MARITIME BIG DATA ANALYSIS WITH ARLAS

W. Gautier*, S. Falquier, S. Gaudan

Data Science Department, Gisaïa
2 avenue de l'Escadrille Normandie Niemen
31700 Blagnac – FRANCE
(willi.gautier@gisaia.com, sylvain.gaudan@gisaia.com, sebastien.falquier@gisaia.com)

KEY WORDS: Maritime Data, Automatic Identification System (AIS), Vessels Traffic, Origin Destination, Big Data, Open Source, Data Visualization, Machine Learning

ABSTRACT:

The maritime industry has become a major part of globalization. Political and economic actors are meeting challenges regarding shipping and people transport. The Automatic Identification System (AIS) records and broadcasts the location of numerous vessels and delivers a huge amount of information that can be used to analyze fluxes and behaviors. However, the exploitation of these numerous messages requires tools based on Big Data principles.

Acknowledgement of origin, destination, travel duration and distance of each vessel can help transporters to manage their fleet and ports to analyze fluxes and focus their investigations on some containers based on their previous locations. Thanks to the historical AIS messages provided by the Danish Maritime Authority and ARLAS PROC/ML, an open source and scalable processing platform based on Apache SPARK, we are able to apply our pipeline of processes and extract this information from millions of AIS messages. We use a Hidden Markov Model (HMM) to identify when a vessel is still or moving and we create "courses", embodying the travel of the vessel. Then we derive the travel indicators.

The visualization of results is made possible by ARLAS Exploration, an open source and scalable tool to explore geolocated data. This carto-centered application allows users to navigate into the huge amount of enriched data and helps to take benefits of these new origin and destination indicators. This tool can also be used to help in the creation of Machine Learning algorithms in order to deal with many maritime transportation challenges.

1. INTRODUCTION

Maritime activity has significantly increased over the last two centuries. The management of the goods and people transportation has become an important challenge for political and economic actors. 90% of transportation of goods is globally carried out by more than 80,000 service vessels. To manage these fluxes and improve the vessel's safety, the Automatic Identification System (AIS) records and broadcasts their location in almost real time. It delivers more than 520 million messages per day from more than 180,000. It provides a reliable source of information for understanding maritime activity, but it requires powerful tools for handling such voluminous data.

Identifying the vessel's travels and determining their origin, destination and duration can help transporters to optimize their fleet activity and ports to understand the incoming/outcoming traffic and to select containers to inspect in priority for instance. We managed to extract these travel indicators from AIS messages. We use a dataset composed of AIS messages provided by the Danish Maritime Authority. This extraction follows several steps, adapted to the volume of data thanks to ARLAS PROC/ML, an open source and scalable processing platform. The results are visible with ARLAS Exploration, an open-source solution (ARLAS Development Team, 2021a) we have developed to visualize and explore interactively such amounts of geographical data. We will see how this valuable information can be extracted from such amounts of AIS messages thanks to these tools.

2. A PROCESS APPLIED TO ENRICH MARITIME DATA AND ISOLATE TRAVELS

We use the AIS data provided by the Danish Maritime Authority to isolate the travel of the different vessels and identify their origins and destinations. The objective of our process is to transform punctual information into 'course' objects. A course, as we define it, corresponds to the travel of a particular vessel between a start and end timestamps. All the observations corresponding to this course will be used to build a condensed 'course' object.

This processing pipeline has been developed according to several steps.

2.1 Initially, raw data of heterogeneous quality

The AIS data is supposed to be transmitted continuously and at regular time intervals. It contains information about the vessels (name, size, type...) and their position. These messages are received by other vessels and land-based antennas. The Danish Maritime Authority provides the historical AIS data they collect. We use several days of this data (DMA, 2021) to explore AIS information (*in accordance with the conditions for the use of Danish public data*).

The points are not usually sampled regularly. We have different tempos and many temporal "gaps" while locations are not transmitted. We also have outliers, where the ship signal is transmitted from locations far apart from other measurements that cannot be explained by realistic motion, which can then lead to misinterpretation of the displacement. Finally, a lot of information about the vessel (type, name...) is not available in all messages and these fields should be harmonized to analyze the AIS data.

https://doi.org/10.5194/isprs-archives-XLVI-4-W2-2021-71-2021 | © Author(s) 2021. CC BY 4.0 License.

2.2 Local location outliers

Geolocated objects usually obtain their positions through GPS equipment. This data is often noisy (approximate measurements) and occasionally irregular. We have therefore implemented a series of algorithms to denoise it and to identify outliers and ignore them.

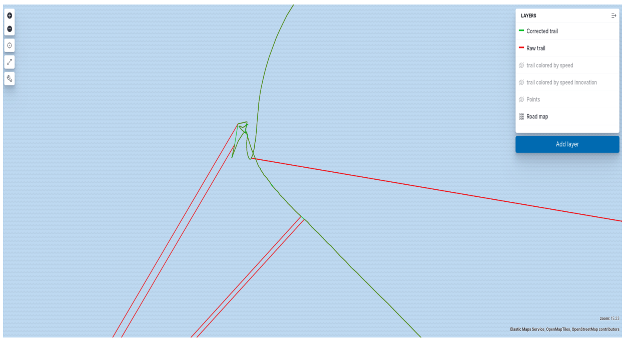


Figure 1. Examples of boat positions identified as outliers (red) and normal (green). The red lines illustrate the "jumps" generated by the GPS

Our filter derived from the Hampel filter (Hampel, 1974) uses median values calculated over sliding windows to identify local anomalies. These treatments allow us to identify and automatically correct most of the geographical outliers contained in AIS data. We adapt and change our algorithms when new types of anomalies are identified.

2.3 Temporal sampling

Once obtained by the ship's GPS system, its location is transmitted by the AIS system via VHF radio frequencies. Class A equipment is supposed to transmit its position every 10 seconds underway and every 3 minutes at anchor, while Class B equipment sends messages every 30 seconds. In practice, the transmission of positions is sometimes very irregular. We observe phases where geo-locations are provided with details, while there are also instances of missing signal.

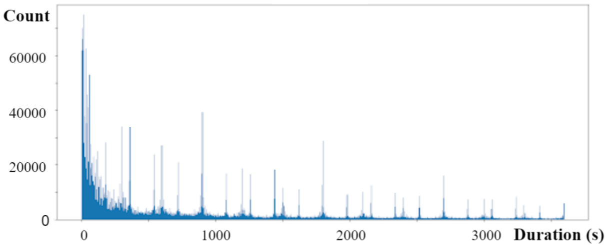


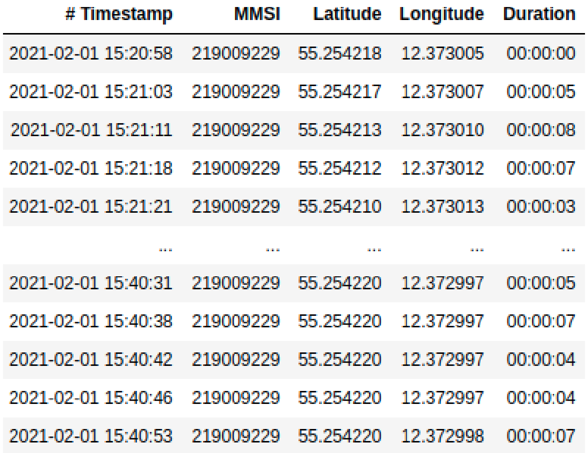
Figure 2. Distribution of durations separating two consecutive AIS messages of a vessel. The multiple peaks can correspond to different emission tempos.

The time separating two AIS messages of a vessel is very heterogeneous. Multiple peaks are distinguishable and they do not correspond to the normal ais time frequency (2-10s, 30s or 3min). It is necessary to detect and homogenise these different frequencies to facilitate the application of machine learning algorithms. Hidden Markov Model (Rabiner, 1986) -HMM- are used to identify the "tempos" in the broadcast sequences. Each tempo is modelled by a Markov state. Since the probabilities of transitions between states are lower than the probabilities of remaining in the same state, we have an inertia that avoids rapid transitions between different tempo's, thus adding some robustness to the noise of the measurements.

- Notion of fragments

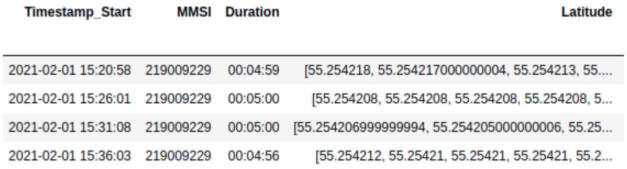
We call ‘fragment’ a portion of the travel of a vessel between two timestamps. The smallest fragment is the one between two AIS messages of a vessel

This type of data identification transmission times not only allows for application of diverse treatments to this particular transmission frequency, but also to harmonize and compress the information by reducing redundancy without loss of quality. Consecutive short fragments are then linked to form harmonized fragments of 5 minutes duration.



# Timestamp	MMSI	Latitude	Longitude	Duration
2021-02-01 15:20:58	219009229	55.254218	12.373005	00:00:00
2021-02-01 15:21:03	219009229	55.254217	12.373007	00:00:05
2021-02-01 15:21:11	219009229	55.254213	12.373010	00:00:08
2021-02-01 15:21:18	219009229	55.254212	12.373012	00:00:07
2021-02-01 15:21:21	219009229	55.254210	12.373013	00:00:03
...	...	...	...	...
2021-02-01 15:40:31	219009229	55.254220	12.372997	00:00:05
2021-02-01 15:40:38	219009229	55.254220	12.372997	00:00:07
2021-02-01 15:40:42	219009229	55.254220	12.372997	00:00:04
2021-02-01 15:40:46	219009229	55.254220	12.372997	00:00:04
2021-02-01 15:40:53	219009229	55.254220	12.372998	00:00:07

Figure 3. Example of AIS data (220 messages) for a particular vessel



Timestamp_Start	MMSI	Duration	Latitude
2021-02-01 15:20:58	219009229	00:04:59	[55.254218, 55.254217000000004, 55.254213, 55....
2021-02-01 15:26:01	219009229	00:05:00	[55.254208, 55.254208, 55.254208, 55.254208, 5...
2021-02-01 15:31:08	219009229	00:05:00	[55.254206999999994, 55.254205000000006, 55.25...
2021-02-01 15:36:03	219009229	00:04:56	[55.254212, 55.25421, 55.25421, 55.25421, 55.2...

Figure 4. Resulting resampled data (4 fragments) from example above

These extracts show the compression of data without losing spatial details. The sequences of latitude and longitude are stored and used to represent the vessel track.

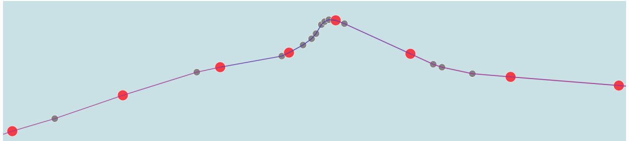


Figure 5. Example of resampled fragments ends (red) versus raw geopoints (grey). Resampled fragments are more homogenous along the vessel track.

These algorithms are designed and fully implemented by our engineers to be massively distributed on computer farms.

2.4 Movement detection

Before identifying complex movement patterns of a vessel or inferring its paths, it is necessary to automatically detect

whether the vessel is moving or stationary from its geolocation and dynamics data.
Sometimes the geolocated object also transmits its speed, but this is not always the case. Regardless of this aspect, an object is not necessarily moving when its velocity (also noisy signal) is not zero or when its position evolves. This detection must take into account the noise of the signal and the evolution of the position of the object when it is stationary. Here again we use Markov Chains (Figure 6) which provide a robust response to these variations (Figure 7).

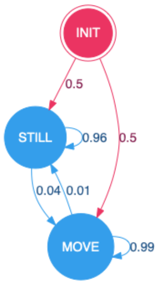


Figure 6. Graph of transition probabilities between the states still and move

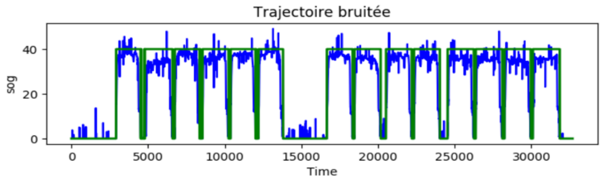


Figure 7. Time evolution of the speed (blue) of a ship making round trips between two ports. In green the result of the noise-robust stop/move phase classification.

Once the "still" and "move" states are identified, they are plotted on the actual ship tracks.

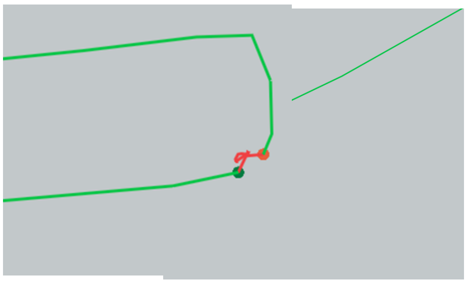


Figure 8. Examples of boat tracks identified in motion (green) and at rest (red) by **ARLAS PROCML.** On the left, we recognise the red trace of a boat held by an ink.

Another data compression step takes place to aggregate all the points associated with a stop and greatly reduce information redundancy.
This step therefore allows both the analysis of vessel mobility and the optimization of the space occupied by the data.

2.5 Travels Origins and Destination

Detecting the stops makes it possible to identify the movements between these areas as paths of the ship. Sometimes, certain routes are punctuated by shortstops, pauses, which must be identified so as not to interrupt what we call a "mission", a logical movement between an origin and the target destination.
The study of the origins and destinations of these journeys allows for a better analysis of the movement flows of the fleet. Clustering algorithms (unsupervised classification) such as DBScan or K-mean can be used to group trips with similar origins/destinations (Figure 9).

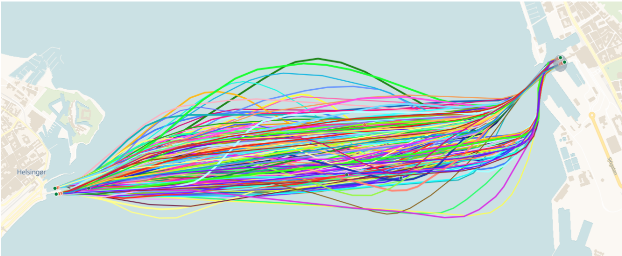


Figure 9. Set of similar routes between two ports

We used a geoservice to get the address of these origins and destinations location in order to get the names of the ports when available.

2.6 Machine learning for activity detection

Once the data has been cleaned, de-noised, re-sampled and the movement/stop phases detected, the result is a sequence of fragments for each vessel. Each fragment represents a portion of the object's movement between two instants. Once harmonised, the short fragments were concatenated to manipulate unified fragments of 5 minutes. These fragments are ideal models for the application of machine learning algorithms. It allows the dynamics of moving objects to be understood as time series. Activity detection, including supervised classification algorithms, exploits these time series in different ways.

The use of temporal convolutions like taking into account neighbouring fragments, allows the upstream and downstream historical context of the object's movement to be considered for prediction of the behaviour on the current fragment. The fragments are then ordered in time and a time window is applied to calculate indicators on each fragment sequence.

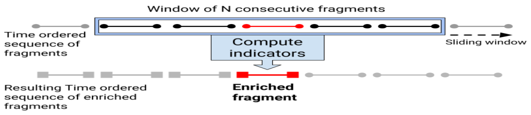

Window of N consecutive fragments
Time ordered sequence of fragments
Compute indicators
Sliding window
Resulting Time ordered sequence of enriched fragments
Enriched fragment

Figure 10. Schematic of processing using time convolution over fragments.

We can then calculate aggregations such as total distance, speed averages, and also direct distances between the beginning of the first fragment and the end of the last one. Such indicators can be computed over different window sizes, allowing the encoding of multiple movement dynamics. These enriched fragments are then provided as input to classification models.

The choice of the classification model and its parameterization are clearly essential in defining the behavior that needs to be identified. In a fishing detection challenge, our experience as well as reviewing scientific literature on the subject, informs our choice of several relevant algorithms (Random Forrest, XGBoost, Recurrent Neural Network, LSTM). In order to implement these models, we use Scikit-Learn, Keras/Tensorflow, SparkML libraries as well as some libraries that we developed to exploit HMM and kalman filters. We train the models on the dataset and experiment with multiple parameters. To be comparable, all models are evaluated using the same metrics. These metrics are computed by comparing the predicted classes to the real classes. All parameters describing the experiment and the metrics obtained are stored in the open-source MLFlow tool (MLFlow, 2021), which then allows all experiments to be plotted and compared with each other.

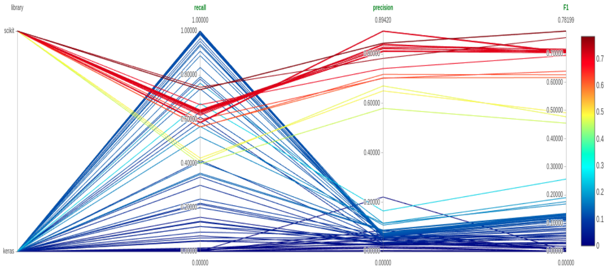


Figure 11. Comparison of experiments according to different metrics (recall, precision, F1-score) with MLFlow. Each line represents an experiment.

Each new attempt is recorded and the multiplication of experiments facilitates step by step convergence towards a machine learning algorithm that provides optimal classification results.

3. BIG DATA-PROOF ANALYSIS

3.1 Data compression

The creation of these 'course' objects allows an important compression of the information. Indeed, all the static information about the vessel is much less redundant and the aggregated dynamic information is much smaller. Moreover, we applied a geometry simplification algorithm to lighten the travel geometry without losing information. We also compute the geohashes covered by the travel geometry, to facilitate the geo exploration of these courses (only these geohashes could be enough to observe the flux, but we chose to keep also the detailed travel). From more than 31 million GPS locations representing 7 days of data, we eventually get around 14,000 course objects. In our context of Big Data analysis, this compression is really valuable.

3.2 A scalable process based on ARLAS PROC/ML

We've implemented the whole process described above as a suite of Apache Spark applications, relying on the open ARLAS PROC/ML framework from CSV files ingestion to ARLAS Exploration data loading. ARLAS PROC/ML is now open source and contains the different scalable processes that can be applied to mobile data. It allows the creation of processing pipelines adapted to the use case.

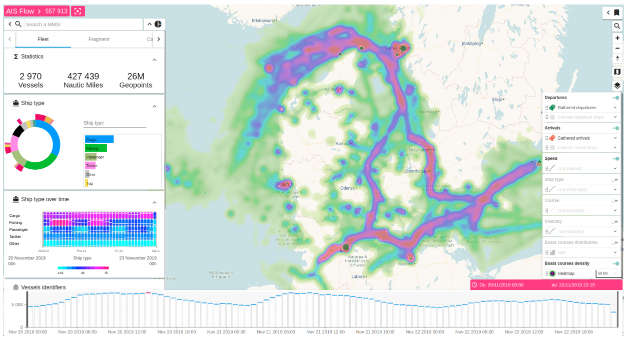
Depending on the step of the process, ARLAS PROC/ML is taking advantage of Spark to group AIS data by vessels and build their time series. It also extensively uses Spark's window computing to process specific parts of these time series with maximum efficiency in a distributed computing environment. As a result, data are partitioned differently at each step of the process according to this step's specificity. This makes the computation scalable by increasing (or decreasing) the number of executors that are able to compute partitions.

For instance, to get 14,000 courses from a dataset of 31 million GPS positions, the cluster was configured with 4 nodes of 8 cores and 52 Go memory, split into 6 executors (computation workers) of 4 cores and 30 Go memory. Since executors need sufficient memory on some steps of the process that require loading the entire vessel's time series, we had to limit parallelization with few executors to assign more resources for each one of them. That being said, the whole dataset was processed in about one day with this setup. If we need to ingest a larger dataset or to compute this one faster, we can add nodes to the computation cluster and/or give them more cores and memory.

This scalable processing chain gives us enriched data containing information such as the origin, destination, duration and distance of travel for all vessels, which can be really useful to analyze maritime traffic. However, the optimal exploitation of these results requires a visualization tool that suits the volume of data.

3.3 ARLAS Exploration, an open source geo-analytic tool

ARLAS Exploration is an open-source scalable solution to explore huge amounts of geolocated data. It is highly configurable thanks to a builder facilitating the creation of dashboards. We use this tool to analyze and understand the result of this processing chain (ARLAS Development Team, 2021b). ARLAS Exploration is a map-centered application which allows users to navigate in the map and explore the spatial dimension of the data. A system of aggregation is used to represent the flux of all the vessels.


AIS Flow > 557 913

Statistics: 2 970 Vessels, 427 439 Nautic Miles, 26M Geopoints

Du 20/11/2019 00:00 au 22/11/2019 23:15

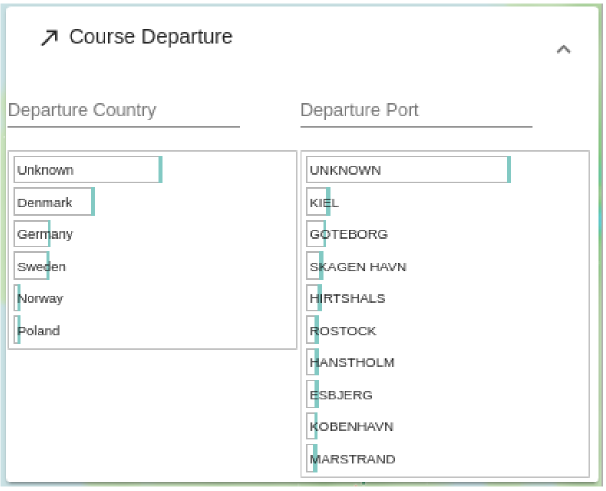
Figure 12. Data visualization of AIS courses with ARLAS Exploration

The temporal dimension of the data is analyzed by the timeline, which shows the temporal distribution of the courses and allows to select time ranges. The other components of the courses, such as mmsi, vessel type, duration or distance are visible in the widget space on the left of the application. These widgets make

https://doi.org/10.5194/isprs-archives-XLVI-4-W2-2021-71-2021

it possible to select parts of the data, and each filter updates the entire application. All these elements make the navigation easy and interactive, even with a huge amount of data.

It is possible to filter the courses on their origins or destinations.


↗ Course Departure

Departure Country

Unknown
Denmark
Germany
Sweden
Norway
Poland

Departure Port

UNKNOWN
KIEL
GOTEBORG
SKAGEN HAVN
HIRTSHALS
ROSTOCK
HANSTHOLM
ESBJERG
KOBENHAVN
MARSTRAND

Figure 13. Origins of Ship Trips in ARLAS Exploration. We can directly select trips starting from given ports or countries or type names in the search bar.

If we focus on the departures from the port of Kiel (Germany) for example, we directly visualize the main maritime roads linked to this port.

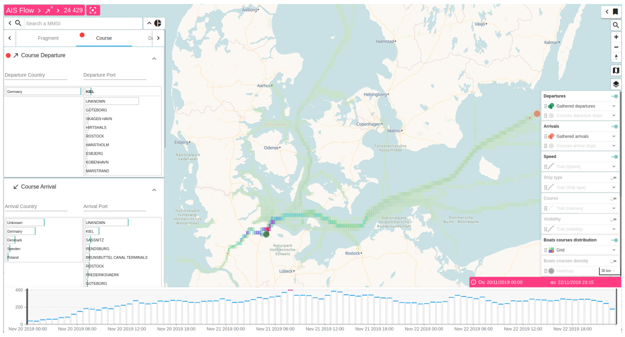


Figure 14. Selection of courses with departure in Kiel

We can also visualize the detailed itineraries starting from this port.

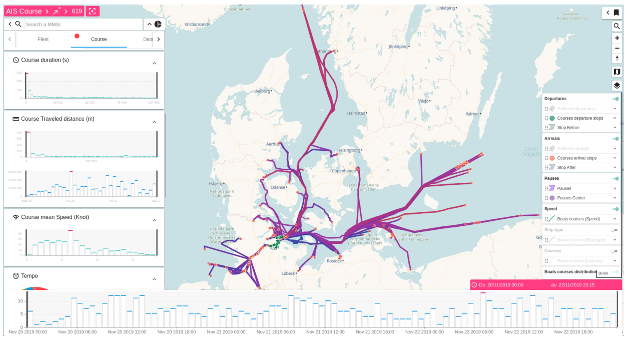


Figure 15. Plots of ship routes from the port of Kiel in ARLAS Exploration

In Figure 15, the itineraries are colored by their mean speed. If desired, courses may be colored from other attributes, mmsi or vessel type for instance. It is also possible to draw geographic selection. Only courses crossing this selection are selected (other rules can be chosen, such as keeping only the course starting inside the selection). We can also apply time selection thanks to the timeline cursors.

ARLAS Exploration works perfectly to explore AIS data and exploit the results of our origin/destination process. This open-source solution can also be used to explore any maritime data in a Big Data environment.

3.4 ARLAS Exploration for machine learning

Moreover, ARLAS Exploration provides labelling data functions, so users can set a label to a given selection. It helps data scientists to create and evaluate Machine Learning algorithms, particularly for detection/classification tasks. In the case of supervised learning, ARLAS allows data curators to create a training set thanks to this tagging system. It is possible to quickly select a part of the data with ARLAS Exploration and “label” it. It is particularly relevant for pattern detection. Experts can observe data, recognize patterns and set a suitable label (for fishing detection for instance).

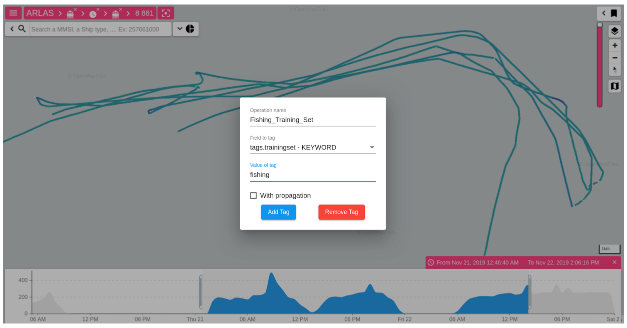

Operation name
Fishing_Training_Set
Field to tag
tags.trainingset - KEYWORD
Value of tag
fishing
□ With propagation
ARLAS
8 881
Search a MMSI, a Ship type, ... Ex: 257061000
Add Tag
Remove Tag
From Nov 21, 2019 12:46:40 AM
To Nov 22, 2019 2:06:16 PM
400
200
0

Figure 16. Example of use of the labelling system for fishing detection

Then we can download the labelled data with ARLAS API, available in Python, among others. These data are then used to train a Machine Learning model. In addition, it is possible to label the result of the algorithm prediction in ARLAS with the API, in order to explore the results and discover where classification mismatches the real behavior.

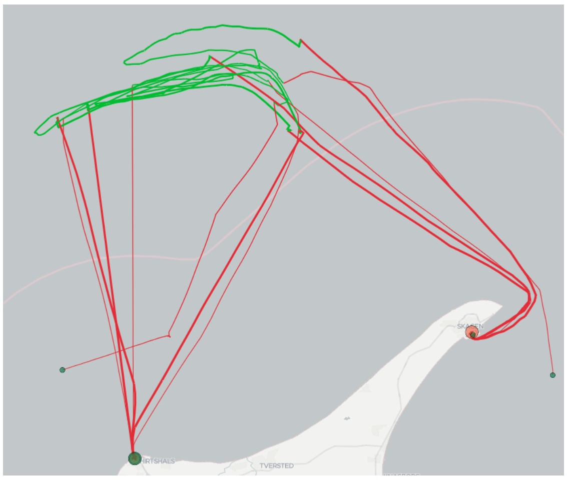


Figure 17. Examples of fishing vessel tracks identified in fishing (green) and moving (red) activity

In case of unsupervised learning, the manual labelling is not necessary, but the exploration of classification results is still valuable.

4. CONCLUSION

We were able to transform the data, compress it and isolate origin, destination and key indicators of travels from the dynamics of vessels. ARLAS PROC/ML allows us to apply this process to any amount of data.

Then, the result of this transformation can be discovered and analyzed with ARLAS Exploration, a powerful and open source geospatial analytic tool. Knowing the multiple origins of a vessel can help the authorities to adapt their goods control policy to sensitive places and better understand its frequentation. And the course statistics can help transporters to optimize their fleet activity.

Besides, ARLAS Exploration allows data scientists to label the data, which is useful for creating training sets. In addition, the ability to explore the results of predictions makes it a useful tool for developing Machine learning algorithms to effectively deal with maritime transportation challenges.

ACKNOWLEDGMENT

I would like to thank the Gisaïa team for this collective effort, particularly Laurent Dezou for leading a company that values innovation and research for societal issues, Mohamed Hamou for his wonderful help with ARLAS Exploration and Dolphine Rambaud for her advice in the redaction of this article.

REFERENCES

ARLAS Development Team (2021a), ARLAS software documentation, arlas.io

ARLAS Development Team (2021b), ARLAS demo with AIS course, demo.cloud.arlas.io

Danish Maritime Authority (2021), AIS Historical Data, www.dma.dk.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of the american statistical association, 69(346), 383-393.

MLFlow (2021), An open-source platform for the machine learning lifecycle, mlflow.org

Rabiner, L. R., & Juang, B. H. (1986). A tutorial on hidden markov models. IEEE ASSP Magazine, 3(1), 4-16.