Analyzing Sakura Flowering Trends Under Climate Change*

Predicting Bloom Dates Using Historical Records and Temperature Data

Mariko Lee

December 2, 2024

This study uses historical and modern data to analyze the impact of temperature and long-term trends on sakura flowering in Japan. Linear regression highlights a significant shift toward earlier flowering dates, while generalized additive models (GAMs) capture non-linear relationships. Model comparison via AIC favors the GAM, showing its effectiveness in handling complex patterns. Predictions under future climate scenarios (+1°C, +2°C, +3°C) indicate flowering dates may advance by 5–15 days by 2070. These results underscore the sensitivity of sakura flowering to climate change and its broader ecological and cultural implications.

1 Introduction

In Japan, cherry blossoms, or "sakura," have been a significant cultural and historical symbol, representing the transient beauty of life. The practice of "hanami," or flower viewing, dates back centuries. During the Heian period (794-1185), aristocrats would gather to appreciate the blossoms' fleeting beauty while composing poetry (Japan National Tourism Organization (n.d.)). Hanami has developed into a deeply ingrained cultural tradition that transcends social classes, with its economic significance growing in modern times. Recent research by Katsuhiro Miyamoto indicates that the economic impact of hanami in Japan is expected to double this year, highlighting its role as a cultural event and an economic activity (Kaneko (2024)).

However, a concerning trend has surfaced in recent decades: sakura blossoming earlier each year, attributed mainly to rising global temperatures. According to The Economist, the earlier blooming of cherry blossoms is closely tied to climate change, raising concerns about the long-term implications for Japan's cultural heritage and biodiversity (The Economist (2017)). This shift in flowering trends serves as a reminder of humanity's impact on the environment and a

*Code and data are available at: https://github.com/leemarik/sakura-flowering-trends.git.

pressing scientific challenge to understand the intricate relationships between climate variables and flowering dates.

The estimand of this analysis is the relationship between average March temperatures, year, and the Sakura flowering dates. By analyzing centuries of historical and modern sakura data, my goal is to determine how changes in temperature and long-term climate trends have impacted the timing of sakura blooming. Additionally, I aim to forecast future flowering dates under various climate scenarios, providing insights into the potential effects of climate change on this culturally significant phenomenon.

Despite the historical documentation of sakura flowering trends, there is a critical gap in understanding how these changes will evolve in the future under various climate scenarios. This analysis addresses this gap by using historical and modern data on sakura flowering dates and temperatures to determine the impact of climate change on flowering dates. I used statistical modeling techniques of linear regression and generalized additive models (GAMs) to explore the relationship between temperature, year, and flowering dates. These models provide insights into past trends and enable predictions for future scenarios under different temperature increases.

The findings indicate that rising temperatures strongly correlate with earlier blooming dates, which have accelerated over the last century. By modeling these patterns, the analysis offers a framework for predicting future sakura blooming dates and assessing the broader implications of climate change on cultural and ecological systems. This research contributes to scientific understanding and highlights the urgency of climate mitigation efforts to preserve cherished cultural practices like hanami.

The structure of this paper is as follows: Section 2 reviews the data sources and preprocessing methods used in this analysis. Section 3 details the modeling approaches, specifically Linear Regression and Generalized Additive Models. Section 4 presents the results, followed by a discussion in Section 5 on the implications, limitations, and future directions.

2 Data

2.1 Overview

This study uses the statistical programming language `R` (R Core Team 2023) to analyze sakura flowering patterns across Japan over the centuries, integrating three datasets obtained from Alex Cookson's repository (Cookson 2020). These datasets include historical Sakura flowering records (812-2015) from Kyoto, modern Sakura flowering data (1953-2019) from across Japan, and temperature data sourced from the Japan Meteorological Agency (Agency 2024).

The historical dataset includes reconstructed flowering dates based on Kyoto hanami records and paleoclimatological studies, with estimated March temperature. The modern dataset provides nationwide observations of flowering and full bloom dates alongside geographic and temperature data. Temperature data for March was sourced directly from Japan Meteorological Agency's climate statistics.

The three datasets were cleaned, standardized, and merged into a unified structure. Missing values were removed, flowering dates were converted into day-of-year (DOY) format, and March temperatures were imputed using reconstructed values where necessary. New variables such as flowering range (early, mid, late) and decade grouping were created for this analysis.

Key R packages used include `tidyverse` (Wickham et al. 2019), `lubridate` (Grolemund and Wickham 2011), `dplyr` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2024), `arrow` (Richardson et al. 2024).

2.2 Measurement

The dataset used in this study was sourced from Alex Cookson's repository, which integrates historical and modern sakura flowering data with temperature observations. The historical data originate from Kyoto, with records spanning 812 to 2015 CE, compiled from diary entries, literary works, and paleoclimatology studies. The modern dataset, spanning 1953 to 2019, was obtained from the Japan Meteorological Agency and includes data collected across Japan, such as flowering and full bloom dates, geographic locations, and observed temperatures. The integrated dataset consists of 3,526 observations after cleaning and merging.

The cleaned dataset comprises variables such as: - `Year`: The year of observation. - `Flowering Date`: The calendar date of initial sakura flowering, converted to day-of-year format. - `Average Temperature`: March temperatures either observed or reconstructed (historical). - `Day of Year`: Numerical representation of the flowering date within the year. - `Decade`: Grouped decade for temporal trends. - `Flowering Range`: Categorical classification of flowering timing (Early, Mid, Late). - `Source`: Indicates the origin of the data (modern or historical).

Historical temperature data for Kyoto were reconstructed using paleoclimatology techniques, providing critical context for trends in sakura flowering dates. While this reconstruction supports long-term trend analysis, it introduces potential biases due to variations in methodology and accuracy over time. Similarly, modern data include direct observations of March temperatures and flowering times recorded at multiple locations, offering broader geographic coverage but differing from the localized Kyoto records in the historical dataset.

To ensure comparability, the flowering and temperature data were transformed into consistent numeric formats, allowing analysis of trends over time. Missing or incomplete entries were removed, and observed temperatures were prioritized over reconstructed values when available. The final dataset was prepared for analysis with derived variables to capture temporal and categorical trends in flowering dates.

2.3 Limitations

Although the dataset allows long-term sakura flowering trend analysis, several limitations must be noted:

- Reconstructed Data: Historical temperature data for Kyoto rely on paleoclimatological reconstructions, which may introduce biases or inaccuracies due to methodological differences.
- Geographic Variability: The historical data is Kyoto-specific, while modern data represents nationwide observations. Differences in regional climate conditions may impact direct comparisons.
- Temporal Gaps: The historical dataset spans over a millennium but includes periods with sparse or no data, particularly for earlier centuries.
- Consistency of Definitions: Variations in how flowering dates were recorded historically versus modern observation standards may affect the accuracy of trend analysis.

Despite these limitations, the dataset analyzes the relationship between climate change and sakura flowering patterns across Japan.

3 Model

My modeling process aimed to predict the sakura flowering dates (DOY) under varying climate conditions and identify critical factors influencing flowering patterns over time. To achieve this, I used both linear regression and generalized additive models (GAMs) to capture relationships between sakura flowering dates, temperature, and time. These models were implemented in `R` using the `lm` (Gelman, Hill, and Vehtari 2020) and `mgcv` (Hastie and Tibshirani 1990) packages.

Given the historical and modern data, capturing variability and trends needed the consideration of linear and non-linear relationships. While the linear regression model offered a straightforward approach, the GAM allowed flexibility in accounting for complex, non-linear trends. Both models were evaluated to ensure a balance between simplicity and accuracy.

3.1 Linear Regression Model

The first step in my analysis was building a linear regression model to establish baseline relationships between flowering dates, average March temperatures, and year. This model was expressed as:

$$\text{day\_of\_year}_i = \beta_0 + \beta_1 \cdot \text{avg\_temperature}_i + \beta_2 \cdot \text{year}_i + \epsilon_i$$

Where: - day_of_year_i: Day of Year for the i^{th} observation (flowering date). - avg_temperature_i: Average March temperature. - year_i: Year of observation. - $\beta_0, \beta_1, \beta_2$: Coefficients for intercept, temperature, and year, respectively. - ϵ_i: Random error term.

This model provided a starting point to explore how flowering dates shifted over time. The coefficients revealed the effect of each variable, while residual diagnostics were used to evaluate the model's assumptions, such as linearity and homoscedasticity. Despite its simplicity, the linear model struggled to capture the non-linear trends evident in the dataset.

3.2 Generalized Additive Model (GAM)

To address the limitations of linear regression, I employed a generalized additive model (GAM) to capture non-linear relationships between the predictors and flowering dates. The GAM was implemented using the `mgcv` package in R with the following expression:

$$\text{day\_of\_year}_i = \beta_0 + s(\text{avg\_temperature}_i) + s(\text{year}_i) + \epsilon_i$$

Where $s(\text{avg\_temperature}_i)$ and $s(\text{year}_i)$ are smooth functions capturing non-linear effects of temperature and year. These smooth terms were estimated using splines, allowing the model to adjust flexibly to changes in the dataset over time.

This approach offered several advantages, including flexibility, where smooth terms allowed the model to adapt to non-linear relationships, robustness in terms of the REML method, ensuring reliability in estimating smoothness in parameters, and interpretability by visualizing smooth terms, where I was able to identify how temperature and year independently influenced flowering dates.

3.3 Model justification

The goal of this analysis was to model and predict the sakura flowering dates and explore how they are influenced by temperature and long-term temporal trends. To address this, I combined linear regression and GAM model to reflect the goal of balancing interpretability while capturing non-linear trends.

The linear regression model was chosen as the baseline to provide a straightforward interpretation of the relationships between sakura flowering dates, average temperature, and year. By incorporating temperature and year as predictors, the model captured the direct effect of warming trends and temporal changes on flowering dates. However, its simplicity limited its ability to capture non-linearities of the data.

The generalized additive model was then used to address the limitations of the linear model, precisely its inability to model non-linear trends. The smooth terms in GAMs offered the flexible modeling of the relationship between predictors and the response variable. GAMs are particularly well-suited for ecological datasets, where non-linear relationships are common due to factors such as thresholds and seasonal variations. The inclusion of smooth terms for temperature s(Temperature) and year s(year) model to adapt to varying relationships across the dataset without overfitting.

By comparing the two models, the GAM was a better choice for capturing the complex relationship of sakura flowering trends, as evidenced by its superior fit and lower AIC.

3.4 Model Evaluation and Validation

To ensure the reliability of the models, I performed the following evaluation and validation steps:

Model Performance Metrics: For the linear regression model, the adjusted `R^2` was 0.0136, indicating a weak fit to the data. The residual standard error was 28.18 days, highlighting high variability in flowering dates not explained by the model. The GAM improved the adjusted `R^2` to 0.0411, explaining 4.36% of the deviance. While small, this improvement shows the GAM's ability to capture non-linear trends lacking in the linear model.

Diagnostics: Residual plots showed a non-random pattern, suggesting the presence of a non-linear relationship and the need for a more flexible model in the linear regression model. For the GAM, the residual checks confirmed model convergence, with all smooth terms for

`avg_temperature` and `year` showing significant contributions ($p<0.001$). The basis dimension tests indicated no under-smoothing, making sure that the model captured the complexity of the data.

Model Comparison: The Aikake Information Criterion (AIC) favored the GAM (AIC = 33466.71) over the linear model (AIC = 33555.97), confirming the GAM's superior fit. The GAM's flexibility allows non-linear trend modeling, providing a more accurate understanding of the relationships in the data.

Validation: The dataset was split into training and testing subsets to evaluate predictive performance. The dataset was split into a training set (80%) and a testing set (20%). Both models were fitted to the training data and evaluated on the testing data using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics.

The GAM slightly outperformed the linear regression model, as indicated by lower RMSE and MAE values for both training and testing datasets. This aligns with its ability to capture non-linear relationships between predictors and sakura flowering dates. Both models presented similar errors on training and testing datasets, suggesting good generalization with no evidence of overfitting. The GAM's better performance, especially on testing data, validates its flexibility in modeling complex relationships.

3.5 Future Predictions

To explore the impact of climate change on sakura flowering dates, the GAM was extended to simulate hypothetical future scenarios. These scenarios predict the timing of sakura flowering under varying levels of temperature increases over the next 50 years. This approach provides critical insights into the sensitivity of sakura phenology to climate change.

Future predictions were based on three hypothetical warming scenarios: - +1°C: A moderate temperature increase. - +2°C: A significant increase in average temperature. - +3°C: A severe climate change scenario.

The scenarios were created by extrapolating 50 years ahead of the latest recorded year in the dataset and incrementing the average March temperature by 1°C, 2°C, and 3°C. The GAM was then used to predict flowering dates for each scenario.

4 Results

5 Discussion

Appendix

References

Agency, Japan Meteorological. 2024. *Japan Meteorological Agency | Tables of Monthly Climate Statistics*. *Jma.go.jp*. https://www.data.jma.go.jp/obd/stats/etrn/view/monthly_s3_en.php?block_no=47401.

Cookson, Alex. 2020. *data/sakura-flowering at master · tacookson/data*. *GitHub*. https://github.com/tacookson/data/tree/master/sakura-flowering.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari. 2020. *Regression and Other Stories*. Cambridge University Press. https://avehtari.github.io/ROS-Examples/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Hastie, Trevor, and Robert Tibshirani. 1990. *Generalized Additive Models*. 1st ed. Boca Raton: Chapman; Hall/CRC.

Japan National Tourism Organization. n.d. "Sakura History: The Story Behind Japan's Cherry Blossoms." https://www.japan.travel/en/au/experience/cherry-blossoms/sakura-history/#:~:text=For%20many%20Japanese%2C%20the%20blooming,watching\XeTeXglyph\numexpr\XeTeXcharglyph"0027\relax{}%20parties%20known%20as%20hanami.

Kaneko, Karin. 2024. *Economic impact of hanami expected to double this year*. *The Japan Times*. https://www.japantimes.co.jp/news/2024/03/15/japan/society/hanami-economic-impact/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

The Economist. 2017. "Japan's Cherry Blossoms Are Emerging Increasingly Early." *The Economist*. https://www.economist.com/graphic-detail/2017/04/07/japans-cherry-blossoms-are-emerging-increasingly-early.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.