Datasheet for 'Raw_Data_WaterQuality.csv'*

Wentao Sun

April 18, 2024

This data set contains water quality monitoring records at multiple measurement points since 1994. The main measurement parameters include salinity, dissolved oxygen, pH value, transparency, water temperature and air temperature, aiming to monitor and evaluate changes in environmental water quality. The data are obtained through regular sampling and cover a wide range of time spans and locations, providing important long-term data resources for studying water environment health and ecological changes. These data can be used for environmental science research, policy development, and to promote sustainable water resources management.

Extract of the questions from Interior (2019).

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.*
    - This dataset was created to analyze water quality in the United States. We were unable to find a publicly available dataset in a structured format that contained the water quality criteria values needed for modeling.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
    - U.S Department of the Interior
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.*
    - None
4. *Any other comments?*

*Code and data are available at: https://github.com/TonySun1107/Water_Quality.git

- None

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.*

    - This dataset represents a combination of environmental measurement instances, taken at different times and potentially different locations, focusing on water quality metrics. The presence of multiple types of measurements (like water temperature, air temperature, pH, etc.) suggests a detailed tracking of environmental conditions over time.

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

    - One

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 does not contain all possible instances, it is a sample of instances from a larger set. A larger set is the various indicators of water quality. The sample is not representative of the larger collection.

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.*

    - The original data includes the year, month and day of the time, water salinity, dissolved oxygen, pH, water depth, water temperature and air temperature.

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

    - None

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.*

    - Data that did not meet the requirements or that seriously did not meet the requirements were processed before the start of the study.

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.*

   - Relationships between individual instances have been clarified, with a direct negative correlation between dissolved oxygen content and temperature in the same waters. An increase in temperature results in a decrease in dissolved oxygen content.

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 data used in this study is already clear and neat and data splitting is not recommended.

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

   - The first step at the beginning of the study for the presence of errors in the dataset is to clean and analyze the obtained dataset.

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 sets are self-contained, ensuring that these resources exist and remain unchanged over time

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.*

    - The dataset does not contain data that may be considered confidential.

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 data set does not contain the above factors.

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 is age-identified.

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

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 data set does not contain the above factors.

16. *Any other comments?*

    - None

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 associated with each instance was obtained through measurements by government professionals. These data are not directly observable. Nor are they reported by subjects or indirectly inferred/derived from other data.

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?*

    - The hardware apparatuses or sensors

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

    - random sampling

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)?*

    - Staff arranged by government departments have been officially paid by the government.

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.*

   - 1989-2019

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.*

   - None

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

   - Derived from United States Government data

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.*

   - None

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.*

   - None

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).*

    - Pass

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.*

    - TBD

12. *Any other comments?*

    - None

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.*
    - There has been no preprocessing/cleaning/labeling of the data.
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.*
    - Pass
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - Pass
4. *Any other comments?*
    - Pass

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - This dataset has not been used for any task before.
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 link, this is data processed out of a government department.
3. *What (other) tasks could the dataset be used for?*
    - Analysis of environmental protection and future sustainable development.
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 composition of the dataset or the way it was collected and pre-processed/cleaned/labeled will not affect future use.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*

- None

6. *Any other comments?*
    - None

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 will not be distributed to the entity that created it. The dataset exists only on the government website.
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 is distributed and promoted via a zip archive on the website.
3. *When will the dataset be distributed?*
    - October 29, 2023
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.*
    - https://ecos.fws.gov/ServCat/Reference/Profile/117348
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 third party imposes IP-based or other restrictions on the data associated with the instance.
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.*
    - The dataset will not be subject to any export controls or other regulatory restrictions.
7. *Any other comments?*
    - None

Maintenance

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

    - U.S. Department of the Interior

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

    - Github: https://github.com/GSA/data.gov.git

3. *Is there an erratum? If so, please provide a link or other access point.*

    - None

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)?*

    - The dataset will be updated with new instances, the frequency of which will be determined by the Department of the Interior.

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.*

    - This dataset is not related to people.

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.*

    - Older versions of the dataset will not be maintained, but new data will be generated. There will be a specialized government department responsible for data measurement and statistics.

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.*

    - Others are not allowed to extend/add/build/contribute to the dataset as this data is collected and processed by official government departments and cannot be manipulated by ordinary people.

8. *Any other comments?*

    - None

References

Interior, U. S.Department of the. 2019. “BKB_WaterQualityData_2020084.csv.” *Water Quality Data*.https://ecos.fws.gov/ServCat/Reference/Profile/117348.