Datasheet for ‘Combined Weather and Climate Data’*

Lexun Yu

2024-11-26

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.*

    - Each instance represents a monthly atmospheric observation at Vancouver International Airport, with variables including temperature, wind speed, atmospheric pressure, precipitation, and gust speed.

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

    - Yu, Lexun.

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 uses public resources licensed under the Open Government Licence – Canada.

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.*

[*] Code and data are available at: https://github.com/yulexun/ClimateChangeYVR.

- Each instance represents a monthly atmospheric observation at Vancouver International Airport, with variables including temperature, wind speed, atmospheric pressure, precipitation, and gust speed.

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

    - There are 610 instances spanning from August 1959 to August 2010.

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 is a subset, focused on data from the Vancouver International Airport. It is representative of atmospheric conditions at this location but not globally.

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 contains both raw measurements (e.g., wind speed, precipitation) and derived features (e.g., logarithmic and Box-Cox transformations).

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

    - The primary target variable is the log-transformed mean temperature.

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.*

    - Yes, missing values in raw data were addressed using imputation and quality control by Environment and Climate Change Canada.

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.*

    - There is no explicit relationships, as the dataset treats instances as independent monthly observations.

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 (70%) and testing (30%) sets for model validation, it is stored as train_data and test_data.

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

    - Zero-inflation was noted in the snow variable, and skewness in several predictors required transformations.

10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)?

    - The dataset is self-contained but is derived from public resources maintained by Government of Canada.

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 contains publicly available climate data.

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, it contains neutral climate measurements.

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.*

    - No.

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 is anonymized and does not include personal information.

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, the dataset contains general climate observations.

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.*

   - The data was collected using sensors and manual measurements from weather stations as maintained by the Meteorological Service of Canada.

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?*

   - Weather instruments such as anemometers and rain gauges were used, and adjustments were made for consistency in the AHCCD dataset.

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 dataset includes monthly aggregated observations from the Vancouver International Airport station.

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)?*

   - Environment and Climate Change Canada.

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.*

   - From August 1959 to August 2010.

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.

7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*

   - It is obtained from sensors.

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.*

   - Sensors are not notified because they do not have conscious.

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.*

   - NA

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).*

    - NA

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.*

    - 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, significant preprocessing was performed:
     - variables (e.g., wind speed, precipitation, pressure, gust speed) were transformed using logarithmic and Box-Cox methods to stabilize variance and reduce non-linearity.
     - Missing values were imputed, and quality control flags from the original datasets were used to exclude invalid entries.
     - Column names were standardized for consistency, and dates were unified to a common format (yyyy-mm-dd).
     - The raw historical weather dataset and AHCCD dataset were merged using the date-time variable as the key.

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 raw data is publicly available through Canadian Centre for Climate Services and Meteorological Service of Canada resources. Adjustments and cleaned data are documented in the project repository.

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, the analysis and preprocessing were done in R. The R scripts and documentation are available in the GitHub repository: ClimateChangeYVR.

Uses

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

   - Yes, the dataset was used to develop and validate a polynomial regression model predicting the mean temperature at Vancouver International Airport based on atmospheric predictors.

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.*

   - No.

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

   - The dataset could be used for:
     - Climate change impact studies.
     - Developing predictive models for weather variables.
     - Evaluating relationships between atmospheric conditions and airport operations.
     - Educational purposes in data science and meteorology.

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 dataset's focus on a single location limits its generalizability to global or non-coastal contexts. Care should be taken when using it for studies outside Vancouver's climatic conditions. The use of adjusted data may also mask localized variations.

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

   - The dataset should not be used to infer causal relationships due to the observational nature of the data. It is also unsuitable for real-time operational weather forecasting as it represents historical data.

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, the dataset can be accessed through the linked GitHub repository under an open government license.

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

   - There is no DOI assigned. It is distributed on Github.

3. *When will the dataset be distributed?*

   - It is currently available in the Github repository.

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 is distributed under the Open Government Licence – Canada.

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 additional restrictions beyond those specified in the Open Government Licence.

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.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*

    - The GitHub repository owner, Yu, Lexun, currently maintains the dataset.

2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*

    - Via the GitHub repository issues page.

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.*

    - No specific retention limits; the dataset will remain accessible through GitHub.

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.*

    - Only the most recent version will be maintained. Older versions can be accessed through GitHub's version control.

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 can be submitted via pull requests on GitHub.

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.