STA4042 24-25T1 Project: AIS constructors clustering

Inquery: 223040237@link.cuhk.edu.cn

November 21, 2024

Important indications:

Your work on this project should be summerized in an all-in-one Jupyter notebook using Python with no more that 60 cells (code and markdown) – but it could be far less. You can work by binome (2 people max) and should not copy anything from an other team. You can use the method of any python library you find useful for your project.

We will evaluate your work followingly:

- Decision (global structure of your method): 50%
- Illustration: 20%
- Performance: 20%
- Clarity/Correctness/**Concision**: 20%

The project should be handled back before November, Friday the 15$^{\text{th}}$, at 23:59.

Automatic identification system (AIS)

Automatic identification system (AIS) is an is an automatic tracking system following international regulations that uses *transceivers* (also called *transponders*) on ships to assist a vessel's watchstanding officers and allow maritime authorities to track and monitor vessel movements.

AIS Transceivers share the radio bandwidth (161.975 MHz and 162.025 MHz) allocated to AIS operation using Time Division Multiple Access (TDMA) techniques. AIS typically operates on two parallel VHF Marine Band radio frequency channels and each channel is shared in time between multiple users by dividing channel access into 2250 'slots' per minute.

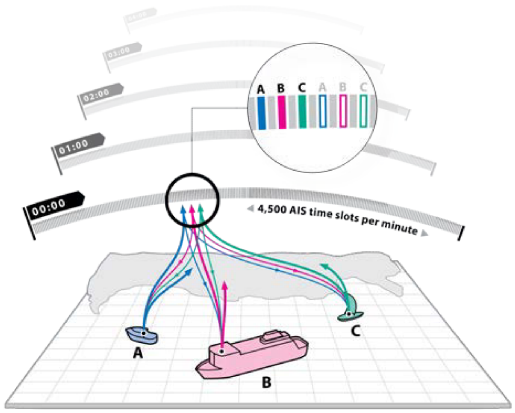
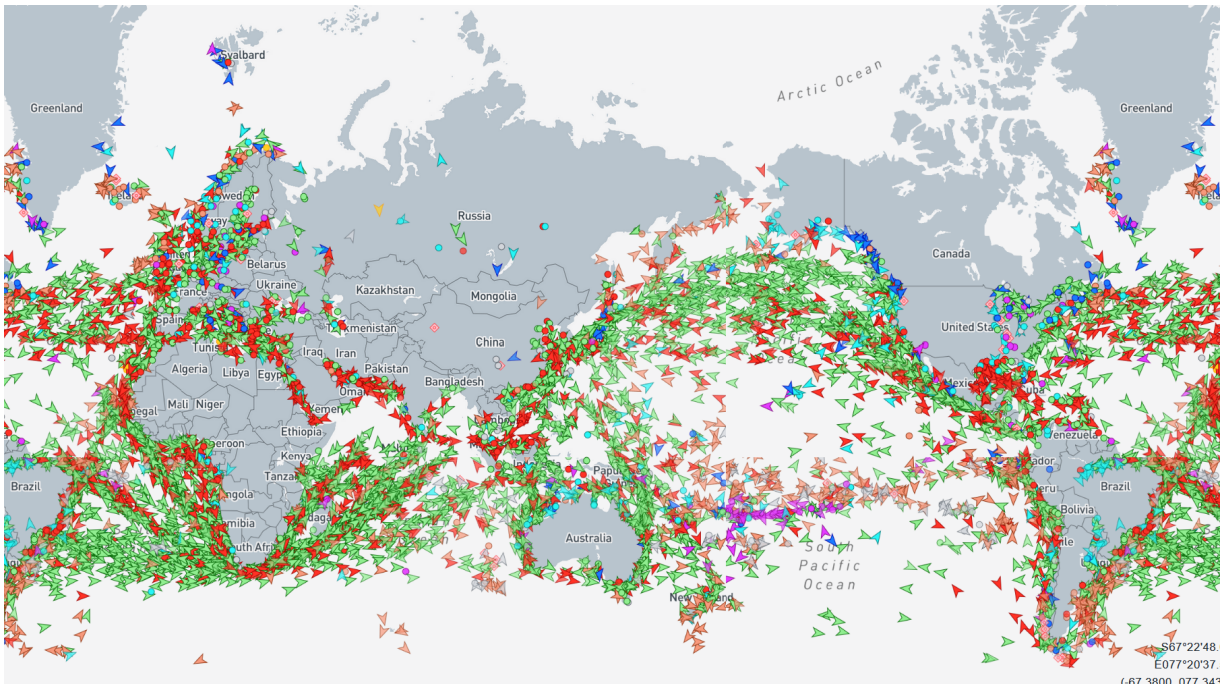


Figure 1: AIS TDMA system

In a typical TDMA system (such as GSM) a controlling entity (a GSM Base Station) is used to allocate transmission slots to each user. TDMA slot timing is derived from GPS UTC time ensuring all AIS *transceivers* share a common accurate time reference.

As AIS must operate far offshore the system cannot rely on a controlling entity to allocate time slots to each user. This means that each AIS transceiver must determine its own TDMA slot allocation, and critically, it must avoid using a slot that is in use by another vessel within reception range in order to avoid transmissions clashing. Extensive documentation on the different AIS standards can be found here: AISinfo



Description of the data

The data folder provided to you contains 425 csv, each csv corresponding to the signal received from one boat. You can find it following this link: data. In columns you can find:

1. `channelNumber`: One of the two channels.
2. `id`: Different behaviors of the boat. 1: straight line constant speed, 3: acceleration, 5: static information every 6min (departure, destination, boat identity).
3. `NavStatus`: 0 = under way using engine, 1 = at anchor, 2 = not under command, 3 = restricted maneuverability, 4 = constrained by her draught, 5 = moored... 15 = undefined.
4. `SlotOffset`: When `STO` = 0 difference between the new slot number and the old one plus 2250.
5. `SlotNumber`: Slot number in AIS message (between 0 and 2249 – a lot of `NaN` values).

6. `STO`: Initially sampled between 5 and 7 then incrementally decrease until it reached 0, then a new time slot must be chosen.

7. `SlotIncrement`: When acceleration (`id`= 3) need to produce message every 3.3s. Each channel produces a message every 6.6s, the slot increment is therefore generally around $\frac{6.6 \cdot 2250}{60} = 247$.

8. `KeepFlag`: Set to `TRUE` if the slot remains allocated for one additional frame.

9. `TS`: Computed Slot number (almost no `NaN` value).

10. `Lat`: Latitude.

11. `Lon`: Longitude.

12. `Sog`: Speed.

13. `thresholdSog`: Different discrete value depending on the speed (if `Sog`$\in [0, 3]$: 0, if `Sog`= 3: 1, if `Sog`$\in (3, 14)$: 2, if `Sog`= 14: 3, if `Sog`$\in (14, 23)$: 4, if `Sog`= 23: 5, if `Sog`$>$ 23: 6...)

14. `Course`: Direction (different with Heading, if there is some sea current).

15. `Heading`: Heading.

16. `changeHeading`: Change of Heading.

17. `RepeatIndicator`: Some rare messages sent when a boat transmits the message of an other boat.

18. `SpecialManoeuvre`: 0 = not available = default 1 = not engaged in special maneuver, 2 = engaged in special maneuver (i.e. regional passing arrangement on Inland Waterway).

19. `AISVersion`: Version of AIS used.

20. `toa`: Time of emission.

21. `DiffToa`: Difference in time with previous emission.

The -1 in the csv are non documented values (`NaN` values with `numpy`)

Globally the information sent by each boat looks the same because they should follow AIS regulations. But some differences can be noticed between AIS transponders for instance concerning:

- synchronization errors,
- regularity of messages depending on the speed/acceleration,
- emissions characteristics during special maneuvers,
- sampling of STOs and Slot numbers,

- reservation of a timeslot before sending a message with `id`= 5,
- alternation between channels.

These small differences are called "constructors signatures" and we will investigate them in this project.

Project's task

Our objective with this project is to confront you with real data and let you make decisions and be creative to study them efficiently. The final goal is to:

Cluster the different constructors of the AIS transponders

The main difficulties or challenges that you should keep in mind and solve are:

1. Maybe not all of the data are useful for the final objective.
2. Some data might need to be preprocessed.
3. Need to deal with nan values.
4. Some messages are lacking.
5. Remove or treat outliers.
6. Cluster the constructors and not the boat.
7. Number of classes not given, find arguments to justify your decision.
8. Since the problem is not supervised, find a method to evaluate the validity of your final solution.