Cherry Blossom Prediction

Miaoshiqi Liu Siyue Yang

February 28, 2022

Contents

Abstract

This project focuses on the prediction of the cherry blossom date of Kyoto, Liestal-Weideli, Washington, and Vancouver, where the varying-coefficient regression model and semi-supervised learning method are utilized separately for different cities. ARIMA model is applied to forecast the weather data in the next 10 years, which allows us to obtain the predictions for the peak blossom days.

1 Introduction

With a distinctive increase in the global temperature, a seasonal advance of the cherry blossom has been observed during recent years. To construct a model that predicts the peak blossom date, the existing literature generally considers both the chilling process and the heating process. Legave et al. (2008) tried different combinations of chilling sub-models and heating sub-models to study the apple flowering time. Similarly, Chung et al. (2011) used the bud-burst model which divides the flowering process of deciduous trees into two stages:

dormancy including rest and quiescent periods during autumn and winter, and the flowering period following bud-burst in spring. This stimulates the necessity to find indicators of the heating effect, such as the temperature in Spring, and the chilling effect, such as the temperature in Winter.

Following the description of chill days in the bud-burst model, we constructed a variable named "Accumulated Freezing Degree Days (AFDD)", which is computed as the accumulated number of days when the temperature falls below the threshold $0°C$ in Winter. For a better prediction, the potential covariates also include the time (year), average maximum/minimum temperature in Winter/Spring, and the average precipitation in Winter/Spring. These candidates are analyzed using a preliminary analysis presented below, and are finalized further during each variable selection process in the modeling.

1.1 Preliminary Analysis

A preliminary analysis is conducted to examine the characteristics of peak bloom and the potential covariates (weather data). The bloom peak data is only available for Kyoto, Liestal, and Washington DC. We extracted Vancouver bloom peak date during 2004-2021 from the National Park Website. Weather data including daily temperature and daily precipitation are extracted from an R package with built-in API connected to the National Oceanic and Atmospheric Administration climate database. We noticed that weather data is missing across several years, thus using the Kalman Smoothing time series model to impute the daily temperature, available by Moritz and Bartz-Beielstein (2017). Figure 3 in the Appendix shows the result of the imputation.

From Figure 1, the distribution of peak bloom days is different across the four locations, which suggests that a common model to quantify the peak bloom is not feasible, and separate models should be considered to characterize the change. Peak bloom days in Kyoto are more concentrated, while the days in other three locations are more spread out. Cherry blossom is earlier in Vancouver and Washington DC than Kyoto and Liestal; this may due to the differences in the locations and climate features.

For weather data, we summarised daily weather data into accumulated growing degree days (AGDD), first growing days of year (FGDDY), last growing days of year (LGDDY) [1], accumulated freezing degree days (AFDD), first freezing days of year (AFDDY), last freezing days of year (LFDDY), average maximum temperature in Winter (Tmax-W), average maximum temperature in Spring (Tmax-S), average minimum temperature in Winter (Tmin-W), average minimum temperature in Spring (Tmin-S), average precipitation in Winter (PRCP-W), and average precipitation in Spring (PRCP-S).

The correlation across variables is shown in the Figure 4. Several covariates are highly correlated and they were treated with delicate carefulness when we include them in the model. It also shows that the correlation across sites are different, which further confirms that each site needs a different model. We mainly considered covariates that are highly

[1] AGDD is defined as the number of days that the average temperature is above $0°C$. FGDDY is defined by the days between Jan 1 and the first day that the average temperature is above $0°C$. LGDDY is defined by the days between Jan 1 and the last day that the average days is above $0°C$. Similar for the freezing days and date of the year.

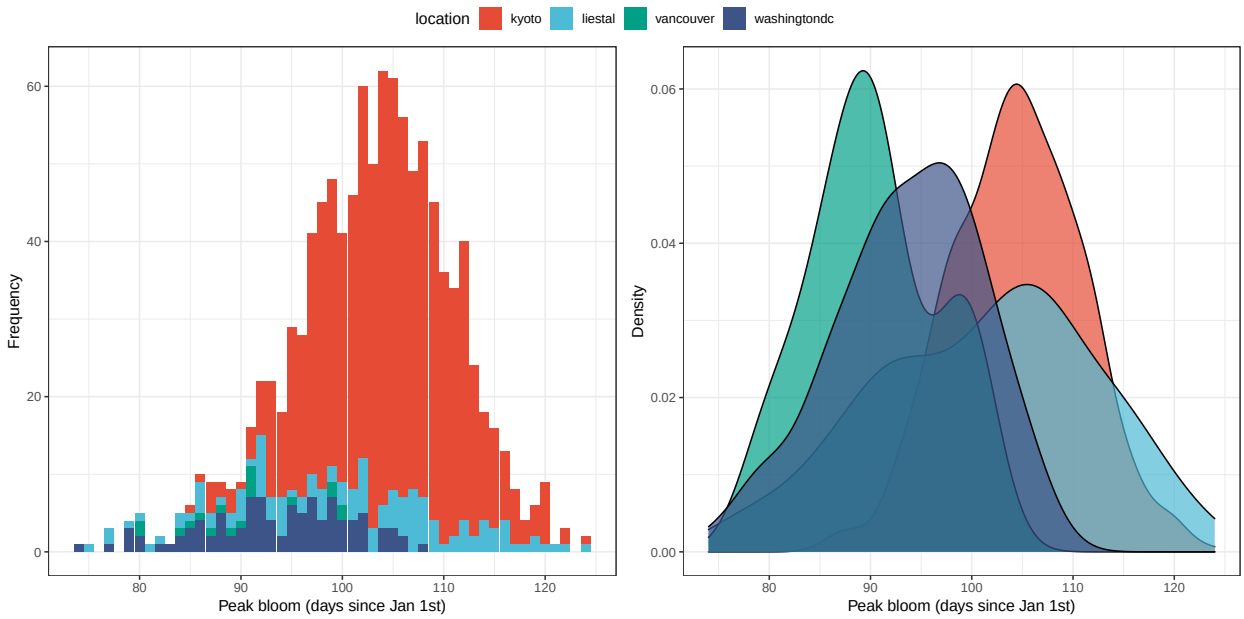

Figure 1: Distribution of peak bloom days across the four sites.

correlated with the bloom days in the models. Additional exploratory data analysis and plots of weather data trends across time are summarised in the Appendix.

2 Methods

Considering the geographical difference, species difference of the cherry trees and the different definition of blossom, we analyze the 4 locations separately. For the three locations: Kyoto, Liestal-Weideli and Washington DC, we utilize a varying-coefficient regression model to illustrate the time-varying relationship between the potential factors and the bloom date of year.

For Vancouver, we extracted additional peak bloom data from the National Park Website. Together with the weather data, these make the Vancouver dataset perfect for the setting of semi-supervised learning, where we have both a small set of labeled data (2004-2021) and a larger set of unlabeled data (1950-2003). The semi-supervised regression co-trained by a set of machine learning models is applied to leverage the valuable information stored in the large amount of unlabeled data to improve the estimation efficiency, see Garcia-Ceja (2019).

2.1 Varying coefficient model

Denote the bloom date of year as y_i, and the possible covariates as X_i's, then the varying coefficient model can be represented as follows:

$$y_i = \beta_0(t_i) + \beta_1(t_i)X_{1,i} + \beta_2(t_i)X_{2,i} + \cdots \beta_p(t_i)X_{p,i} + \epsilon_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

where ϵ_i's are i.i.d. random errors with mean 0 and variance σ^2.

In this way, we take into account the time structure of the data, which is more realistic compared to the traditional multivariate linear regression model. Therefore, our goal is to obtain the estimators of the coefficients $\hat{\beta}_0(t), \hat{\beta}_1(t), \cdots, \hat{\beta}_p(t)$, which further enable us to estimate or predict the bloom date $\hat{y}_i$.

To do this, we considered two nonparametric methods: local linear estimation and spline estimation.

2.1.1 Local linear estimation

For simplicity and efficiency, we consider the Epanechnikov kernel $K(\cdot)$, and let $K_{b_n}(\cdot) = K(\cdot/b_n)$, where b_n is the bandwidth, then the local linear estimator of the coefficients $\hat{\boldsymbol{\beta}}(t) = (\hat{\beta}_0(t), \cdots, \hat{\beta}_p(t))^\top$can be obtained via

$$\left(\hat{\boldsymbol{\beta}}(t), \hat{\boldsymbol{\beta}}'(t)\right) = \underset{\eta_0, \eta_1 \in \mathbb{R}^p}{\arg\min} \left[\sum_{i=1}^{n} \left\{y_i - \mathbf{x}_i^{\mathrm{T}}\eta_0 - \mathbf{x}_i^{\mathrm{T}}\eta_1 (t_i - t)\right\}^2 K_{b_n}(t_i - t)\right]. \tag{2}$$

Generally, as the bandwidth increases, the bias of the estimators increase while the variances decrease, which is also referred to as "bias-variance trade-off". To tackle with this, we use the rule of thumb $b_n = c \cdot n^{-1/5}$ for some coefficient c. By modifying the coefficient, we tried $b_n = 0.1$ and $b_n = 0.15$ separately, and the bandwidth is finalized by choosing the one that minimized the mean absolute error on the training set.

2.1.2 Spline estimation

As for spline estimation, it aims to approximate the coefficients $\beta_i(t)$ by spline functions $S_i(t)$, where $S_i(t) = \sum_{j=1}^{k} P_j(t)$ is a linear combination of the order n spline basis $\{P_j(t)\}_{j=1}^{n}$. The merit of this method lies in that it, unlike the kernel method, provides a uniform expression of the coefficients over the time interval. While it takes effort to choose the appropriate knots, spline order and the basis, the function gam() provides a convenient implementation of the method by automatically selecting the best combinations.

2.2 Machine learning models

Common machine learning methods, such as random forest (RF), and support vector machine (SVM), are also considered and compared with the two nonparametric approaches.

In order to evaluate the performance across methods, we split the data into the validation set (latest 10 years) and the training set (otherwise).

Location	Kyoto				Liestal				Washington			
Model	LL	spline	RF	SVM	LL	spline	RF	SVM	LL	spline	RF	SVM
R^2	0.73	0.67	0.45	0.60	0.88	0.77	0.53	0.75	0.72	0.40	0.14	0.31
Train MAE	1.52	1.85	2.55	1.73	2.84	4.55	6.36	4.00	2.25	3.68	5.65	3.75
Test MAE	3.70	3.00	3.70	3.80	8.40	6.64	6.00	6.36	4.60	5.70	7.50	6.70

Table 1: Comparison of different models on Kyoto, Liestal-Weideli and Washington: Models considered are local linear estimation (LL), spline method (Spline), random forest (RM), and supporting vector machine (SVM).

From Table 1, the nonparametric methods generally outperform the machine learning methods by demonstrating a higher R^2 and better interpretability. Additionally, we can see that a higher R^2 and smaller mean absolute error (MAE) on the training set can be obtained via local linear estimation, which is a direct result of the small bandwidth. However, when it comes to the MAE on the validation set, the spline method enjoys better performance. Considering that the spline method expresses the coefficients as a linear combination of the spline basis, we finally chose spline method as the estimation technique in that it facilitates the prediction process by directly giving us the exact form of the coefficient functions.

2.3 Semi-supervised algorithm for Vancouver

For Vancouver, we incorporated the peak bloom date of Vancouver from 2004 to 2021 and extracted weather information from 1950 to 2021. However, the fully labeled training data only has 17 years of data (2004-2021). Due to the limited sample size and the relatively large number of covariates, training a common statistical or machine learning model with such dataset might result in overfitting, further leading to low out-of-sample accuracy. Therefore, we are considering using the semi-supervised learning techniques to train models on both labeled and unlabeled data. When the weather information contained in the unlabeled data is related to the bloom days, the model estimation efficiency will be improved. In practice, this means we need fewer samples to get a desired estimating standard error.

Following the similar spirit of the Abdel Hady et al. (2009), we implemented the self-learning and co-training by committee semi-supervised algorithms from a set of regressors, including the linear regression, random forest, and support vector machine. We split the data into validation data (latest 5 years) and training data (otherwise), and the semi-supervised model has adequate performance (MAE = 5.4).

2.4 Prediction of the covariates

In order to predict the bloom date of the next 10 years, we also need to first of all give predictions of the covariates. The covariates used in our model are generally related to the temperature in Spring and Winter, which is a non-stationary time series and makes it hard to predict. The approach we considered here is the ARIMA model, where the time structure of the temperature is analyzed and applied to inpute the temperature of the next 10 years.

3 Results

After finalizing the methods, we use all the available data as the training set so as to incorporate all the existing information. For each city, the models give us separate estimates of the coefficients. A summary of the covariates used in each finalized model is provided in Table 2.

Using ARIMA time series model, we obtained the predicted covariates for the next ten years. The particular plots showing the predictions of the average maximum temperature in Spring and the AFDD are demonstrated in Figure 6, see Appendix. As a direct result, the

	Kyoto	Liestal-Weideli	Washington	Vancouver
Year	–	–	✓	–
Tmax-S	✓	✓	✓	✓
Tmin-S	–	–	–	✓
PRCP-S	✓	✓	–	–
Tmax-W	✓	✓	✓	–
Tmin-W	–	–	✓	–
PRCP-W	✓	✓	–	–
AFDD	–	–	–	–

Table 2: Covariates used in each model on Kyoto, Liestal-Weideli, Washington, and Vancouver

predicted bloom dates for the next 10 years are derived by the model and the covariates, as shown below in Figure 2.

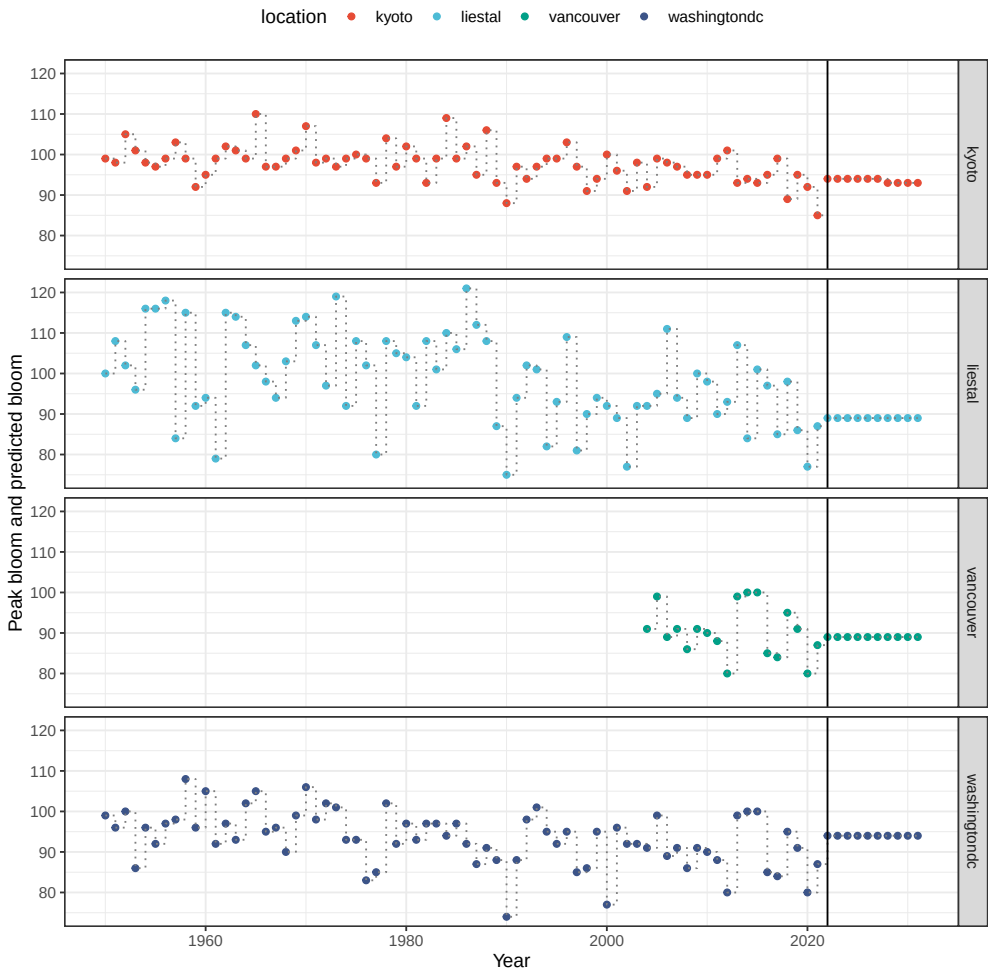

Figure 2: Peak bloom dates of Kyoto, Liestal-Weideli, Washington, and Vancouver: the predicted ones (2022-2031) are presented after the vertical line.

4 Discussion

Although it's extremely difficult to disclose the accurate mechanism of cherry blossom, our study chose appropriate models based on different data features. For the three locations with complete record, varying-coefficient model with a spline method to estimate is easy to interpret and applies to a more general realm compared to the traditional multivariate linear model. While for Vancouver where the data are mostly unlabelled, we tried semi-supervise learning method in order to strike a balance between interpretability and prediction accuracy.

Despite the inevitable limitations due to the intrinsic complexity of the phenological event, the models selected could be improved by increasing the sample size and polishing the prediction of the temperature for the next 10 years. It would also be of interest to incorporate more climate information so as to investigate the possible interplay between the cherry blossom and global warming.

References

Cherry Blossom Festival (U.S. National Park Service). https://www.nps.gov/subjects/cherryblossom/bloom-watch.htm. Accessed: 2022-02-28.

Mohamed Farouk Abdel Hady, Friedhelm Schwenker, and Günther Palm. Semi-supervised learning for regression with co-training by committee. In *International Conference on Artificial Neural Networks*, pages 121–130. Springer, 2009.

Uran Chung, Liz Mack, Jin I Yun, and Soo-Hyung Kim. Predicting the timing of cherry blossoms in washington, dc and mid-atlantic states in response to climate change. *PloS one*, 6(11):e27439, 2011.

Enrique Garcia-Ceja. *ssr: Semi-Supervised Regression Methods*, 2019. URL https://CRAN.R-project.org/package=ssr. R package.

Jean-Michel Legave, I Farrera, Tancrède Alméras, and Michel Calleja. Selecting models of apple flowering time and understanding how global warming has had an impact on this trait. *The Journal of Horticultural Science and Biotechnology*, 83(1):76–84, 2008.

Steffen Moritz and Thomas Bartz-Beielstein. imputets: time series missing value imputation in r. *R J.*, 9(1):207, 2017.

5 Appendix

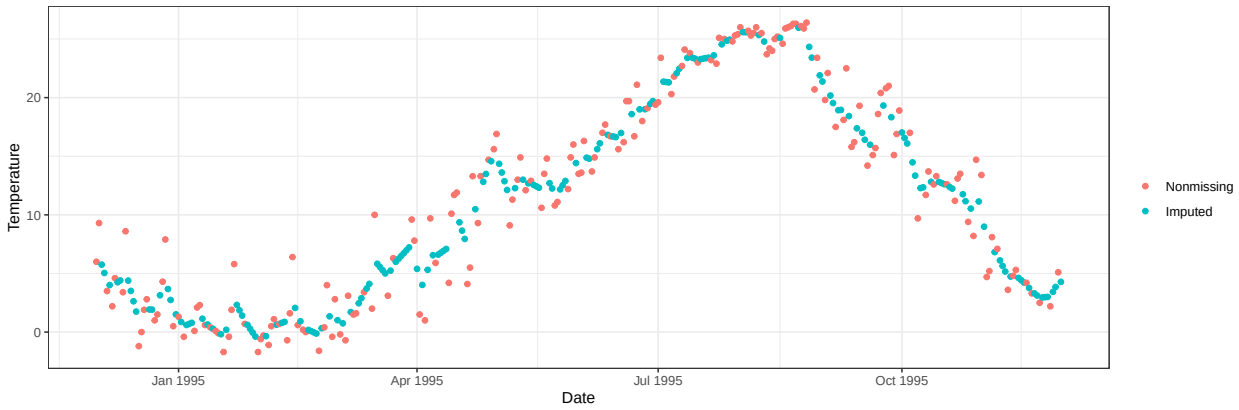

Figure 3: An example of the imputation on missing daily temperature using the Kalman Smoothing on structural time series models (or on the state space representation of an arima model). The figure shows the imputed values (green) and the original non-missing values (red) during 1995 in Kyoto.

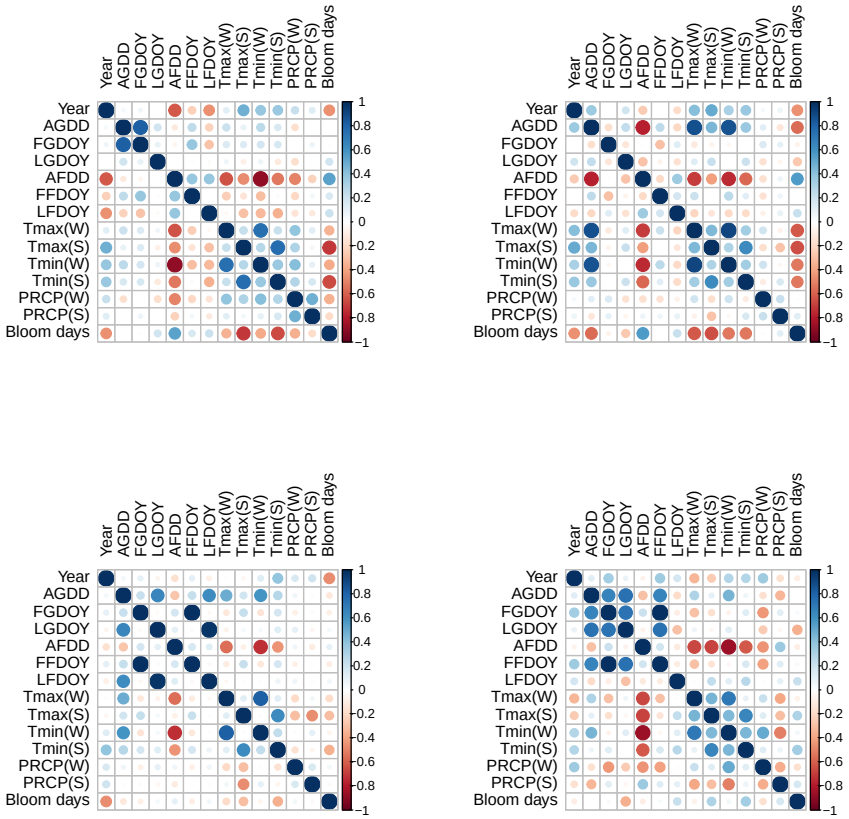

Figure 4: Correlation plots of the variables across four locations. The left top panel is for Kyoto, Japan; the right top panel is for Liestal, Switzerland; the left bottom panel is for Washington DC, US; and the right bottom panel is for Vancouver, Canada.

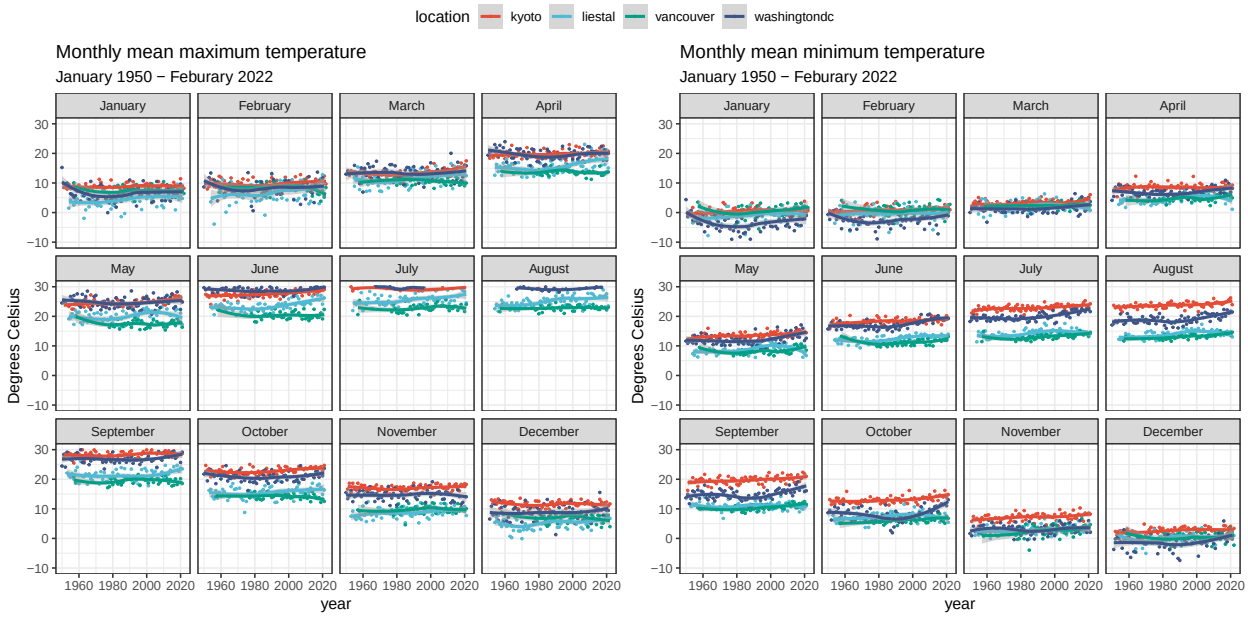

Figure 5: Monthly average maximum and minimum temperature across four site. Due to the different temperature trends, separate models or hierarchical models should be considered for forecasting the bloom dates. We used separate models in our analysis.

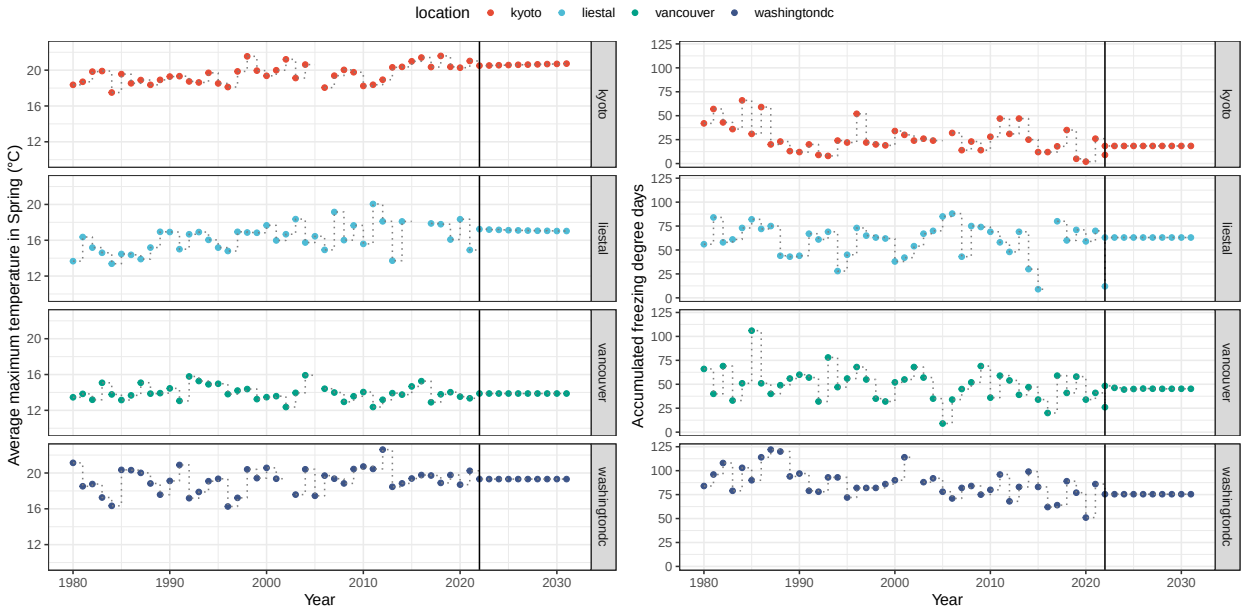

Figure 6: Average maximum temperature and the accumulated freezing degree days in Spring of Kyoto, Liestal-Weideli, Washington, and Vancouver: the predicted ones (2022-2031) are presented after the vertical line.