate data used in the analysis was sourced primarily from the National Oceanic and Atmospheric Administration (NOAA). Recorded peak bloom dates were then used along with this data to fit multiple decision tree based models before finalizing a model that yielded the most accurate predictions. Additionally, as peak bloom dates have trended earlier in the season in conjunction with a rise in global temperatures, our final model suggests that peak bloom for 2022 will be on March 28th for Washington DC, April 1st for Vancouver, April 5th for Kyoto, and April 8th for Liestal.

Introduction

Yoshino cherry trees are a tranquil flowering plant that has been celebrated in Japan for centuries, as the history of cherry trees goes back to the seventh century when it is believed that cherry trees were carved into existence by a mountain deity. Most Americans are familiar with cherry trees from the ones planted along the Potomac River in Washington D.C. These were a gift from Japan as a gesture of goodwill in the early 20th century, brought upon by the American diplomat and journalist, Eliza Scidmore. After encountering cherry trees on a trip to Japan to visit her brother, she is quoted to have said that the cherry trees were "the most beautiful thing in the world." She subsequently spent almost twenty-five years advocating the United States government to bring them from Japan. Historically, the cherry trees in Washington D.C. have reached their "peak bloom" in late March or early April. This peak bloom date is defined as the date when the cherry trees have seventy percent of their flowers blooming.

Unfortunately, the peak bloom dates in Washington D.C. have continuously been observed over the past few decades as occurring earlier in the year as global temperatures have increased. This is not a local phenomena, as earlier peak bloom dates have been recorded in Kyoto, Japan and Liestal, Switzerland. This change in the cherry tree phenology is considered evidence of global warming's impact, and an example of the changing environment in response to atmospheric change (as shown in Figure 1). A successful predictive model that would be able to forecast global peak bloom dates would provide continued evidence of global warming and possibly provide information to populations on how to adapt to changes in their livable space.

Forecasting atmospheric and phenological patterns is extremely difficult due to the multitude of confounding factors. In addition, given the regime change observed in the past thirty years, it is plausible to believe historical values may be limited in their predictive power of future peak bloom dates. This report discusses the various obstacles that were encountered during the case study, and documents our results. The first section describes the data sources we use to build our model, and the various techniques to impute missing values and create additional variables we deem necessary in the final predictive model. The second section details the alternative models, and the third section outlines our results with the final model, where we discuss the conclusions we arrive at.

Section 1: Data Sources and Imputation

The RNOAA package in R is the main data source for weather data used in our analysis. The RNOAA package catalogs atmospheric data sourced from the National Oceanic and Atmospheric Administration (NOAA). The data available includes daily temperature data on all four locations covered in the case study: Washington D.C (USA), Kyoto (Japan), Liestal (Switzerland), and Vancouver (Canada). The locations are organized in the dataset by stations, in which each station corresponds to the longitudinal and latitude coordinates of where the climate data is captured. Temperature data is supplied to the NOAA by nearby airports. We noticed that some weather stations in the RNOAA dataset contained more information than others. While most stations had historical data going back to the 1940s, not all stations captured the same variables. To maximize the amount of information used from the RNOAA dataset and in an attempt to create a more unified model, we only consider variables that were present across all four stations of interest. These included daily maximum temperatures, minimum temperatures, total precipitation, and total snow depth. Additionally, for these variables we focused on data captured from December through April since those are defined periods on when weather patterns impact cherry tree bloom cycles. For instance, it does not make sense to use data patterns in July when no cherry trees are observed to bloom in July. The latest observed period in which peak bloom dates were observed occurred in May. In order to work with the daily temperature data provided in the RNOAA dataset, we pulled the minimum and maximum of each day's recorded temperature across the four locations.

Additional candidate variables were identified with the CDC's Wide-ranging Online Data for Epidemiologic Research (WONDER) database. Specifically, daily sunlight data (measured in KJ/m^2) was pulled for Washington D.C. and Whatcom County, WA as this was the closest location to Vancouver available. Similar to the RNOAA dataset, sunlight was averaged by month for December through April for all years available (1979-2011). As we were not able to find reliable sunlight data for Liestal or Kyoto, this information was used solely to improve predictions for Washington D.C. and Vancouver. While some variables were ultimately ruled out from the analysis or only included for certain locations, our starting dataset contained 266 observations of 41 variables, and included 1,952 missing values (representing 18% of observations, respectively).

Data imputation was needed for these missing data points from the RNOAA and CDC sunlight data. Our approach focused on using decision trees and it was the primary method used for most of the data. It uses supervised learning in order to predict values on missing data points. The initial imputation is the median for the variable of interest, which is then improved by using proximity matrices generated after iterating through a chosen number of random forest models. Another method that was applied for years where peak bloom date was unknown, which uses a pseudo supervised model where the variable with missing values is selected as the response variable.

Given the average minimum temperature and average maximum temperature, we classified each day as either "hot" or "cold" based on the average temperature (i.e. the midpoint between the average minimum temperature and average maximum temperature observed on a given day). If the average temperature was less than or equal to 11 °F then the day was classified as a "cold" day, and if the average temperature was greater than or equal to 50 °F then the day was classified as a "hot" day. The classification was used in order to count the number of days that would be considered cold or hot within a specific month. Recent weather events are relevant

to when cherry trees reach peak bloom, since cherry trees react to the immediate environment. By measuring the occurrence of recent hot days in a given month, our model assumes the cherry trees determine when it is the best period to peak based on recent temperatures.

The thresholds were determined based on a multivariate analysis with the year, location, hot and cold as the independent variables and the dependent variable as the number of days from January 1st until peak bloom. The threshold that produced a model with the lowest sum of absolute difference and lowest R^2 were then used. The predicted versus actual values in the linear regression with that criteria is shown in Figure 8. The final model produced a sum of absolute difference and R^2 for 2011-2020 of 167.9 and 32.4%, respectively.

For variable selection purposes, all of our preliminary and final predictive models used temperature as an independent variable, whether as a direct or indirect measurement. Additional models without temperature were not considered since it is well-known that plant phenology is highly correlated with temperature, even without considering the presumed impact of global warming. Additionally, if a predictive model that uses temperature is not possible to construct, then it would provide evidence that global warming may not be a prevalent factor impacting earlier peak bloom dates.

Section 2: Alternative Models

As stated in the previous section, temperature was used in all predictive models. Providing additional evidence of the importance of temperature as a model parameter, a simple time series model was fitted to the peak bloom data on the three locations in which reliable peak bloom data was available (Washington D.C., Kyoto, and Liestal). The historical pattern in the peak bloom dates from the early 1990's to the present show a clear linear trend. In order to model this trend, the data was separated into a training set from the years 1950 to 2000 to test against the testing set from the years 2001 to the present, contingent on model fit.

Each model using this linear trend as a factor demonstrated a negative coefficient on the trend variable, as expected given the historical data. The coefficients were measured as -0.0384, -0.0008, and -0.0613 for Washington D.C., Kyoto, and Liestal and neither showed statistical significance at the 5% confidence level. These results could allude to a poor model fit or that the trend may not be differentiable from the random variation within the data. Scatterplots of the observed peak bloom values with their lagged one, two, three, and four observed values showed there was a correlation between observations that was possibly linear, but a correspondent model without temperature was determined to be unsuccessful.

With the large number of variables, and our focus on predictive accuracy, we chose to consider decision tree algorithms in addition to the linear trend model discussed in the previous section. Random forests and gradient boosted regression tree ensembles were both considered and validated on data from 2011-2021. In order to establish a baseline to compare model performance, we fitted two linear regression models and evaluated our final model against the observed test error for the 2011-2021 subsetted observations. The first linear model was replicated from the demo analysis with peak bloom date regressed on year and location. This model yielded a test error of 201.46. The second linear model regressed peak bloom date on our calculated Spring and Winter average maximum temperature variables and were included in the model through an interaction term. This second linear regression yielded a test error of 196.26. With the baseline test errors established, we proceeded with the decision tree algorithms.

For the random forest approach, parameters were chosen for m, the number of predictors considered at each split and, n, the number of trees to build. According to standard guidance m is often chosen to be approximately the square root of the number of predictors. In our case this was roughly $m = 6$ given the nearly 40 variables available. The number of n trees was set to 5,000 given a test set was set aside for 2011-2021 and we weren't as concerned with overfitting. Several options for m were back tested to identify the best test error, as defined in our baseline cases - the sum of the absolute difference of the residuals for Washington D.C., Kyoto, and Liestal between 2011-2021. The model with the best results yielded a test error of 118.86 with $m = 10$. It's informative to note that the variables which lead to the greatest decrease in mean square error were the average maximum temperatures in March, average minimum temperatures in March, average maximum temperature in February, and year.

For the gradient boosted regression ensemble, we focused on optimizing parameters for the number of trees, B, the shrinkage parameter or learning rate, λ, and the depth of each tree, d. To stay consistent with the random forest approach we selected $B = 5{,}000$ and chose a small learning rate of $\lambda = 0.001$ to allow for best performance given the large number of trees. For the depth we tested $d = 1, 2, 3, 4$. Again the performance of the model was evaluated based on the test error as defined previously. After iterating through the different options our best results yielded a test error of 100.61 using a depth of $d = 2$. Looking at the relative influence plot we saw that the most important predictors were again average maximum and minimum temperatures in March, and February.

Both the random forest and gradient boosted regression models outperformed the two baseline linear models, with the gradient boosted regression model showing the most promising results given the lowest test error. We decided to move forward making our final predictions with the gradient boosted regression model**.**

Section 3: Forecasting

To forecast the majority of our covariates from 2023-2031 (i.e. forecasted temperature), a Monte Carlo simulation was conducted to estimate sampling distributions of the covariates and to generate a distribution of temperature variables by year to use for predicting bloom dates between 2023-2031. Each of the temperature covariates were assumed to have a Normal distribution, and the precipitation and snow depth covariates were assumed to have a Poisson distribution so as not to yield any negative values (refer to Figure 6). Each of the covariates was regressed onto year and location to obtain the predicted distributions. Next values were randomly drawn from these distributions, and utilized in the gradient boosted regression model to generate a prediction for each year and location. This process of randomly drawing values for the predictors and generating a prediction was repeated 1,000 times. The median value for each of the 1,000 simulations was chosen for the final prediction for the forecasted years, 2023-2031. The median had more variation than the mean values and was chosen given the variability observed in the historical record.

We used a different approach to forecast the hot and cold days. When forecasting hot and cold days by multivariate linear regression it performed poorly, hence we forecasted average temperature and then used that model to forecast the hot and cold days. To check the forecasting performance on the hot variable we back predicted the hot variable using the actual average temperature and predicted average temperature. As shown in Figure 7, the performance of the forecasting methods (blue line) closely followed the actual distribution of the hot variable (red

line). Overall, forecasting hot covariate performed well in back predicting so theoretically should be accurate in predicting the peak bloom day for the future. The two drawbacks when trying to forecast these variables is that when predicting the hot variables, the number of hot days never went above 5 for that season which makes it biased because some locations, especially Washington D.C., had a total of 13 hot days in one season. Therefore, it is important to notice that there are limitations in this forecasting model.

Conclusion

In the feature analysis on our gradient boosted model, it was observed that the number of hot days in a month preceding peak bloom was one of the most significant features. This was not surprising given the plant phenology, and our results provided additional evidence that warming global temperatures are impacting peak bloom dates. Warming global temperatures are a much-discussed phenomena with polarizing debates in the political arena about its validity or causation. Cherry trees blooming earlier in the season may seem to be an insignificant example of this catastrophe, but it epitomizes the changing nature in which humans will have to adapt cultural practices with the new, warmer environment. Changes in the phenological cycles may also cause disruptions that are difficult to currently predict. Additionally, if humanity is able to correctly predict changes in the environment such as peak bloom dates for cherry trees, then there may be sufficient methods available to predict other weather patterns that may have a more economical or dire consequence on society.

Finally, we're happy to report that our analyses yielded peak bloom date predictions for 2022 of March 28th for Washington DC, April 1st for Vancouver, April 5th for Kyoto, and April 8th for Liestal.

Appendix

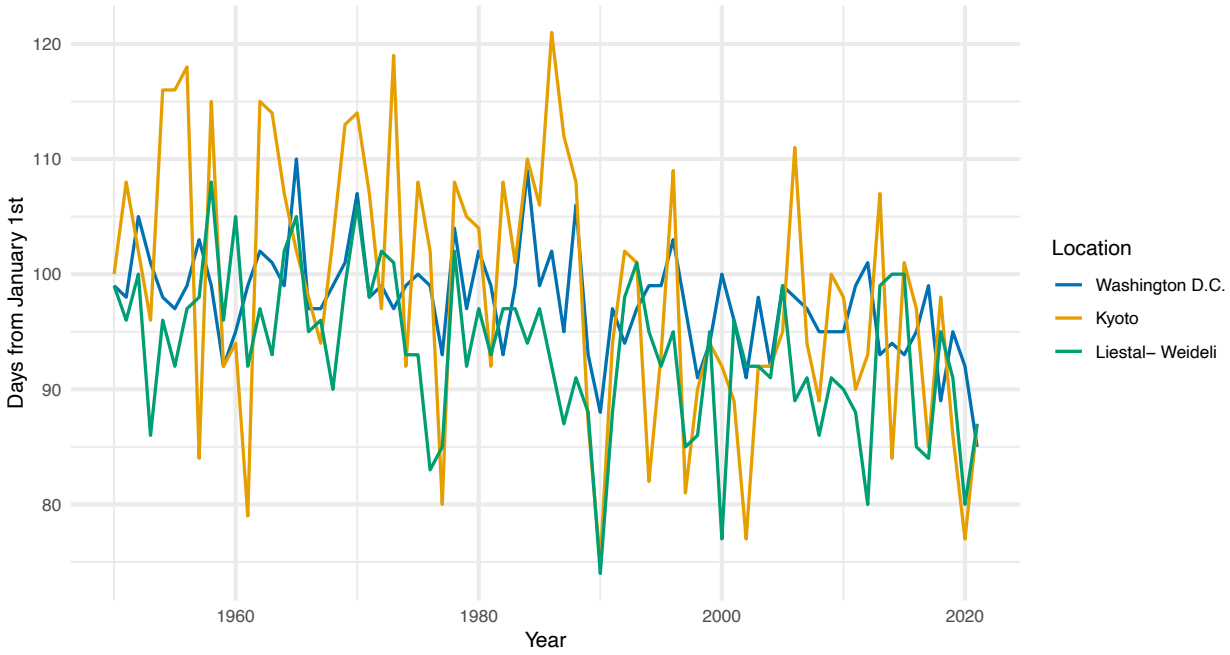

Figure 1: Peak bloom days starting January 1st of three different locations from 1950 to 2021

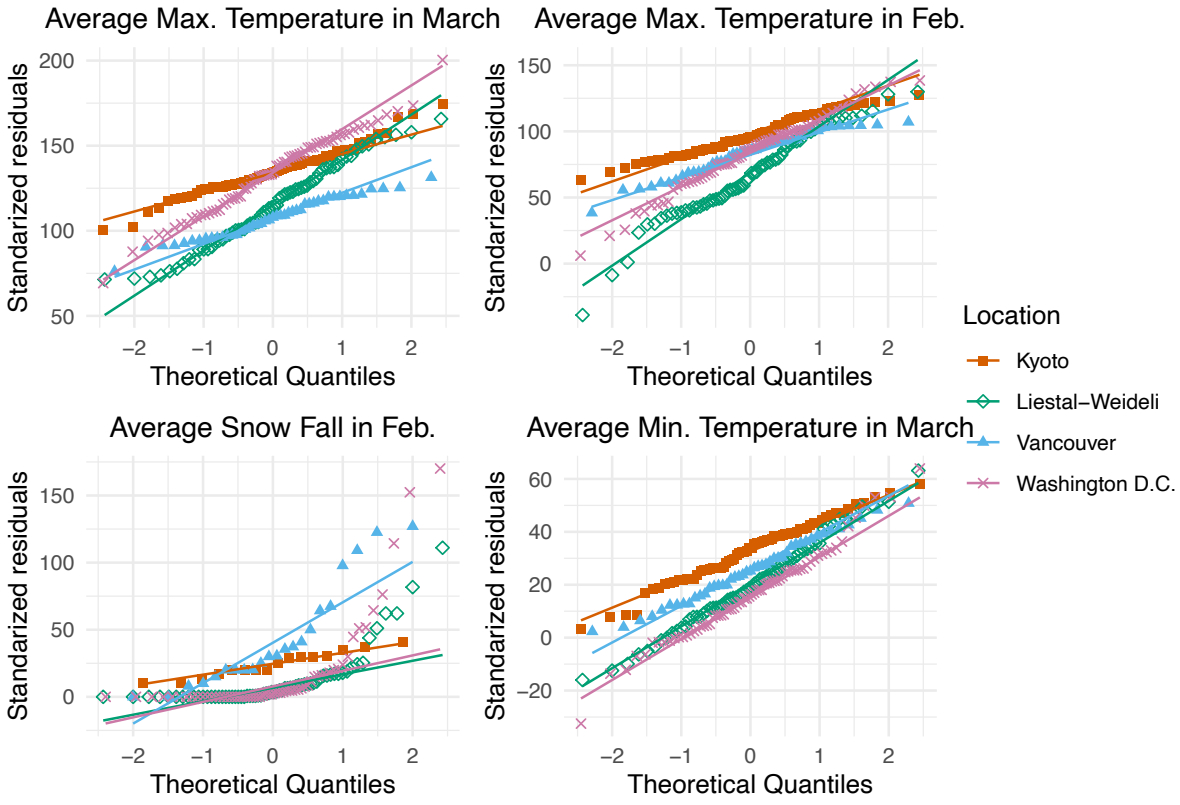

Figure 6:Q–Q plots of four forecasted variables from the different locations.

Figure 7: Line plot comparing the three different methods of predicting the number of hot days.

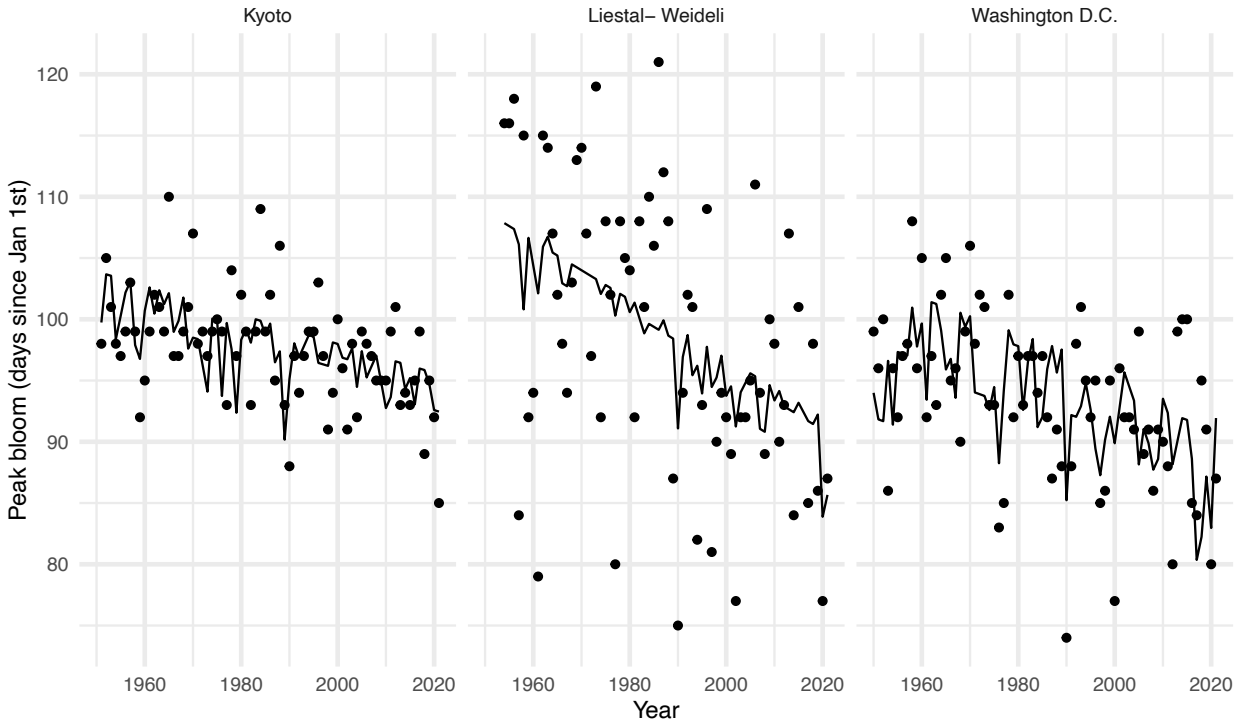

Figure 8: Scatterplot of actual bloom days. The lines show the predicted bloom days as obtained from the linear regression model.