> "No catalogue of techniques can convey a willingness to look for what can be seen, whether or not anticipated." J.W. Tukey

1 Introduction

In view of the difficulty that exists in making accurate flowering predictions and since the majority of approaches reported in the literature link meteorological data, in the development of my analysis I tried to link complementary sources of information to climatic data and delved into finding behavioral patterns. that revealed discrepancies between countries, between species and between locations. *Exploratory Data Analysis - EDA* was fundamental in my approach, beyond trying to build a good model and making an accurate prediction, I wanted to investigate the underlying relationships that govern the phenomenon.

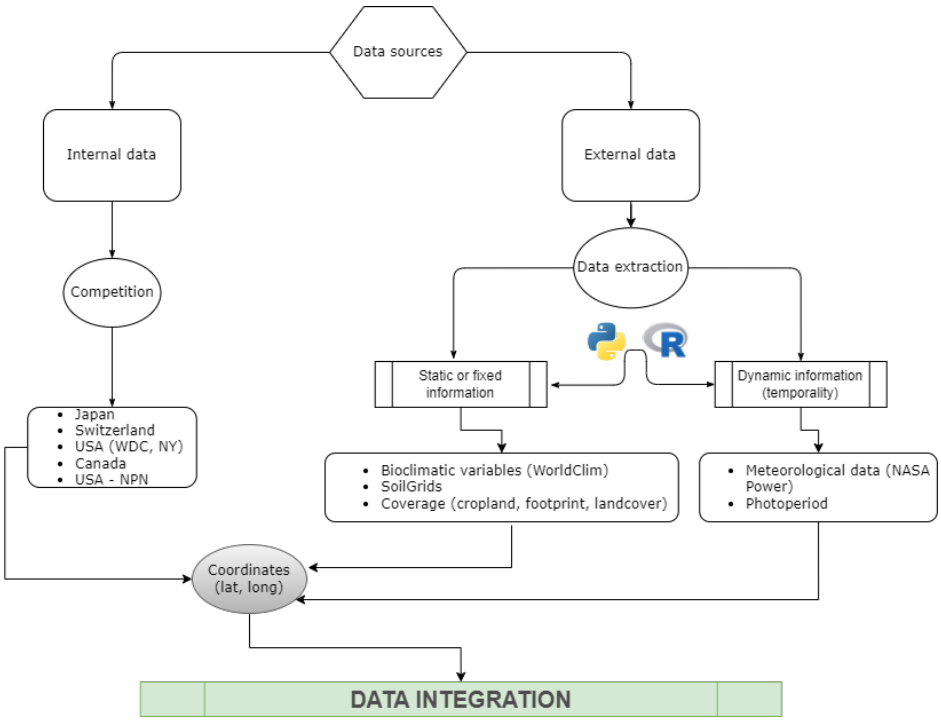
2 Methodology

The methodology I followed to develop the analysis can be divided into three stages:

- **Stage 1:** data extraction and integration. Consolidation of information necessary for EDA and modeling.
- **Stage 2:** exploratory data analysis (EDA). Search for distributional, temporal patterns and interactions.
- **Stage 3:** feature engineering, selection of predictors, construction of models and predictions for the year 2024.

2.1 Data sources

The following diagram shows the data sources, illustrates how the extraction was carried out and finally how the databases were integrated, always taking into account the coordinates (latitude and longitude). I classified the sources of information into *internal data* and *external data*, from the former I used all the data that was provided for the competition, including those provided by *USA National Phenology Network (NPN)*. In the gray circle at the bottom of the image, the coordinates are highlighted as *key* variables or *common identifiers* between the internal and external data. I divide external data sources into *static or fixed* and *dynamic (with temporality)*, the latter change over time.



2.1.1 Static or fixed information

The summary of 19 bioclimatic variables in the years 1970 to 2000 (*static information*) are extracted from WorldClim. A resolution of 30 seconds/0.5 minutes (~1 km2) is used for each country. The extraction of this data was carried out through the library `geodata`. en R. Through the library `geodata` the following information is obtained:

- **Cropland:** Cropland data (*cropland*). This data is derived from ESA WorldCover.
- **Footprint:** "The 'human footprint' is an estimate of direct and indirect human pressures on the environment. Human pressure is measured using eight variables including urbanized environments, population density, electric power infrastructure, cropland, grasslands, roads, railways and waterways. It is expressed on a scale from 0 (low) to 50 (high footprint)." `geodata` library. - Scientific article 2009
- **Land cover:** data was obtained for tree, grassland, shrub and water cover.

In addition to the previous fixed variables, for each coordinate ten soil variables were obtained from SoilGrids. In all of them the averages were extracted for the depth of 0-5 cm, except in the variable `ocs` (*Soil organic carbon stock*) which is only available for the depth of 30 cm.

2.1.2 Dynamic information (temporality)

The meteorological data were obtained with R from the Prediction of Worldwide Energy Resources (POWER) project, using the `nasapower` for extraction of time series (from 1981-01-01 to 2024-02-25) with daily frequency of the following variables:

- Average Earth Surface Temperature (TS)
- Maximum temperature of the earth's surface (TS_MAX)
- Minimum land surface temperature (TS_MIN)
- Evaporation from the earth's surface (EVLAND)
- Frost days (FROST_DAYS)
- Precipitation (PRECTOTCORR)
- Soil moisture profile (GWETPROF)
- Relative humidity at two meters (RH2M)
- Soil moisture in the root zone (GWETROOT)
- The total amount of ozone in a column extending vertically from the Earth's surface to the top of the atmosphere (TO3)

As the POWER project is aimed at three user communities, in this case the Agroclimatology community (**AG**) is used.

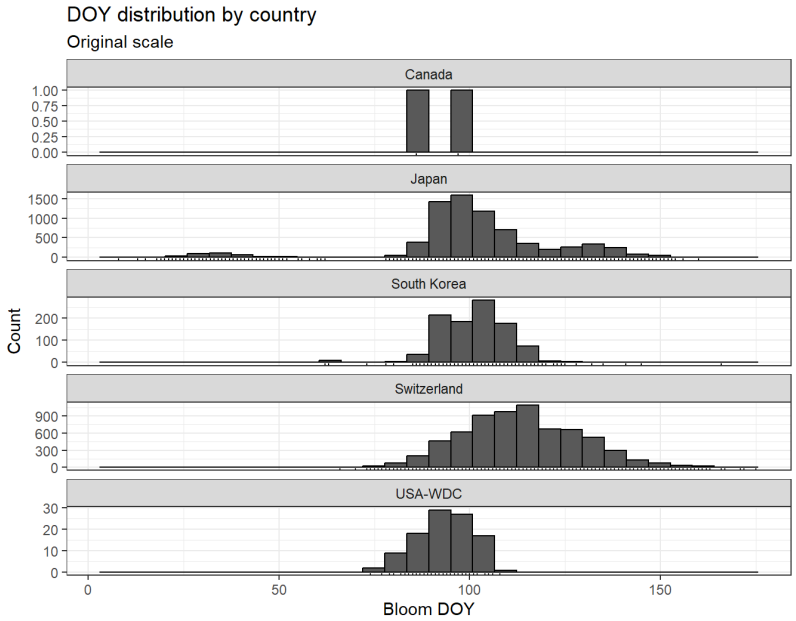
The photoperiod is obtained with the library `meteor` using the `photoperiod()` function, which receives as parameters Enter the date and latitude. In this library the approach proposed by Forsythe et al., 1995. is used.

2.2 Software

With *Python* I executed the information extraction from SoilGrids, with *R* I carried out the rest of the analyzes (data extraction, EDA and modeling).

2.3 Exploratory Data Analysis (EDA)

In the exploratory data analysis I used graphical representations to examine distributional patterns, calculation of correlation coefficients, scatter diagrams with trend lines obtained through *Generalized Additive Models - GAM*, principal component analysis, autocorrelation in the time series of the *DOY* and graphs that highlight temporal trends. Interaction effects between variables were evaluated through scatter diagrams. The distribution of the response variable `bloom_doy` in 5 locations is shown in the following graph.



2.4 Modeling and prediction

2.4.1 Feature Engineering

I create several data sets with **fixed** predictor variables and summary metrics (*average*, *median*, *standard deviation* and *sum*) for the **dynamic** predictor variables (climatic and photoperiod) . All predictor variables are at the coordinate level.

- **`data_predictors_fixed`:** fixed information for each coordinate. In this database, predictors that summarize climate information (*bio_*) from 1970 to 2000 are considered. Cover variables and soil information (*SoilGrids*) are also included..
- **`data_predictors_summary_weather[w]M`:** I generate five databases with summary climate information (10 climate variables) from a time period t_0 to one day before flowering $t_{\gamma-1}$. I summarize with the *average*, *median* and *standard deviation* of all the climatic variables except `FROST_DAYS`, on which I calculate the sum of days with frost and obtain the proportion. I also obtained the sum (accumulation) of temperature, precipitation and evaporation, with these accumulations I calculated a *rate* of *temperature* and *precipitation* by dividing the first by the second, another *rate* between *precipitation* and *evaporation* was calculated by dividing *evaporation* by *precipitation*. For example, if at a coordinate c_i the flowering date t_γ was recorded on "1992-03-24" and a time window $w = 3$ (in months) is selected, then t_0 is "1991-12-25", the summary metrics θ for variables X will take date ranges between *"1991-12-25"* and *"1992-03-23"*. I programmed this whole process in the `FEWeatherSummary()` function and it can be implemented for

any coordinate. To obtain the five databases with a climate summary for months prior to flowering, I used $w = 1, 3, 6, 9, 12$, that is, I summarized the climate for a month, quarter, semester, 3/4 of year and 1 year before flowering.

$$\theta_i = \sum_{t_0}^{t_{\gamma-1}} X_i$$

Where:

$$t_0 = t_\gamma - (30 * w)$$

- **`data_predictors_summary_photo[w]M`:** for the photoperiod I implement the same strategy previously described for the predictors derived based on meteorological information. I built the function `FEPhotopSummary()` for this purpose and the underlying process is the same, choosing a time window w which is used to summarize the photoperiod (X) with the metric θ. I generate five databases with a summary of the photoperiod of months prior to flowering, with $w = 1, 3, 6, 9, 12$.
- **`data_predictors_gdd`:** to calculate the *growing degree-days:* (GDD) I use three approximations. In order to simplify the problem, I use January 1 of each year as the starting date to do the calculations, however, the function `featureEngGDD()` receives an argument named `start_date` (t_0) with which You can choose a different date. I choose the *basal temperature* or limit (t_b) to consider thermal absorption as 5°C, however, the function in question has an argument named `t_base` that allows testing with different values. The daily temperatures x_t (variables `TS`, `TS_MAX`, `TS_MIN`) are used for this calculation. Due to time issues I was not able to experiment enough to estimate optimal values of t_b and t_0. The three ways of calculating GDD are described below (Piña, R. A. et al., 2021; McCaster, G. S. & Wilhelm , W.W., 1997):
  - **Equation 1:** classic approach in calculating GDD (Piña, R. A. et al., 2021). If a basal temperature $t_b = 5$ is exceeded then the difference in degrees Celsius will be taken into account to calculate the accumulated GDD (`AGDD1`).
  - **Equation 2:** Equation 2 (Piña, R. A. et al., 2021) represents the *triangular model of GDD*, this model represents a non-linear triangular function based on temperatures. This equation takes into account the minimum, maximum temperature and an *optimal temperature*. Since I do not have enough information to consider an optimal temperature, I use the same basal temperature with a value of 5°C.
  - **Equation 3:** this equation is another classic approximation (McCaster, G. S. & Wilhelm, W.W., 1997) for the calculation of GDD, where the maximum and minimum temperature is used. In this case *equation 1* coincides with this, however, for the first I use the average temperature directly, in *equation 2* I use the average of the maximum and minimum temperatures. I verified that the results are not the same from the three methods.

$$GDD_{1(x_t)} = \begin{cases} 0\ if\ x_t < t_b \\ x_t - t_b\ if\ x_t \geq t_b \end{cases} \tag{1}$$

$$GDD_{2(x_t)} = \begin{cases} 0\ if\ x_t < t_{min} \\ \frac{x_t - T_{min}}{T_{opt} - T_{min}}\ if\ T_{min} \leq x_t \leq T_{opt} \\ \frac{x_t - T_{max}}{T_{opt} - T_{max}} if\ T_{opt} \leq x_t \leq T_{max} \\ 0\ if\ x_t \geq T_{max} \end{cases} \tag{2}$$

$$GDD_{3(x_t)} = \begin{cases} 0\ if\ \left[\frac{T_{max} - T_{min}}{2}\right] - t_b < t_b \\ \left[\frac{T_{max} - T_{min}}{2}\right] - t_b\ if\ \left[\frac{T_{max} - T_{min}}{2}\right] - t_b > t_b \end{cases} \tag{3}$$

With the previous equations, the $GDD_{(x_t)}$ are obtained for each day that is between the initial date t_0 and one day before flowering $t_{\gamma-1}$, then they are added for each coordinate and the accumulated growth degree days (`AGDD1`, `AGDD2` and `AGDD3`) are obtained.

$$AGDD_i = \sum_{t_0}^{t_{\gamma-1}} GDD_i$$

2.4.2 Feature selection

Given that the number of predictor variables is relatively high and given that it was previously observed in the exploratory analysis that some predictors have a high correlation, I implement lasso regression to select a smaller set of predictors and mitigate multicollinearity problems. Finally, I take this subset of variables into account for building the model. I use the library `tidymodels` to fit the lasso regression model. I adjust the hyperparameters to find the optimal value. The reference book I use to guide the predictor selection strategy is Feature Engineering and Selection: A Practical Approach for Predictive Models by *Max Kuhn and Kjell Johnson* . To validate the precision of the model after selecting the predictors, I use the last flowering year that each coordinate has. Before adjusting the models I select values whose variance is greater than 0.

2.4.3 Modeling

In total I fit three models using two algorithms: lasso regression and XGBoost. Initially I fitted a lasso regression model with all the predictor variables, with this method I managed to select a subset of 48 predictor variables and with these variables I optimized a first model with XGBoost, however, in view of the findings in the exploratory analysis I decided to use a

smaller subset of predictors (24 in total), this third algorithm was used to make predictions for the year 2024 since it showed better results.

All models were fitted with the `tidymodels` library in R, using stratified cross validation (by the response variable) with $k = 5$ and I used the MAE metric to evaluate and compare the models. Data preprocessing included imputation of numerical variables through the median, elimination of variables with zero or close to zero variance, YeoJohnson transformation to correct problems. of asymmetry and the numerical variables were normalized to avoid problems in the variance caused by the units of measurement of each variable.

The confidence intervals were obtained using the residuals through resampling techniques. The confidence level of the intervals is 90% and they were obtained with the percentiles.

3 Results

3.1 Correlations

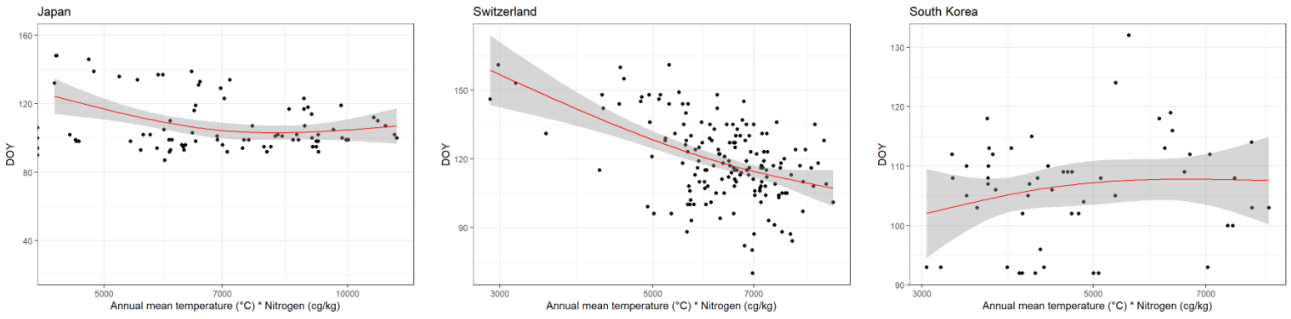
The scatter plots allowed us to show linear and non-linear relationships between some of the bioclimatic variables and the flowering day. Associations were also observed between soil composition, for example, between pH and *DOY*, the relationships are not significant or clear. Temperature and flowering day present a relationship that we could suggest is between linear and quadratic, independent of the country; earlier flowering has been observed at higher temperatures. This behavior is similar in Japan and Switzerland. However, South Korea presents some differences, for example, in the variable *bio11* for Japan and Switzerland. , the quadratic type relationship is evident, but in South Korea it is not. Precipitation also shows interesting relationships with flowering date, with slight differences between countries, for example, annual rainfall (*bio12*) shows an inverse linear relationship with *DOY*, in Switzerland, on the contrary, we observe a quadratic relationship positive and in South Korea the relationship is not clear. This result may suggest that temperature and precipitation are factors that interact and have a joint effect. The seasonality of temperature shows a different pattern of behavior between countries, in Japan the relationship is positive and linear, in South Korea the association is not clear and in Switzerland it shows an inverse quadratic relationship. It was also observed that places where there are more days with frost, the flowering of cherry trees is expected to be slightly delayed.

Regarding soil composition, nitrogen, soil water pH, soil organic carbon content (*SOC*), bulk density (*BDOD*) and cation exchange capacity (*CEC*) show trends in the relationship with *DOY*. Finally, the information on cropland (*CROPLAND*), human footprint (*FOOTPRINT*) and soil cover does not present any type of relationship that could be considered of interest. Furthermore, some of these variables (for example *SHRUBS*) lack information for the coordinates under analysis.

Correlation profiles were constructed for each country and for all countries together. In Japan and South Korea, cherry trees located further north seem to take longer to bloom, the variable *latitude* is the first in the correlation profile of these countries, however, in Switzerland it is not latitude that is the linear factor. The greatest influence is *altitude* followed by soil composition variables and in the top 10 there are three variables that collect precipitation information. In all countries, the annual temperature variable (*bio1*) coincides in being the largest inverse linear factor, followed by other temperature indicator variables. We can say that if we think about the factors that are associated with **delay** in flowering, the profiles between countries are very different, however, if we think about the factors that are associated with **precocity** in the flowering, the profiles between countries are very similar. In conclusion, faster flowering seems to have common and global *"causes"* linked to temperature, however, delayed flowering could be more difficult to establish and generalize.

3.2 Interactions

A first interesting result of the interactions evaluated is that the behavior differs between countries, for example, the interaction *temperature-precipitation* shows an inverse relationship with *bloom_doy*, but it is not linear in the three countries, in South Korea the relationship tends to be quadratic. We can observe something similar in the interaction *temperature-nitrogen*, in Japan and South Korea it does not present any relationship with the target variable, however, in Switzerland an inverse linear relationship is evident. In the three countries the interaction *temperature-apparent density* is negatively related to the response variable, but in Japan the linear dependence that exists between the variable y and the predictor derived from this interaction is evident. The following graph shows the differences for each country.



3.3 Principal Component Analysis (PCA)

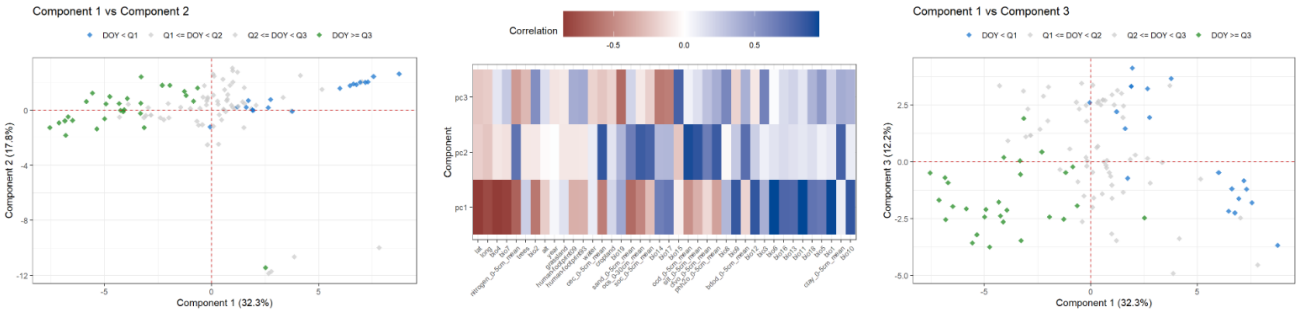
Taking advantage of the fact that there is a high correlation between the predictor variables, I run principal components analysis to reduce the dimensionality and try to see if in a smaller

space of variables a behavioral pattern related to the response variable is exhibited. To facilitate visualizations I discretize the response variable into the following levels (ordinal):

- **`Q1:`** `bloom_doy` values lower than the value of quartile 1: $DOY < Q1$.
- **`Q2:`** values of `bloom_doy` greater than or equal to the value of quartile 1 and less than the value of quartile 2: $Q1 \leq DOY < Q2$.
- **`Q3:`** values of `bloom_doy` greater than or equal to the value of quartile 2 and less than the value of quartile 3: $Q2 \leq DOY < Q3$.
- **`Q4:`** `bloom_doy` values greater than or equal to the value of quartile 3: $DOY \geq Q3$.

I introduce the discretized response variable `doy_categ` to the PCA as a qualitative supplementary variable. - Regardless of the country, the blooms that do not exceed Q1 (`DOY < Q1`) and those that take longer than Q3 (`DOY >= Q3`) exhibit differences when graphed in the first three main components, being located in different places . It is important to mention that the pattern is less visible in South Korea. Blooms that are between Q1 and Q3 tend to be similar. This differentiation seen between early flowering and late flowering could be indicative of differential profiles in the ecological niche or environmental environment to which cherry trees are exposed. In the case of Japan, it is observed that component 1 is positively associated with variables that collect information on temperature (bio1, bio5, bio10, bio11) and precipitation (bio12, bio13, bio16), indicating that values greater than zero in We can associate component 1 with high values of temperature and precipitation, precisely under these conditions we observe rapid flowering (blue dots), on the other hand, this component is negatively associated with latitude, longitude, bio4 (seasonality of the temperature), bio7 (annual temperature range) and to a lesser extent with nitrogen, indicating that values greater than zero of this component will be related not only to high values of temperature and precipitation but also to less variation in temperature (bio4) and lower annual temperature range (bio7), therefore, the described profile is the one that applies to early flowering, just the opposite occurs with late flowering, where low temperatures and precipitation are expected, with greater variations in temperature throughout the year , we could also affirm that plants that are located further south in Japan tend to have delayed flowering. Due to the location of the points we could also affirm that early flowering in Japan is modulated by changes in temperature, however, other factors such as cation exchange capacity (*cec*) and organic carbon density (*ocd*) could be influential, since it is observed that these variables have a positive correlation with component 2, indicating that values greater than zero in component 2 are associated with soils with high cation exchange capacity and soils with greater density of organic carbon. In Switzerland, the difference between early flowering and late flowering is also evident; however, the correlation of the components with the variables allows us to infer behavioral patterns that are slightly different from those in Japan and South Korea. Component one has a positive correlation with variables that collect temperature information (bio4, bio5, bio10, bio11) and exhibits a negative correlation with altitude, soil organic carbon content (*soc*), soil organic carbon inventory (*ocs*) and to a lesser extent with precipitation variables (bio12, bio16, bio17), these correlations allow us to understand that early blooms occur in conditions of high temperature (also the annual variation) and in places where there is less rainfall annual, lower soil organic carbon

content and located at low altitudes. Cherry trees that exhibit late blooms tend to be located at high altitudes, sites with greater amounts of annual precipitation, higher soil organic carbon contents. In South Korea, a behavior similar to that of Japan and Switzerland is observed in that temperature is the determining factor in flowering times. However, component 1 is negatively associated with variables that collect temperature information (bio1, bio6, bio9, bio11) and positively with variables that collect information on the annual variation in temperature (bio4, bio7). Late blooms are located to the right of 0 in component 1, that is, these plants are expected to be in places with greater variations in temperature throughout the year and the temperatures of these sites are also expected to be lower. Based on the positive association that component 3 has with precipitation (bio12, bio15, bio16) and observing the location of the green dots in the graph of CP1 vs CP3, it is possible to intuit that delayed flowering is also associated with greater precipitation.



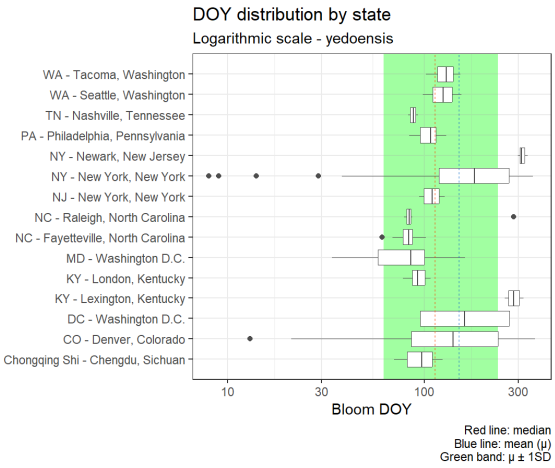
3.4 NPN

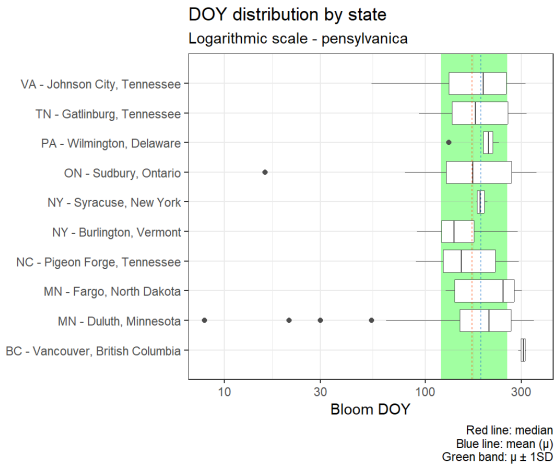
NPN data reflect high variability in cherry blossom depending on location and species. The following graph shows the distribution of `bloom_doy` by state for the species *Yedoensis* and *Pensylvanica*. The distribution of the species *Pensylvanica* is much more homogeneous compared to *Yedoensis*.

3.5 Distribution of days between flowering

To calculate the difference (in days) between flowering t_i with t_{i-1} I take into account coordinates that meet the following characteristics:

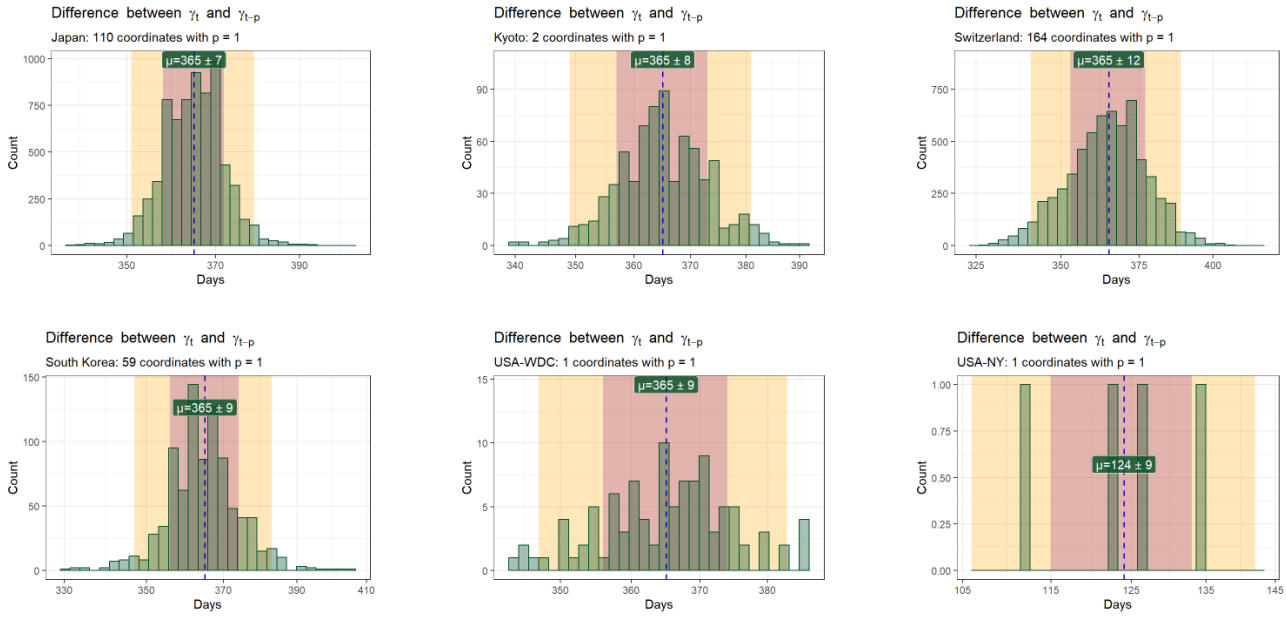
- Minimum 2 records ($n > 1$)
- Difference between years equal to 1, that is, I only keep coordinates whose blooms have been reported with an annual frequency, it does not matter if it was 20 years ago, the important thing is that there is information year after year.





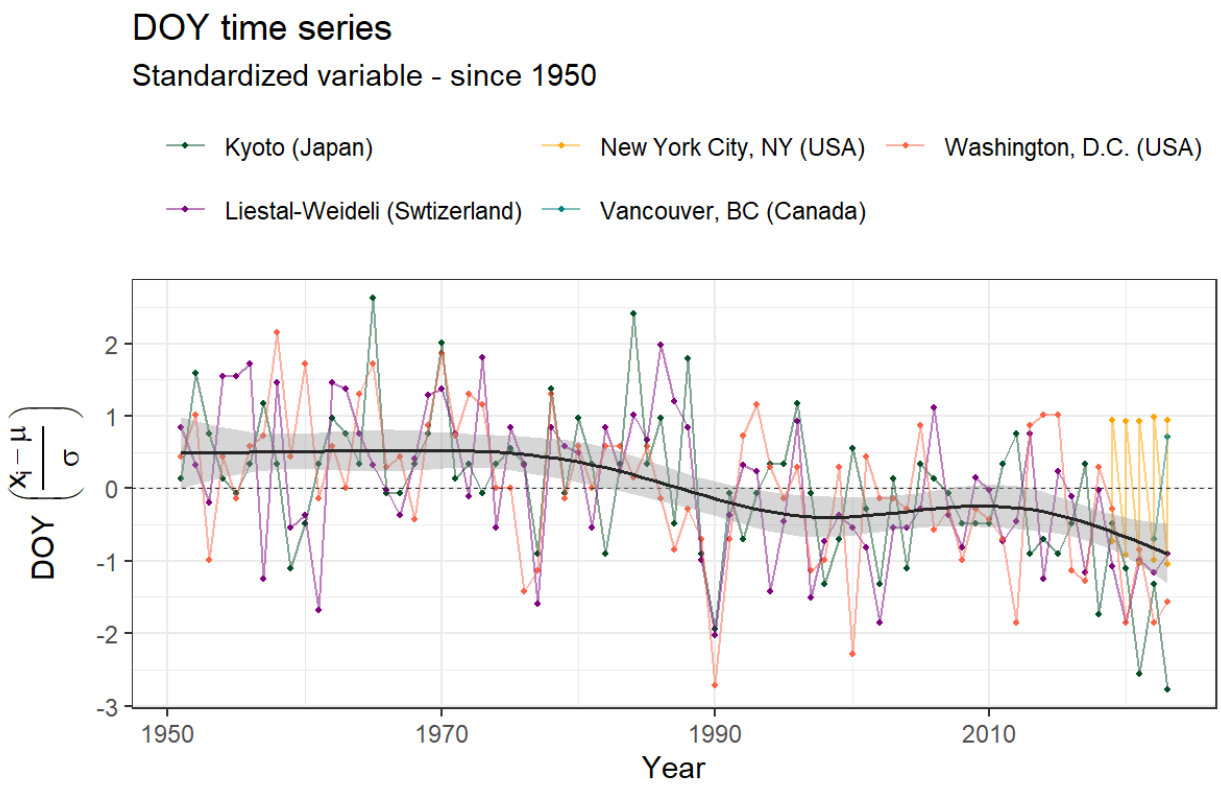
If we call γ the bloom date (`bloom_date`) that has measurements in a year t, with $t = 1, 2, ..., k$, where k is the number of records that meet the two restrictions described previously, the distribution shown in the following graphs is the difference of γ_t with γ_{t-1}, that is, the lag operator with $p = 1$ on the difference in days between flowerings. It is important to mention that the user will be able to choose the lag value to generate (γ_{t-p}), in which case for the first condition n is expressed as $n = p + 1$; The second condition remains the same regardless of the value of p.

These results exhibit a symmetrical distribution centered on 365 days with a standard deviation that ranges between 7 and 12 days, which means that regardless of whether cherry trees delay or advance their flowering, they tend to bloom with an annual frequency with few variations throughout weather. With more time I would like to take advantage of this result, I think that based on these findings we could delve a little deeper, for example, in Monte Carlo simulations based on these probability distributions.



3.6 Temporal trends

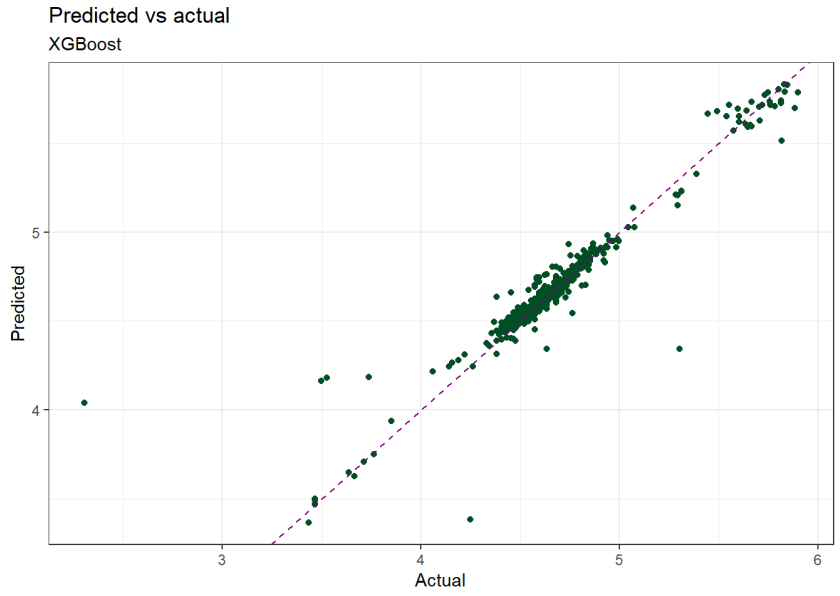
In the following graph you can see the trend of the last 30 years regardless of the site of interest. To facilitate comparison, the variable *GIVE* was standardized with $mu = 0$ and $\sigma = 1$. Since 1990, few blooms have exceeded the average (value of zero).



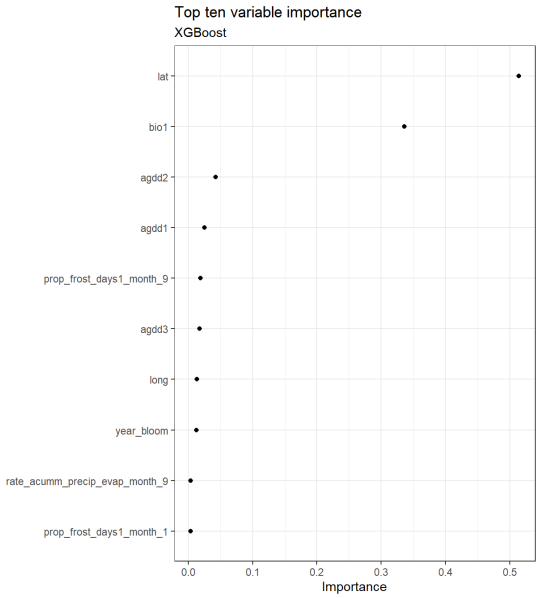
3.7 Modelos y predicción

The MAE on the test set for the lasso regression model was 6.81 days, for the first XGBoost model it was 2.88 days, and for the third XGBoost model it was 5.35 days. In this last model, the response variable was transformed through the natural logarithm and at the end the inverse transformation was applied to return to the original units. The second XGBoost model showed overfitting and for that reason I did not choose to make predictions with it.

The following graph shows the relationship between the actual values and those predicted by the final model.



The 10 most important variables for this model are shown in the following graph:



The table with the final predictions for the year 2024 is shown below:

location	prediction	lower	upper

kyoto	102	98	105
liestal	114	110	117
vancouver	97	94	100
washingtondc	106	103	110
newyorkcity	105	101	108