Datasheet for 'Toronto Beaches Observations Dataset'*

A comprehensive dataset for analyzing environmental, temporal, and spatial factors affecting waterfowl counts on Toronto beaches

Yizhe Chen

December 3, 2024

The Toronto Beaches Observations Dataset is designed to analyze environmental, temporal, and spatial factors influencing waterfowl counts across Toronto beaches. It contains rich features including weather variables, water clarity, and seasonal indicators, enabling robust ecological and environmental modeling.

Extract of the questions from Gebru et al. (2021). Some information are from Parks, Forestry & Recreation (2024).

Motivation

1. For what purpose was the dataset created?
    - The dataset was created to collect related observations on Toronto beaches, including environmental data. The specific goal is to observe ecological situation of Toronto beaches.
2. Who created the dataset and on behalf of which entity?
    - The dataset was curated by the City of Toronto as part of its public environmental monitoring initiative.
3. Who funded the creation of the dataset?
    - The dataset is publicly available and funded by the City of Toronto through its environmental monitoring programs.

Composition

1. What do the instances in the dataset represent?

[*] Code and data are available at: https://github.com/YizheChenUT/Waterfowl-on-Toronto-Beaches.git.

    - Each instance represents daily observations of environmental and ecological conditions at a Toronto beach, including waterfowl counts and weather parameters.

2. How many instances are there in total?
    - The dataset comprises over 9,300 instances collected from multiple beaches.
3. Does the dataset contain all possible instances or is it a sample?
    - The dataset represents a complete collection of available observations for the given time frame and locations.
4. What data does each instance consist of?
    - Each instance includes variables such as waterfowl counts, air temperature, water temperature, wind speed, rain, wave action, water clarity, beach name, date, and month.
5. Is there a label or target associated with each instance?
    - Yes, the primary target variable is `water_fowl`, representing the count of waterfowl observed.
6. Is any information missing from individual instances?
    - Missing values in numeric variables (e.g., temperatures) were removed during preprocessing.
7. Are relationships between individual instances made explicit?
    - Relationships between observations are implicit based on shared temporal and spatial contexts.
8. Are there recommended data splits?
    - Users may split the data temporally (e.g., by year) for training, validation, and testing.
9. Are there any errors, sources of noise, or redundancies in the dataset?
    - Noise may be present due to variability in manual observation methods.
10. Is the dataset self-contained?
    - Yes, the dataset is self-contained and does not rely on external resources.
11. Does the dataset contain data that might be considered confidential?
    - No, all data is publicly available and anonymized.
12. Does the dataset contain potentially offensive or sensitive data?
    - No, the dataset does not contain sensitive content.

13. Does the dataset identify any sub-populations?
    - Sub-populations are identified by `beach_name`, representing geographic locations.
14. Is it possible to identify individuals from the dataset?
    - No, the dataset does not contain personally identifiable information.
15. Does the dataset contain sensitive data?
    - No, the dataset contains only environmental and observational data.

Collection process

1. How was the data associated with each instance acquired?
    - Data was collected through manual and automated observations by the City of Toronto's environmental monitoring team.
2. What mechanisms or procedures were used to collect the data?
    - Observations were recorded using field measurement tools and standardized reporting templates.
3. If the dataset is a sample, what was the sampling strategy?
    - Not applicable; the dataset represents all available observations.
4. Who was involved in the data collection process?
    - Data collection was performed by trained personnel from the City of Toronto.
5. Over what timeframe was the data collected?
    - The dataset spans multiple years, providing seasonal and annual coverage.
6. Were any ethical review processes conducted?
    - No ethical review was required as the dataset involves public environmental observations.
7. Did you collect the data from individuals directly?
    - No, the dataset was collected from public environmental observations.
8. Were the individuals in question notified about the data collection?
    - Not applicable.
9. Did the individuals consent to the collection and use of their data?
    - Not applicable.
10. If consent was obtained, was there a mechanism to revoke consent?

- Not applicable.

11. Has an analysis of the potential impact of the dataset been conducted?
    - No formal impact analysis has been conducted.

Preprocessing/cleaning/labeling

1. Was any preprocessing/cleaning/labeling of the data done?
    - Yes, preprocessing included removing missing values, standardizing variable names, and filtering outliers.
2. Was the raw data saved in addition to the preprocessed/cleaned data?
    - Yes, raw data is preserved in its original form.
3. Is the software used to preprocess/clean the data available?
    - Yes, preprocessing scripts are included in the project repository.

Uses

1. Has the dataset been used for any tasks already?
    - Yes, it has been used for ecological modeling and research on waterfowl populations.
2. Is there a repository that links to any or all papers or systems that use the dataset?
    - Currently, this paper represents one of the primary uses of the dataset.
3. What (other) tasks could the dataset be used for?
    - The dataset can be used for climate change studies, habitat management, and wildlife behavior analysis.
4. Are there tasks for which the dataset should not be used?
    - The dataset is unsuitable for uses outside its ecological and environmental scope.

Distribution

1. Will the dataset be distributed to third parties?
    - The dataset is publicly available through the City of Toronto's Open Data portal.
2. How will the dataset be distributed?
    - Via CSV files downloadable from the portal.
3. When will the dataset be distributed?
    - It is already available.

4. Will the dataset be distributed under a copyright or IP license?
    - Yes, the dataset is distributed under the City of Toronto's Open Data License.
5. Have any third parties imposed restrictions on the data?
    - No restrictions are known.
6. Do any export controls or regulatory restrictions apply?
    - None.

Maintenance

1. Who will be supporting/hosting/maintaining the dataset?
    - The City of Toronto.
2. How can the owner/curator/manager of the dataset be contacted?
    - Through the City of Toronto's Open Data portal.
3. Is there an erratum?
    - None known.
4. Will the dataset be updated?
    - Updates occur periodically to include new observations.
5. If the dataset relates to people, are there retention limits?
    - Not applicable.
6. Will older versions of the dataset continue to be supported?
    - Yes, historical versions are archived.
7. If others want to extend/augment/build on the dataset, is there a mechanism?
    - Users may fork the repository or use the dataset as permitted under its license.

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.

Parks, Forestry & Recreation. 2024. "Open Data Dataset." *About Toronto Beaches Observations*. https://open.toronto.ca/dataset/toronto-beaches-observations/.