Datasheet for SAR annotations of AIS data

See:

* https://data.dtu.dk/collections/AIS\_Trajectories\_from\_Danish\_Waters\_for\_Abnormal\_Behavior\_Detection/6287841
* <https://github.com/KristofferOlesen/Datasets-of-AIS-Trajectories-from-Danish-Waters/>

**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).
   * Using Deep Learning for detection of maritime abnormal behaviour in spatio temporal trajectories is a relatively new and promising application. Open access to the Automatic Identification System (AIS) has made large amounts of maritime trajectories publically avaliable. However, these trajectories are unannotated when it comes to the detection of abnormal behaviour.

The lack of annotated datasets for abnormality detection on maritime trajectories makes it difficult to evaluate and compare suggested models quantitavely. With this dataset, we attempt to provide a way for researchers to evaluate and compare performance.

We have manually labelled trajectories which showcase abnormal behaviour following an collision accident. The annotated dataset consists of 521 data points with 25 abnormal trajectories. The abnormal trajectories cover amoung other; Colliding vessels, vessels engaged in Search-and-Rescue activities, law enforcement, and commercial maritime traffic forced to deviate from the normal course

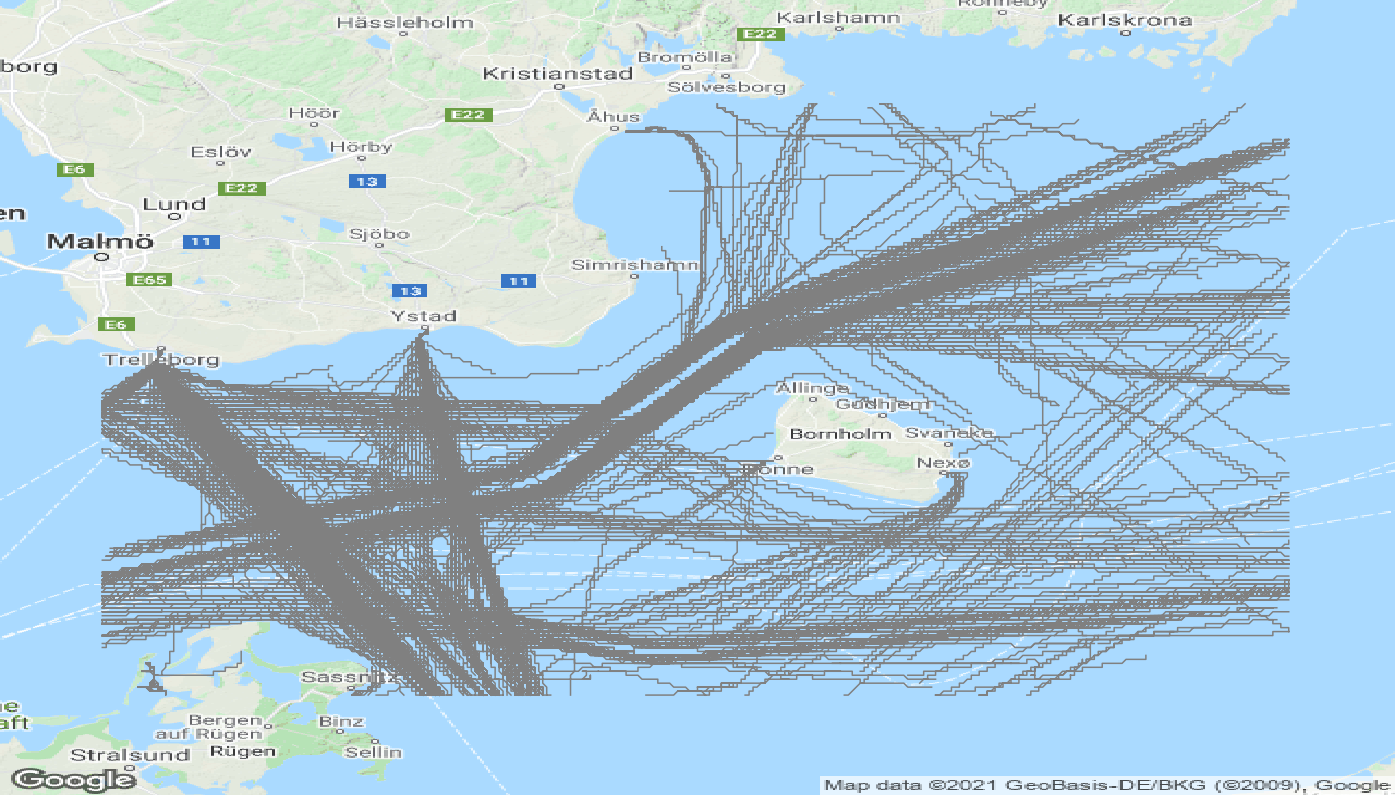
1. Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?
   * The dataset has been annotated by a PhD-student at the Technical University of Denmark for their research on detection of maritime abnormal behaviour using Deep Learning. The data was collected, stored, and provided by Terma A/S.
2. 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 Danish Ministry of Defence Acquisition and Logistics Organisation, grant no. 4600005159
3. Any other comments?
   * The research project is a collaboration between The Technical University of Denmark and Terma A/S,

**Composition**

1. What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions be- tween them; nodes and edges)? Please provide a description.
   * The data are sequences of maritime trajectories defined by their; timestamp, latitude/longitude position, speed, course, and unique ship identifer MMSI. In addition, the dataset contains metadata related to creation parameters
2. How many instances are there in total (of each type, if appropriate)?
   * The dataset consists of a total of 521 trajectories of which 25 is labelled as abnormal.
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 (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).

* The dataset has been limited to a specific time period, ship types, moving AIS navigational statuses, and filtered within an region of interest (ROI). Trajectories were split if exceeding an upper limit and short trajectories were discarded (see preprocessing Q1). All values are given as metadata in the dataset and used in the naming syntax.
  + Naming syntax:
    - data\_AIS\_Custom\_STARTDATE\_ENDDATE\_SHIPTYPES\_MINLENGTH\_MAXLENGTH\_RESAMPLEPERIOD.pkl
  + For more information on AIS shiptypes codes and navigational statuses see:
    - https://help.marinetraffic.com/hc/en-us/articles/205579997-What-is-the-significance-of-the-AIS-Shiptype-number-
    - https://help.marinetraffic.com/hc/en-us/articles/203990998-What-is-the-significance-of-the-AIS-Navigational-Status-Values-
* The ship types used in this data are: Cargo, Tanker, Fishing, Military, Passenger, High Speed Vessel, Pleasure, and Sailing vessel. Then vessels of types Search-And-Rescue, Law Enforcement and Other (Barge) has been added based on their relation to the abnormal collision accident.
* The data is saved using the pickle format for Python
* Each dataset is split into 2 files with naming convention:
  + datasetInfo\_XXX
  + data\_XXX
    - Files named "data\_XXX" contains the extracted trajectories serialized sequentially one at a time and must be read as such. Please refer to attached utility functions for examples.
    - Files named "datasetInfo" contains Metadata related to the dataset and indecies at which trajectories begin in "data\_XXX" files.
* The dataset is an example of a SAR event and can not be considered representative of a large population of all SAR events. The maritime traffic consists of 521 vessels on a single day in a specific region. The remaining normal traffic is representative of the traffic during the winter season. The normal traffic in the ROI has a fairly high seasonality related to fishing and leisure sailing traffic. Below is a table of mean values and standard deviation for all features during 3 different periods and an image of the ROI with a number of randomly sampled trajectories plotted.

|  |  |  |  |
| --- | --- | --- | --- |
|  | June 2021-November 2021 | December 2021 | December 13th 2021 |
| Latitude | 55.20 ± 0.32 | 55.20 ± 0.30 | 55.20 ± 0.29 |
| Longitude | 14.46 ± 0.76 | 14.53 ± 0.73 | 14.49 ± 0.69 |
| Speed | 3.90 ± 3.25 | 3.84 ± 3.22 | 3.72 ± 3.31 |
| Course | 172.43 ± 107.37 | 159.85 ± 102.35 | 160.41 ± 103.26 |



1. What data does each instance consist of? “Raw” data (e.g., unprocessed text or images) or features? In either case, please provide a de- scription.
   * Each instance provide timestamp, latitude/longitude position, speed, course, and unique ship identifer MMSI of a resampled maritime trajectory. The resampling frequency is provided as metadata
2. Is there a label or target associated with each instance? If so, please provide a description.

* Each instance is labelled as normal or abnormal based on manual inspection. The abnormality decision was made by judging the impact of a collision accident, which is regarded as abnormal.

1. 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.
2. Are relationships between individual instances made explicit (e.g., users’ movie ratings, social network links)? If so, please de- scribe how these relationships are made explicit.
   * Relationships between instances may occur but is not made explicit. Trajectories may be continuations of other trajectories in cases of very long trajectories.
3. 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.
   * The provided metadata contains indecies for a 80/20 random split of the data. However, we recommend the data be used in entirety as a test set for models trained on other data
4. Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.
   * Trajectories unrelated to the collision accident may incorrectly be labelled as normal trajectories based on other events than the SAR of interest. In addition, earlier/later trajectories of related vessels may be incorrectly labelled as abnormal.
5. 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 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 can be used in isolation. However, we recommend it is used in collaboration with other larger datasets used for learning/training. We also provide unlabelled larger datasets with same preprocessing for training purposes.
6. Does the dataset contain data that might be considered confiden- tial (e.g., data that is protected by legal privilege or by doctor– patient confidentiality, data that includes the content of individ- uals’ non-public communications)? If so, please provide a description.
   * Not to our knowledge
7. Does the dataset contain data that, if viewed directly, might be of- fensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.
   * Not to our knowledge, but the collision accident in question did have fatal outcome for people involved. This information is not directly visible in the dataset.

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

Does the dataset identify any subpopulations (e.g., by age, gen- der)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. Is it possible to identify individuals (i.e., one or more natural per- sons), either directly or indirectly (i.e., in combination with other data) from the dataset? If so, please describe how.

Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals race or ethnic origins, sexual ori- entations, religious beliefs, political opinions or union member- ships, 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.

Any other comments?

**Collection Process**

1. 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 the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.
   * Automatic Identification System (AIS) is a mandatory self-reporting system for maritime vessels of a certain size. This data has been collected live from AIS listining station in Denmark and indexed by the unique vessel identifyer MMSI.

1. What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?
2. If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?
   * Dataset was filtered as described above

1. 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)?
   * PhD-student and supervisors.
2. 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 time- frame in which the data associated with the instances was created.
   * The data was collected in real time.
3. Were any ethical review processes conducted (e.g., by an institu- tional 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.

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

* + Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?
  + Were the individuals in question notified about the data collec- tion? If so, please describe (or show with screenshots or other informa- tion) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.
  + 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.
  + If consent was obtained, were the consenting individuals pro- vided 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).

• Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been con- ducted? 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.

• Any other comments?

**Preprocessing/cleaning/labeling**

1. 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 miss- ing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.
   * The data was filtered as described above. Trajectories were split if the time between two consecutive updates surpassed 10 min. Trajectories shorter than 10 min were discarded. Trajectories were then resampled and linearly interpolated to every 2 min. Trajectories longer than 12 hours were split into equal segments shorter than 12 hours.

1. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please pro-  
   vide a link or other access point to the “raw” data.
   * Raw data is saved but can not be provided. Raw data files should be available through the AIS system. The Danish Maritime Authority publicizes historical AIS data, however, we can not account for any differences in the raw data due to cleaning or filtering by the Danish Maritime Authority before publication.
     + https://dma.dk/safety-at-sea/navigational-information/ais-data
2. Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.
   * The software used to store and preprocess the raw data is confidential and thus can not be provided.
3. Any other comments?

**Uses**

The questions in this section are intended to encourage dataset creators to reflect on the tasks for which the dataset should and should not be used. By explicitly highlighting these tasks, dataset creators can help dataset consumers to make informed decisions, thereby avoiding potential risks or harms.

1. Has the dataset been used for any tasks already? If so, please provide a description.
   * The data set has been used to evaluate models for detection of abnormal maritime trajectories.
   * Unlabelled versions of the dataset has been used for clustering of maritime trajectories.
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. ‘
3. What (other) tasks could the dataset be used for?
   * Clustering of maritime trajectories for the purpose of maritime pattern recognition.
   * Classification of maritime vessels based on the trajectories.
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 (e.g., stereotyping, quality of service issues) or other risks or harms (e.g., legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?
5. Are there tasks for which the dataset should not be used? If so, please provide a description.
6. Any other comments?

**Distribution**

Dataset creators should provide answers to these questions prior to distributing the dataset either internally within the entity on behalf of which the dataset was created or externally to third parties.

1. Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
2. How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
   * Data will be hosted on servers owned by the Technical University of Denmark and is avaliable at:
   * <https://figshare.com/s/0012239be1c55e988a32>
3. When will the dataset be distributed?
   * Data is avaliable
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.
   * Data is distributed under a CC By 4.0 License.
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.
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.
7. Any other comments?

**Maintenance**

As with the questions in the previous section, dataset creators should provide answers to these questions prior to distributing the dataset. The questions in this section are intended to encourage dataset creators to plan for dataset maintenance and communicate this plan to dataset consumers.

* Who will be supporting/hosting/maintaining the dataset?
  + The data is hosted on Data will be hosted on servers owned by the Technical University of Denmark and is maintained by the authors.
* How can the owner/curator/manager of the dataset be contacted (e.g., email address)?
  + The owner of the dataset can be contacted using either:
    - kvol@dtu.dk
    - lkhc@dtu.dk
    - [kristoffer.olesen@gmail.com](mailto:kristoffer.olesen@gmail.com)
* Is there an erratum? If so, please provide a link or other access point.
* Will the dataset be updated (e.g., 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 (e.g., mailing list, GitHub)?
  + Any updates may be made if possible upon request to the emails above. Updates will not be announced.
* If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., 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.
* 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.
* 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 commu- nicating/distributing these contributions to dataset consumers? If so, please provide a description.
* Any other comments?