Datasheet: *Air Quality e-Reporting Annual Statistics*

Original authors/creators: ***European Environment Agency***
Organization: ***European Environment Agency***
Source: ***https://discomap.eea.europa.eu/App/AirQualityStatistics/index.html***

I. Motivation For Datasheet Creation

A. Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

To analyze data published by the European Environment Agency with respect to how different pollutants correlate with their surrounding environments

B. Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

Yes, the EEA (European Environment Agency) has numerous data visualizations including fact sheets of every individual member nation (`https://www.eea.europa.eu/data-and-maps/daviz#b_start=0&c4=air`)

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

The dataset could be used to ensure nations keep their obligations in reducing air pollution, thus holding them accountable.

D. Who funded the creation dataset?

Was originally proposed by the EU Commission and is actively maintained by the European Environment Agency

E. Any other comment?

None

II. Datasheet Composition

A. What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

There is only one type of instance. Each instance is data that was measured by an air quality measurement station at a certain point in time.

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

Total of 4 511 571 instances

C. What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

Dataset contains features describing the nature of the measurement station (location, label, responsible nation, etc...), the means of aggregation over an entire year and the aggregated amount of a given pollutant

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

No, there is only a label associated with the measuring station

E. Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

As far as we can tell, all information has been made public

F. Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

No, no such links between measurements are made explicit

G. 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 (e.g., 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 (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

No, this dataset is the full version and thus contains all possible instances

H. Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

There are none, because we will not perform any training on the dataset

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

There could be inherent calibration biases and noise between air quality measurement stations depending on how accurate the pollutant sensors are

J. Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., 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 (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? 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 calculated by aggregating raw data produced by the air quality measurement stations over an entire year. Only some measurements have links to this aggregated data and these are provided as a field in the dataset. There is no written guarantee that these links will remain reliable, however, given that the dataset is generated by the EEA, we can assume some level of reliability in the future

K. Any other comments?

None

III. Collection Process

A. What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The data was collected via the use of air quality measurement stations. Each EU member nation is responsible for the maintenance of their own stations

B. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

Measured data was aggregated over the course of a year

C. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The data is the full version

D. Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Each nation is independently responsible for reporting the data collected by their own measurement stations. It is assumed that the employees responsible for maintaining the stations are compensated accordingly

E. Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The first reported measurements date back to 1969, however there are very few such observations. The dataset is updated annually to reflect new measurements reported

IV. Data Preprocessing

A. Was any preprocessing/cleaning/labeling of the data done (e.g., 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 remainder of the questions in this section.

None

B. Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

C. Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

E. Any other comments

V. Dataset Distribution

A. How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The data is distributed through an online service called "Discomap". This service allows for simple sorting and filtering options. Additionally, a .csv copy of the dataset can be downloaded from the webservice

B. When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset is already publicly distributed on EEA's website

C. Are there any copyrights on the data?

The dataset does not contain any intellectual property and therefore does not have any copyrights

D. Are there any fees or access/export restrictions?

None

E. Any other comments?

None

VI. Dataset Maintenance

A. Who is supporting/hosting/maintaining the dataset?

The European Environment Agency hosts the data on their website

B. Will the dataset be updated? If so, how often and by whom?

The dataset is updated annually in order to stay up to date

C. How will updates be communicated? (e.g., mailing list, GitHub)

The EEA has not made available any means of receiving updates on the dataset

D. If the dataset becomes obsolete how will this be communicated?

The dataset contains objective measurements of air quality. The data should not become obsolete

E. Is there a repository to link to any/all papers/systems that use this dataset?

All the products derived from the data (at least by the EEA) can be found on the webpage that hosts the dataset in the related links section (`https://www.eea.europa.eu/data-and-maps/data/aqereporting-9`)

F. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

There is none. The dataset is updated by the EEA after EU member nations have reported their respective measurements

VII. Legal and Ethical Considerations

A. Were any ethical review processes conducted (e.g., 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.

Because the dataset only contains objective measurements, there is no need for an ethical review process

B. Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? If so, please provide a description.

No

C. 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

D. Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No

E. Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

F. Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

G. Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial 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.

H. Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

I. 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.

J. 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.

K. 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).

L. Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., 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.

M. Any other comments?

None