Datasheet for 'Analyzing Sakura Flowering Dates and Climate Trends'*

A Phenological Study Using Historical and Modern Data

Mariko Lee

3 December 2024

This datasheet documents a dataset on sakura flowering dates and associated climate variables. The dataset integrates historical and modern records to examine phenological shifts driven by climate change, analyzing the impact of rising temperatures on sakura flowering patterns.

Extract of the questions from Gebru et al. (2021).

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*

    - The dataset was created to enable analysis of sakura flowering trends. It addresses the gap in structured, longitudinal data combining historical and modern phenological records.

2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*

    - The `sakura-historical` dataset is from Prof. Yasuyuki Aono from Osaka Metropolitan University and `sakura-modern` and `temperature-modern` dataset is from the Japan meteorological Agency.

3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*

    - The dataset relies on publicly available historical records and institutional data, requiring no external funding.

4. *Any other comments?*

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

- The dataset are retrieved from Alex Cookson [https://github.com/tacookson/data/tree/master/sakur flowering] GitHub.

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

    - Each instance represents a sakura flowering observation, along with corresponding temperature data and metadata (e.g., year, source).

2. *How many instances are there in total (of each type, if appropriate)?*

    - The dataset contains 3,526 observations spanning over 1,200 years.

3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*

    - The dataset includes historical records for Kyoto and modern observations across Japan. While not exhaustive, it provides broad geographic and temporal coverage.

4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*

    - Instances consist of variables such as flowering date (doy), full bloom date, monthly mean temp, and geographic coordinates (latitude, longitude)

5. *Is there a label or target associated with each instance? If so, please provide a description.*

    - The primary target variable is flowering date `day_of_year`.

6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*

    - Some March temperature data is unavailable for historical observations due to gaps in records.

7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- The dataset links flowering dates to climatic variables and geographic regions.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*

    - The dataset is split into 80% training and 20% testing subsets for modeling purposes.

9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*

    - Historical data may include biases from estimation methods. Modern data may reflect urban heat island effects.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

    - Data sources include: Prof. Yasuyuki Aono's website (Yasuyuki (2015)), Japan Meteorological Agency (JMA) phenological and climate statistics (Agency (2024)), NOAA's National Centers for Environmental Information (National Centers for Environmental Information (NCEI) (2024))

11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*

    - No, the dataset is public

12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*

    - The dataset focuses on ecological phenomena and does not contain sensitive personal data.

13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*

    - Historical data is limited to Kyoto, while modern data covers a broader geographic range.

14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*

- No personal information is included in the dataset.

15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*

    - The dataset does not raise ethical concerns or pose risks of harm or offense.

16. *Any other comments?*

    - No

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

    - Historical flowering data was extracted from Prof. Yasuyuki Aono's studies and modern data from JMA records, collected by Alex Cookson.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*

    - Data was collected manually and through automated extraction methods.

3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*

    - Comprehensive for Kyoto historical data and geographically representative for modern Japan.

4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*

    - Historical data was compiled by researchers; modern data was sourced from JMA records.

5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*

- The historical dataset spans 812–2015, and the modern dataset spans 1953–2019.

6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - No review was necessary as the dataset does not involve human subjects or private information.
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - No, data was obtained from publicly available research and meteorological records.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - N/A; the dataset does not involve human participants.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - N/A
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - N/A
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - The dataset supports climate research without risks to individuals or groups.
12. *Any other comments?*
    - No

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*

- Missing values were filtered, and flowering dates were standardized into `day_of_year` format for consistency. Historical estimates were aligned with modern records where possible.

2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*

   - The cleaned dataset is available in `.parquet` and `.csv` formats. It is avialable at [https://github.com/leemarik/sakura-flowering-trends/tree/master/data/01-raw_data]

3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*

   - R was used for preprocessing, employing packages like `tidyverse`, `lubridate`, and `arrow`. Data Cleaning script is available at [https://github.com/leemarik/sakura-flowering-trends/blob/master/scripts/03-clean_data.R]

4. *Any other comments?*

   - No

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*

   - The dataset has been used to model trends in sakura flowering dates using linear regression and generalized additive models (GAMs).

2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*

   - Yes, the dataset and analysis code are hosted on GitHub at [https://github.com/leemarik/sakura-flowering-trends]

3. *What (other) tasks could the dataset be used for?*

   - It can be used for climate forecasting, ecological impact analysis, and cultural studies.

4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*

- Dataset consumers should account for biases in historical data and the limitations of temperature proxies.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - No
6. *Any other comments?*
   - No

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - The dataset is publicly available at [https://github.com/leemarik/sakura-flowering-trends.git]
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
   - Available via GitHub as `.csv` and `.parquet` files with documentation.
3. *When will the dataset be distributed?*
   - Distributed upon project completion.
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
   - No. MIT License
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
   - No
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
   - No
7. *Any other comments?*

- No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be maintained by Mariko Lee during the project term.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - For inquiries, contact mariko.lee@mail.utoronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - No
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - N/A
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - No, as the dataset is updated. However, older version will be accessible through update history through GitHub.
7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - Contributions are welcome via GitHub pull requests.
8. *Any other comments?*
    - No

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.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, and Kate Daumé III Haland Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92. https://doi.org/10.1145/3458723.

National Centers for Environmental Information (NCEI). 2024. "Cherry Blossom Phenological Data." https://www.ncei.noaa.gov/access/paleo-search/study/26430.

Yasuyuki, Aono. 2015. *Cherry blossom phenology and temperature reconstructions at Kyoto. .* http://atmenv.envi.osakafu-u.ac.jp/aono/kyophenotemp4/.