Peak Bloom Prediction by Linear Regression and Model Selection

For this project, our group predicted "cherry blossom peak bloom" with linear regression in 3 different approaches. And, to evaluate the prediction model's accuracy, we used the Akaike information criterion and cross-validation. This narrative introduces 3 approaches and illustrates the limitation of the models.

Approach 1

Our first attempt for the prediction includes temperature data. To use this data, missing data should be handled. So our group fit the regression with a generalized linear model. And the result shows that there is a high correlation between `temperature` and `bloom_doy`. However, in 2023, `washingtondc` and `liestal` do not have temperature data for the last 4 months. To solve the problem, our group considered using data imputation, but as it can be biased, we decided to drop this variable.

Approach 2

As the temperature can be explained with sunlight conditions, we tried to compensate for the effect of missing temperature data by specifying trigonometric conversion of year and latitude data. As there is a periodic phenomenon called the solar cycle with 11 years, we added $\text{year}_{\sin} = \sin(\pi \cdot \frac{\text{year}}{5.5})$ and $\text{year}_{\cos} = \cos(\pi \cdot \frac{\text{year}}{5.5})$ terms so that they can explain some portion of the cycle. Similar to this, the intensity of sunlight is proportional to cos(latitude). The final model was selected by the Akaike information criterion (AIC), however, the criteria for this competition is a mean absolute error (MAE). So we assumed that different model fit methods can be even beneficial in this case.

Approach 3

Instead of using AIC, we used cross-validation (CV) that measures MAE. (specifically, leave-one-out cross-validation was used.) Also, to use this function for model selection, we defined a stepwise model-selecting function for MAE-based CV. Considering the characteristic that `/data/vancouver.csv` has only 1 row, it is impossible to use `location` as a categorical variable. So we dropped the `location` variable for the final model.

Discussion

The final model obtained by "Approach 3" was `bloom_doy ~ year + long + year_sin`. That is quite close to the model from "Approach 2", which was `bloom_doy ~ year + location + year_sin`. So it demonstrates that criteria for AIC and MAE-based CV can provide similar error measures for the fitted model.

Regarding limitation of the "approach 3", it still uses the `glm` function, which is based on the least square method. So, using an algorithm with the least absolute deviation (LAD) can improve the error term in this analysis. However, there are more than 1 possible algorithm and solution for LAD, which is decided depending on the situation. So fitting the model with LAD requires more attention, as it can be biased as well.

Also, the Vancouver data set does not have enough rows to make `species` inferences. In other words, this analysis was not able to find a correlation between `species` and `bloom_doy`. Considering the above 2 models, `location` or `long` variables possibly reflects the effects of different species.

Conclusion

Even though there were many factors that made our model unstable, we could find the significance of `year` itself and the cycle of it toward `bloom_doy`. By using this knowledge, our group could make a linear model for prediction.