Datasheet For Air Quality Index By Municipality, Leading Causes Of Death And Average PM2.5 Concentrated Dataset[*]

Vanshika Vanshika Shivank Goel Navya Hooda

March 16, 2024

This datasheet documents the datasets used in a study analyzing the impact of air pollution on health outcomes in Alberta, Canada. The data for the study is sourced from the Alberta government's open data portal. It spans several years and includes variables like AQHI readings, PM2.5 concentrations, and mortality rates linked to specific health conditions. This dataset is helpful for its concise representation to study the environmental health in Alberta. It offers insights into public health implications of air pollution. The data cleaning and processing done using tools like R, ensures its reliability and accuracy. This datasheet is an essential guide to the dataset's structure, collection process, and potential applications.

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 investigate the correlation between air quality, particularly PM2.5 levels, and mortality rates due to respiratory and cardiac illnesses in Alberta. The aim was to understand how air pollution impacts public health in this region.

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

The datasets were created by the Alberta Government, specifically by Jobs, Economy and Northern Development, Service Alberta, and Environment and Parks departments.

[*] Code and data are available at: https://github.com/shivankgoel003/Mortality-in-Alberta.

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 creation of these datasets was likely funded by the Alberta government. No specific grant details were provided.

4. *Any other comments?*

These datasets provide critical insights into environmental and public health concerns in Alberta and are vital for ongoing research and policy development in these areas.

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.*

The instances represent various types of environmental and public health data. The first dataset details the Air Quality Health Indices by municipality in Alberta. The second dataset provides information on the leading causes of death in Alberta. The third dataset offers data on PM2.5 concentrations in Alberta.

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

The exact number of instances is not specified, but each dataset covers multiple years of data across various locations in Alberta.

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 datasets seem to be comprehensive for their respective scopes, covering broadly geographical areas within Alberta. It seems to be representative of the larger set of environmental and health data within the province.

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 dataset has specific labels or targets. For the air quality index, the labels include AQHI values and municipal regions. The causes of death dataset labels include types of illnesses and mortality counts. The PM2.5 dataset's labels consist of concentration levels, locations, and time periods

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

Each dataset has specific labels or targets. For the air quality index, the labels include AQHI values and municipal regions. The causes of death dataset labels include types of illnesses and mortality counts. The PM2.5 dataset's labels consist of concentration levels, locations, and time periods.

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.*

There is no specific mention of missing information. However, as with most large datasets, there may be gaps due to data unavailability or limitations in data collection.

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 datasets do not explicitly mention relationships between individual instances, but they can be inferred. For example, higher AQHI and PM2.5 values may correlate with increased mortality rates from certain diseases.

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.*

There are no specific recommended data splits mentioned. Data splits would depend on the specific research or analysis goals.

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

Specific errors or sources of noise are not mentioned. However, as with any environmental data, there may be variabilities and inconsistencies due to natural fluctuations and measurement limitations.

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 datasets are self-contained and available through the Alberta open data portal. They are subject to the Open Government Licence - Alberta, which allows for reuse and distribution.

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 confidential or personally identifiable information is included in the dataset.

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.*

There is no content in these datasets that would be considered offensive or insulting. The data is factual, focusing on environmental and health statistics.

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 datasets do not specify sub-populations like age or gender. They are more focused on geographical and temporal aspects of air quality and health outcomes.

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.*

Individual identification is not possible with these datasets as they deal with aggregated data at the municipal level or higher.

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 datasets do not contain sensitive personal data. They are focused on environmental measures and aggregated health outcomes.

16. *Any other comments?* N/A

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 were directly observed and collected through environmental monitoring stations and health records. They are likely verified through standard government data validation processes, though specific details are not provided.

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 air quality and PM2.5 data were likely collected using environmental monitoring stations equipped with sensors. The causes of death data were probably compiled from health records.

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 datasets seem to be comprehensive collections rather than samples. They likely represent the entire population of interest (i.e., all municipalities in Alberta for air quality and all recorded causes of death).

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)?*

Data collection was likely conducted by government employees or contractors. Specific details about their compensation are not provided.

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 collection covers several years, as indicated by the datasets (e.g., 2001-2021 for the leading causes of death).

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.*

There is no specific mention of ethical review processes.

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

The data were collected from environmental monitoring stations and health records, not directly from individuals. It is a secondary data 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.*

Individual notification is not applicable, as the data do not pertain to individual persons.

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.*

Individual consent is not relevant for this type of data, as it does not involve personal data collection from individuals.

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).*

As individual consent was not required for this data collection.

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.*

There is no information provided regarding an impact analysis on data subjects, likely because the datasets deal with environmental and health data and not individual-specific data.

12. *Any other comments?*

The data collection process is extensive and critical for understanding the environmental and public health in Alberta. The absence of personal data reduces privacy concerns.

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 is likely some level of preprocessing and cleaning involved, ensuring consistency, and possibly dealing with missing values. However, specific details are not provided in the information given.

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.*

There is no mention of whether the raw data is saved alongside the processed data.

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

Details about the specific software or tools used for preprocessing, cleaning, or labeling are not provided.

4. *Any other comments?*

Proper data preprocessing and cleaning are essential for the accuracy and reliability of analysis results. The datasets, as released, should be ready for use in various analyses and research.

Uses

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

The datasets have likely been used for academic research and study before. They have been used to analyze the relationship between mortality rates and causes of certain diseases.

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 specific repository is mentioned, but the datasets are available through the Alberta open data portal.

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

The data could be used for further environmental health studies, public health planning, and educational purposes.

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 users should consider potential biases and reporting in data collection.Also they should be aware of geographic coverage to avoid misinterpretations.

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

The data should not be used for identifying or targeting specific individuals.

6. *Any other comments?*

The datasets are valuable for understanding trends and correlations in environmental and public health. They should be used appropriately, considering their scope and limitations.

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 datasets are publicly available and can be accessed by third parties. They are distributed under the Open Government Licence - Alberta, which allows for use and distribution by anyone.

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

The datasets are available for download through the Alberta open data portal. There is no mention of a DOI or distribution through other platforms like GitHub.

3. *When will the dataset be distributed?*

The datasets are already available for public access.

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.*

The datasets are distributed under the Open Government Licence - Alberta, which permits a wide range of uses. The license details can be found on the Alberta open data portal.

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.*

There is no mention of any third-party IP restrictions on these datasets.

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.*

There are no specified export controls or regulatory restrictions beyond the Open Government Licence - Alberta.

7. *Any other comments?*

The open access nature of these datasets makes them a valuable resource for a wide range of users, from researchers to policymakers.

Maintenance

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

The datasets are maintained by the respective departments of the Alberta government – Jobs, Economy and Northern Development, Service Alberta, and Environment and Parks.

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

Contact information is likely available through the Alberta government's open data portal.

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

There is no mention of an erratum.

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 datasets are updated annually. Specific communication channels for updates are not mentioned.

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.*

There is no specific information about the maintenance of older versions of the dataset.

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.*

This is not applicable as the datasets do not contain individual-level data.

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.*

There is no mentioned mechanism for external contributions to the dataset. The data are sourced from government monitoring and records, so external contributions might not be applicable.

8. *Any other comments?*

Regular maintenance and updates are crucial for the continued relevance and accuracy of these datasets.

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.