Datasheet for 'Sakura Dataset'*

Shanjie Jiao

28 November 2024

This file recored all the information toward the dataset and usage.

Extract of the questions from (**gebru2021datasheets?**).

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 analyze and model the flowering dates of sakura trees in Japan, focusing on understanding how climatic and geographic factors influence phenology. It addresses the need for a structured, comprehensive dataset combining historical and modern records to study trends related to climate change and its impact on sakura flowering patterns.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - The dataset integrates historical data from Prof. Yasuyuki Aono's work and modern data from the Japan Meteorological Agency (JMA), collected by Alex
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.*
    - TBD
4. *Any other comments?*
    - The temperature are rescraped by Shanjie Jiao, since the origional data provided by Alex existing error in data

Composition

*Code and data are available at: https://github.com/Jie-jiao05/Sakura-Blossom-Prediction-Model.

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.*

    - The instances represent sakura flowering and full-bloom dates, along with associated climatic data (e.g., monthly mean temperatures) and geographic details (e.g., latitude, longitude).

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

    - The historical dataset includes records from 812–2015 (Kyoto), while the modern dataset spans 1953–2019 (Japan-wide), totaling thousands of observations across regions and 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 historical dataset focuses on Kyoto and is limited to available records. The modern dataset includes extensive observations across Japan, covering a wide geographic range.

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.*

    - Each instance includes flowering date (day of year), full bloom date, monthly mean temperatures, and geographic coordinates (latitude, longitude).

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

    - The target variable is the flowering date (numeric day of the 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.*

    - Historical data may have gaps for specific years, especially during periods of war or societal disruption. Some temperature reconstructions may lack direct observational validation.

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.*

- Geographic and temporal relationships are implicit through variables like latitude, longitude, year, and temperature.

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 was split into training (80%) and testing (20%) sets for model evaluation.

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

    - Potential biases exist in historical data due to reliance on subjective records and reconstructed temperatures.

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.*

    - The dataset is self-contained but cites external sources like the JMA and Prof. Aono's research.

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 based on public historical and meteorological records.

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.*

    - No, the dataset focuses on neutral scientific and cultural information.

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.*

    - The dataset indirectly identifies sub-populations by region and time period.

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, the dataset does not contain personal data.

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.*

    - No sensitive data is included.

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 data was reconstructed from historical records, while modern temperature data was directly observed and reported by the JMA. Collected by Alex

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?*

    - Historical data used manual curation of historical documents. Modern data was collected through meteorological stations and systematic phenological observations.

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

    - The modern dataset covers all JMA monitoring stations, ensuring comprehensive coverage of Japan. The historical dataset is specific to Kyoto and reflects available records.

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 collection was led by researchers like Prof. Yasuyuki Aono. Modern data collection was conducted by the JMA.Collected by Alex

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.*

    - Historical data spans 812–2015; modern data 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.*

    - Not applicable, as the dataset consists of non-sensitive historical and climatic data.

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 collected from documented sources and meteorological observations.

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.*

    - Not applicable.

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.*

    - Not applicable.

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).*

    - Not applicable.

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 has been widely analyzed in studies to assess its implications for climate research.

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.*

    - Yes, preprocessing included handling missing values for historical March temperature reconstructions by excluding incomplete entries. Modern data were cleaned to standardize date formats and merge climatic variables with phenological data based on station identifiers. Latitude and longitude coordinates were validated to ensure accuracy.And scraped temperature data from JMA

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.*

    - Yes, the raw data was preserved to ensure reproducibility and support potential future analyses. It is available at [https://github.com/Jie-jiao05/Sakura-Blossom-Prediction-Model]

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

    - Yes, preprocessing was conducted in R, and the cleaning scripts are available on GitHub at [https://github.com/Jie-jiao05/Sakura-Blossom-Prediction-Model].

4. *Any other comments?*

    - No

Uses

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

    - Yes, it has been used for modeling sakura flowering dates and analyzing the impacts of climate change on phenology.

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.*

    - References to related studies are provided in the bibliography. The dataset itself is hosted at [https://github.com/Jie-jiao05/Sakura-Blossom-Prediction-Model].

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

    - It could be used for ecological studies, climate change analysis, tourism planning, and educational purposes in phenology and climatology.

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?*

    - The reliance on historical records for Kyoto introduces potential biases (e.g., confirmation and selection biases) that may impact generalizability. Future users should be cautious when extrapolating beyond the dataset's geographic or temporal scope.

5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

    - The dataset should not be used for individual-level predictions, as it does not account for microclimate variations or tree-specific characteristics.

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.*

    - Yes, it will be made publicly available through [https://github.com/Jiejiao05/Sakura-Blossom-Prediction-Model].

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*

    - The dataset will be distributed via a publicly accessible repository. A DOI will be assigned for citation purposes.

3. *When will the dataset be distributed?*

    - Upon publication of the accompanying study or paper.

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.*

    - Yes, the dataset will be distributed under a Creative Commons Attribution-NonCommercial 4.0 International 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, the dataset compiles publicly available data from reputable sources without additional restrictions.
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 export controls or regulatory restrictions apply.
7. *Any other comments?*
    - TBD

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - The dataset will be maintained by the primary research group, with periodic updates as needed.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Contact information will be provided on the repository page
3. *Is there an erratum? If so, please provide a link or other access point.*
    - The origional temperature data provided by Alex have issue with accuracy, the correct data is rescraped from JWA by Shanjie Jiao.[https://github.com/Jie-jiao05/Sakura-Blossom-Prediction-Model]
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)?*
    - Yes, updates will include corrections, additional instances, or extended temporal coverage. Notifications will be sent via the repository's release page[https://github.com/Jie-jiao05/Sakura-Blossom-Prediction-Model]
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.*
    - Not applicable, as the dataset does not include personal information.

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.*

    - Yes, prior versions will remain accessible alongside newer versions.

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, and all submissions will be validated before inclusion.

8. *Any other comments?*

    - No

1 References