Datasheet for 'NuSEDS-New Salmon Escapement Database System'

Timothius Prajogi

April 19, 2024

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 keep track and monitor salmon spawning records and spawner abundance estimates. It is the primary database used to track salmon for Fisheries and Oceans Canada and it enables future research. It reports spawning observations for 9100_ individual populations.

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

   - The dataset was published by Fisheries and Oceans Canada.

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.*

   - This is a government led program.

4. *Any other comments?*

   - No.

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.*

- An instance in a dataset represents the spawning salmon population observed in a specific waterbody during a specific year.

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

   - There are 413,434 observations made.

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 contains all recorded observations made in the Pacific waterbodies of Canada, and is intended to be a comprehensive dataset.

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 observation notes down amount of salmon observed, when it was observed, the type of salmon, any broodstock removed, and the waterbody the observation was made in.

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

   - Each observation is labeled with a unique identifier labeled `ACT_ID`.

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.*

   - The instance may or may not record certain data regarding the spawning population. For broodstock data, when it is not present it signifies no broodstock was removed from the waterway. Specific date values are sometimes missing, being insignificant data as spawning season generally comes around at the same time, with specific dates recorded only for certain waterbodies.

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 waterbodies and regions are combined to create a unique id which can be used to reference all other observations made about that waterbody and region for any year, species, or otherwise.

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.*

- No datasplits were recommended.

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

   - Historical data that was imported was included but is largely blank. This data just recorded basic observations, with newer observations provided much more detail.

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, hosted on Canada's open data portal.

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.

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.

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.*

    - As mentioned the observations were split by species, waterbody, and region. These attributes are identified in their respective columns.

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.*

- No.

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.*

   - The data was collected as a survey, made from observations gathered from field researchers. Estimates were made as to the exact counts, with the scale of the matter being so large and spanning long periods.

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?*

   - Human curation was necessary to gather the data, along with measurement instruments.

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

   - N/A

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)?*

   - This is not explicitly mentioned, however the data was provided by Fisheries and Oceans 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.*

   - The data is collected yearly, and has been collected since 1950 with the last update being 2022. The dataset is updated quarterly as finalized observations come in.

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 ethical review processes were stated.

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

   - I collected the data directly from their API source.

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.*

   - No they were not, they had licensed the dataset to not require such authorization.

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.*

    - No analysis was stated.

12. *Any other comments?*

    - The background review process for this dataset was not stated, being a government database.

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, N/A values needed to be handled, either treated as 0, or their observations ignored. Depending on the requirements, salmon species were grouped together by year to get an understanding of the grander scale of the population.

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 in this project can be found in `input/data` and can be found from the source here: https://open.canada.ca/data/en/dataset/c48669a3-045b-400d-b730-48aafe8c5ee6.

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, R a statistical programming language was used and the documentation can be found here: https://www.r-project.org/

4. *Any other comments?*

   - No.

Uses

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

   - I imagine so, but usage has not been documented.

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 covers a lot of features that can be used in various aspects of salmon health monitoring. Other research surrounding each of the different features, spawn timings, broodstock, population, can be done.

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?*

   - No.

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

   - It is to be noted that this dataset is only for the Canadian Pacific, and should only be used in analyses that are about that region.

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 is distributed as part of Canada's open data portal.

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

   - It will be downloadable through a provided API endpoint.

3. *When will the dataset be distributed?*

   - It is updated quarterly and was published in 2016

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.*

   - It is published 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.*

   - Not stated.

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.*

   - Not stated.

7. *Any other comments?*

   - No.

Maintenance

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

   - Maintenance is done by the publisher Fisheries and Oceans Canada.

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

   - They have provided an email address to be contacted: Zhipeng.Wu@dfo-mpo.gc.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)?*

   - Not stated.

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.*

   - Not stated.

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.*

   - Not stated.

8. *Any other comments?*

   - No.

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.