Analytic Operators for Trajectories

ESR2.4: Song Wu, started on Sep 1, 2021

Supervisors: Esteban Zimányi (ULB), Mahmoud Sakr (ULB), Kristian Torp (AAU)

The 10th European Big Data Management & Analytics Summer School
The 1st DEDS Summer School
Cesena, Italy; July, 2022

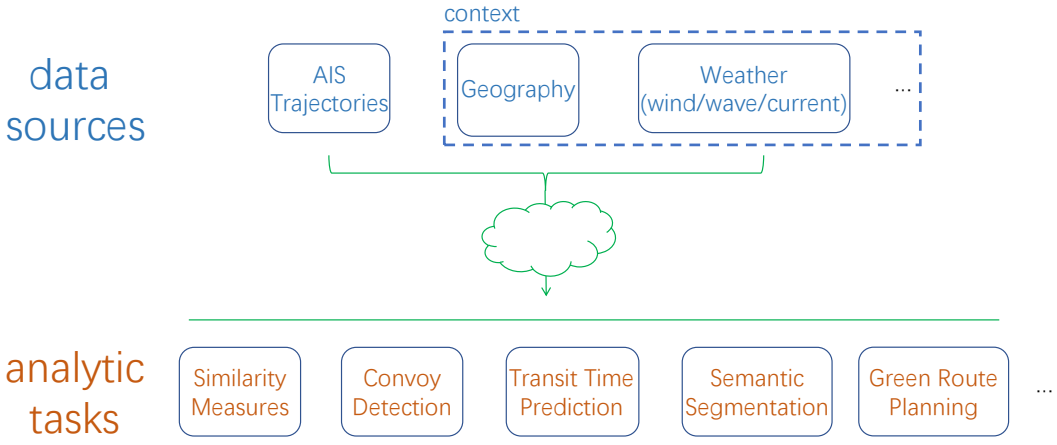
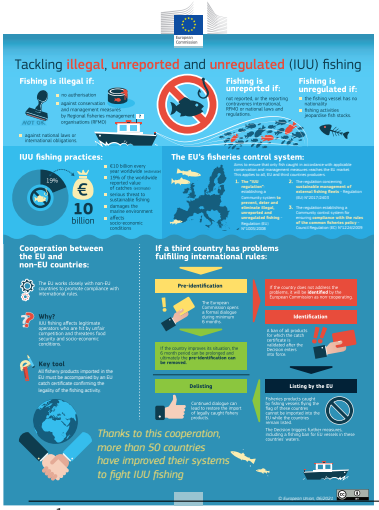

data sources
context
AIS Trajectories
Geography
Weather (wind/wave/current)
...
analytic tasks
Similarity Measures
Convoy Detection
Transit Time Prediction
Semantic Segmentation
Green Route Planning
...

Table of Contents

Illegal, unreported and unregulated (IUU) fishing is becoming an increasing concern[1].



IUU fishing does harm to:

- marine environment
- sustainable use of marine resources
- . . .

common types of IUU fishing:

- using unauthorized gear types
- fishing in prohibited water areas
- . . .

So it is important to know when & where a ship may have performed fishing activities.

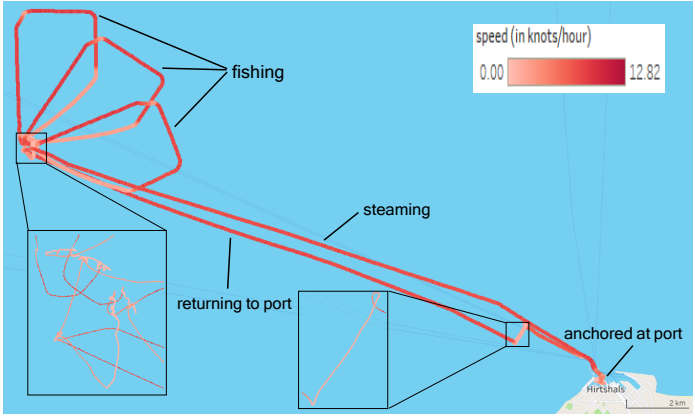
[1] https://ec.europa.eu/oceans-and-fisheries/fisheries/rules/illegal-fishing_en

Definition: A trajectory T is defined to be an ordered sequence of timestamped points $(p_1, t_1), (p_2, t_2), \cdots, (p_{n-1}, t_{n-1}), (p_n, t_n)$, where p_i is the coordinate of a moving object at the timestamp t_i.

Problem Statement: Given a trajectory T, this works aims to split T into a sequence of labelled segments $<(S_1, l_1), \ldots, (S_k, l_k)>$, where S_i is a continuous sequence of points in T, and l_i is the label of segment $\in \{\text{fishing, non-fishing}\}$.

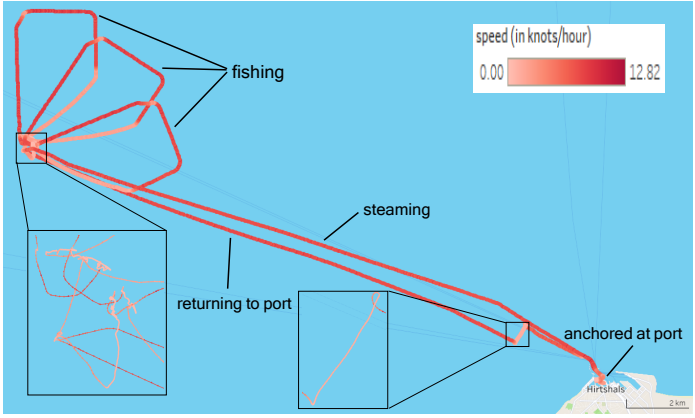
Limitations of existing trajectory segmentation algorithms:

- some work simply treat trajectories as a sequence of stops and moves, such as CB-SMOT[1] and DB-SMOT[2].
- segments are returned without labels, such as GRASP-UTS[3], W-Kmeans[4], SWS[5], WS-II[6].
- Many studies assume that that returned segments should have high homogeneity w.r.t. some spatiotemporal criteria or features of points, such as GRASP-UTS[3].



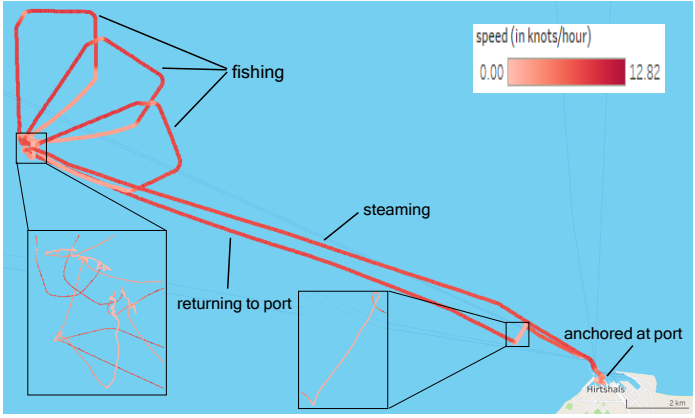
Limitations of existing trajectory segmentation algorithms:

- some work simply treat trajectories as a sequence of stops and moves, such as CB-SMOT[1] and DB-SMOT[2].
- segments are returned without labels, such as GRASP-UTS[3], W-Kmeans[4], SWS[5], WS-II[6].
- Many studies assume that that returned segments should have high homogeneity w.r.t. some spatiotemporal criteria or features of points, such as GRASP-UTS[3].



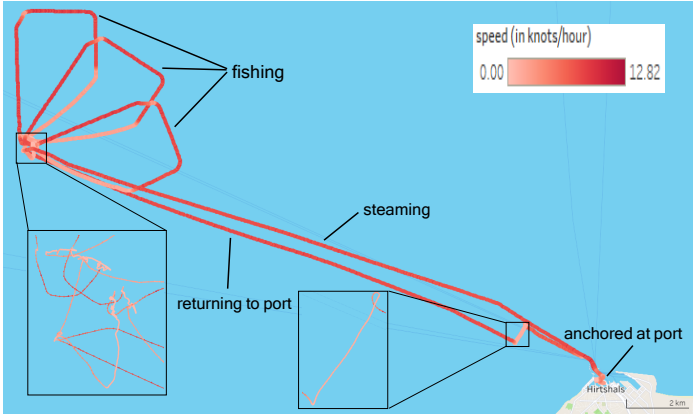
Limitations of existing trajectory segmentation algorithms:

- some work simply treat trajectories as a sequence of stops and moves, such as CB-SMOT[1] and DB-SMOT[2].
- segments are returned without labels, such as GRASP-UTS[3], W-Kmeans[4], SWS[5], WS-II[6].
- Many studies assume that that returned segments should have high homogeneity w.r.t. some spatiotemporal criteria or features of points, such as GRASP-UTS[3].



Limitations of existing trajectory segmentation algorithms:

- some work simply treat trajectories as a sequence of stops and moves, such as CB-SMOT[1] and DB-SMOT[2].
- segments are returned without labels, such as GRASP-UTS[3], W-Kmeans[4], SWS[5], WS-II[6].
- Many studies assume that that returned segments should have high homogeneity w.r.t. some spatiotemporal criteria or features of points, such as GRASP-UTS[3].



Our approach: a combination of ML and run-length encoding

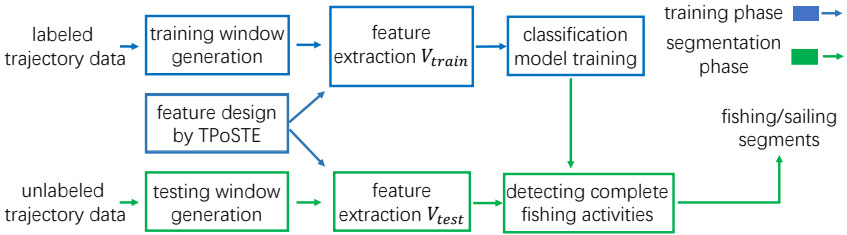
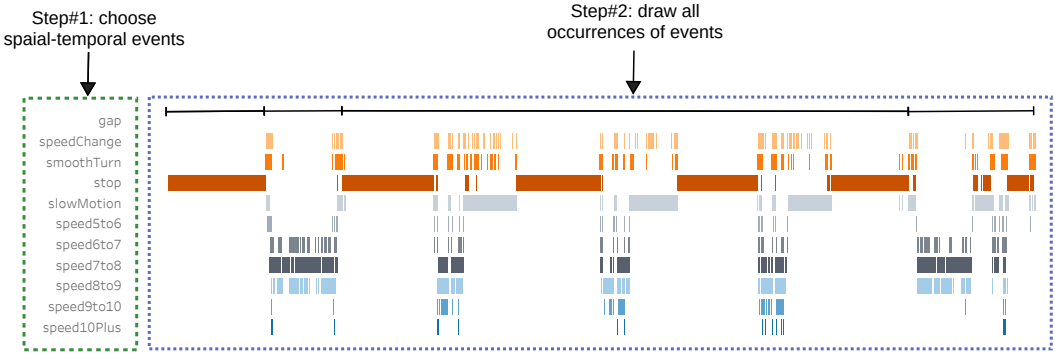


Figure: overall workflow of WBS-RLE



Step#3: gain some insights that help design features capturing movement patterns

34 features are designed using this technique.

Part 1 Methodology: Window-Based Trajectory Segmentation using Run-Length Encoding (WBS-RLE)

Window Generation

- a window is required to contain at least $\boldsymbol{size_w}$ points and its duration is larger than a time threshold $\boldsymbol{t_w}$.
- two adjacent windows have some overlap indicated by ***ratio***.

Run-Length Encoding technique

- an alternating sequence of counts $\ldots, a_{fishing}, b_{sailing}, c_{fishing}, \ldots$ is obtained from the labeled windows.

A complete fishing activity A is a maximal subsequence of counts that:

- A starts and ends with fishing counts.
- each triplet $<a_{fishing}, b_{sailing}, c_{fishing}>$ in A fulfills $a \geq b$ and $b \leq c$ to correct occasional classification errors.

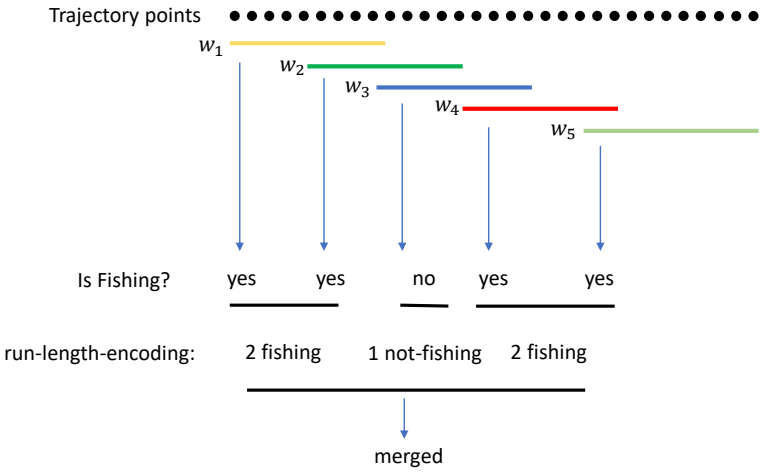
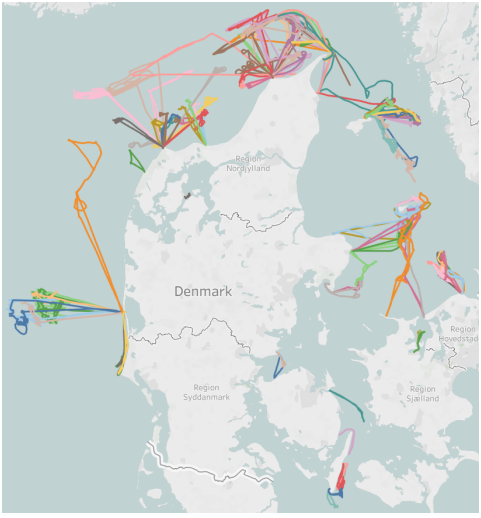


Figure: illustration of WBS-RLE

We manually labeled 128 trajectories between Nov 14, 2021 and Nov 20, 2021 that are likely to contain fishing activities.

- publicly available from Danish Maritime Authority
- average sampling frequency: 10.63 seconds
- # of points: 1,080,220



Among the 128 trajectories, 31 used for training, 97 used for testing (available online)

WBS-RLE achieves:

- the highest harmonic mean
- the closest number of segments w.r.t ground truth, except WKMeans (k=3)

method	purity [3]	coverage [3]	harmonic mean	# of segments
WBS-RLE	0.890	0.974	0.927	2.670
CB-SMoT [1]	0.859	0.885	0.859	5
WKMeans (k=3) [4]	0.878	0.840	0.855	3
WKMeans (k=6) [4]	0.932	0.619	0.741	6
SWS [5]	0.954	0.759	0.837	9.855

Table: average performance on the 97 trajectories

* all segmentation results are available online

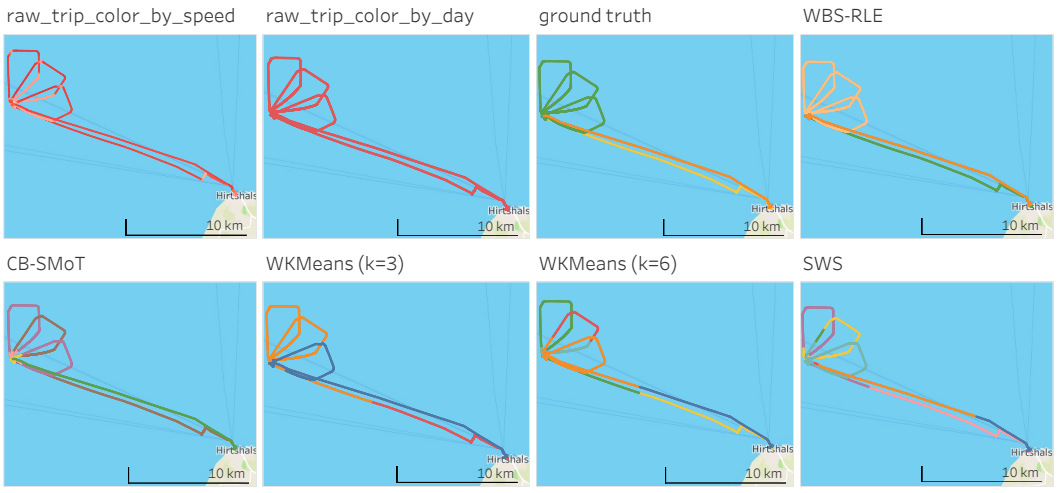


Figure: Segmentation result for the trajectory #220051000-2

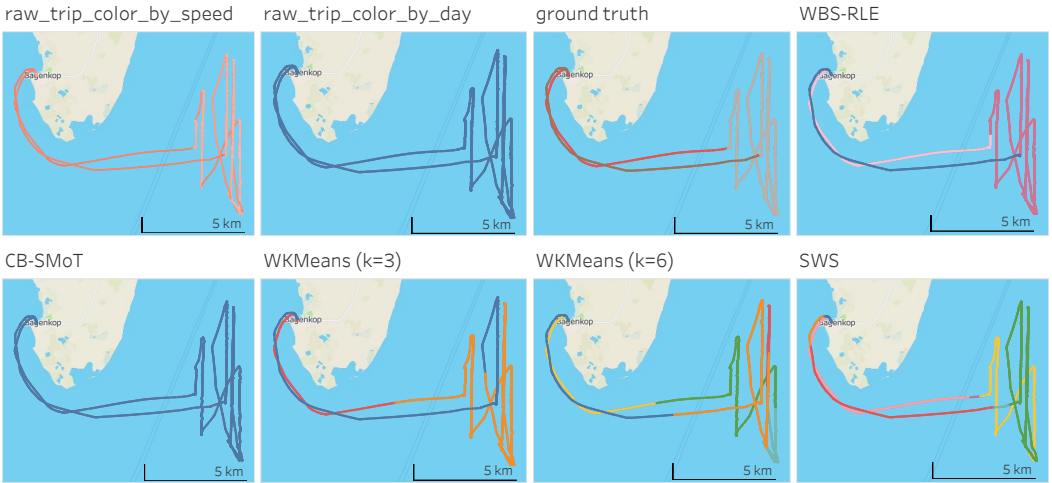


Figure: Segmentation result for the trajectory #219002136-3

One weakness of WLS-RLE: sometimes it returns the whole trajectory as a fishing activity when:

- the sampling frequency is low on some local parts
- a ship returns to the harbour at low speeds

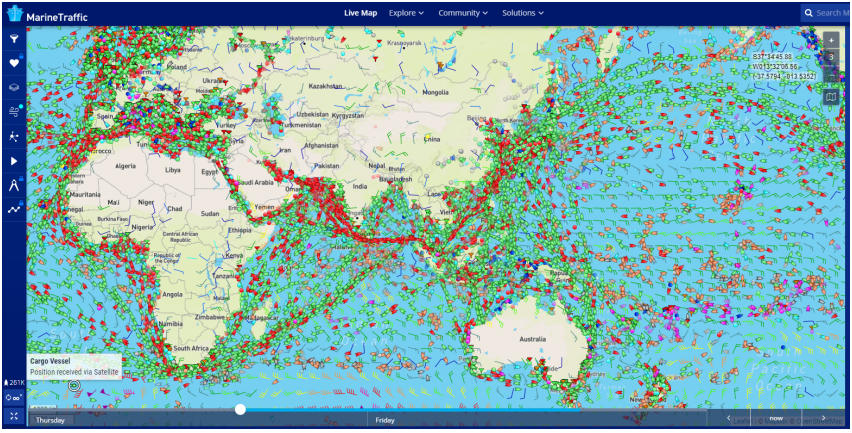
Actually, WLS-RLE returned 1 or 2 segments for 23 of the 97 testing trajectories.

Further improvements include:

- involvement of domain experts for ground-truth dataset
- evaluation on different parameter settings and datasets
- consideration of more events/features, e.g. distance to shore, sinuosity and temporal gap, etc.

What is MarineTraffic?

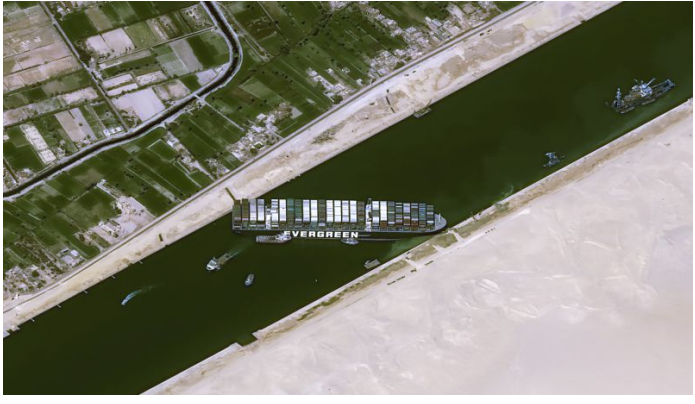
- the world's leading provider of ship tracking and maritime intelligence
- It tracks more than 200K vessels around the globe



Why focus on the Suez Canal context?

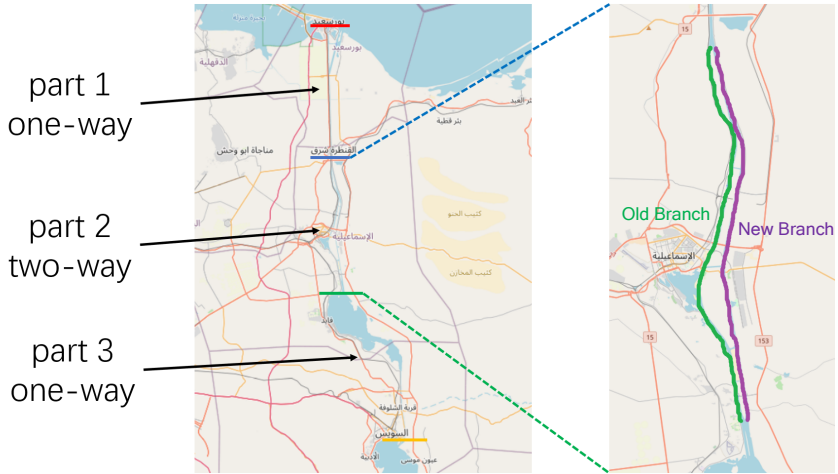
- One of the busiest water channels in the world
- Ship captains need estimates to assist decision-making
- Estimated Time of Arrival (ETA) is one of the most important products delivered by MarineTraffic

In 2021, the Suez Canal was blocked for one week by a giant container ship.



General Problem: How long does it take to transit through the canal?

Based on whether two-way traffic is allowed, Suez Canal can be divided to three parts.



Main Traffic Rules:

- Ships transit through the Suez Canal in the form of **convoys**.
- An **early group** may enter the canal according to the traffic situation, and **then join tail of the convoy later**.

Limit Time of Arrival:

- ships arriving before 23:00 can join the convoy
- ships arriving between 23:00 and 01:00 can join the convoy at a surcharge
- ships arriving after 01:00 can join the convoy at a surcharge, if traffic allows

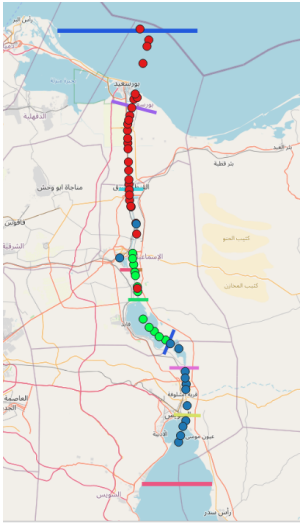
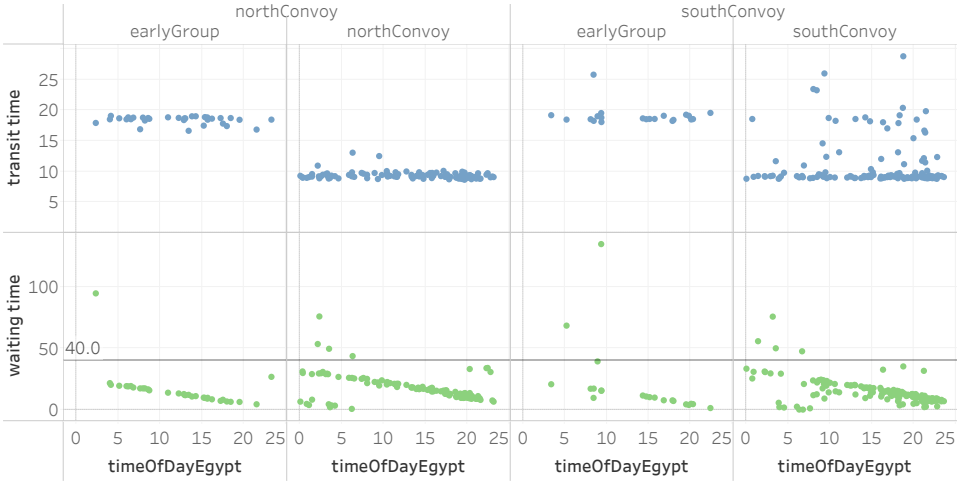


Figure: Convoy on March 21

The total transit time can be divided into two parts.

- **Waiting Time:** the time difference between arrival and entering the canal.
- **Transit Time:** the time taken to go from one end of the canal to the other.

Dataset: one-week AIS data from March 16 to March 22 provided by MarineTraffic



Future improvements include:

- the model will be evaluated on AIS data from July 2021 to Dec 2021.
- reasons behind outliers can be investigated.

To support finer analysis, convoys need to be detected automatically.

Why study the convoy detection problem?

- From the perspective of vessel trackers (e.g. MarineTraffic), it helps make more accurate predictions of ETA, etc.
- However, convoy scheduling is local knowledge: it is planned by the SCA and announced to the passing ships.

In the literature,

- most work regarding convoy detection focus on the algorithm **efficiency**, e.g. ECMA [7] and k/2-hop [8].
- less attention is given to the **quality** of returned convoys.

Three definitions of convoy pattern:

1. **convoy** [9]: a group of at least m objects that are density-connected during at least k consecutive timestamps
2. **valid convoy** [10] and **Fully Connected Convoy** [8] additionally require that convoy objects are density-connected by themselves.
3. **evolving convoy** [11] allows objects to join/leave a convoy.

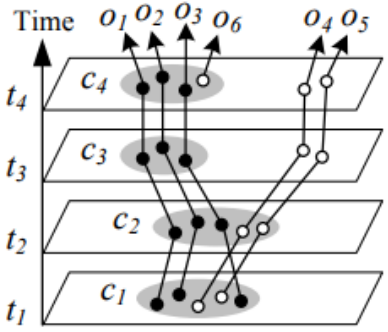


Figure: illustration of invalid convoy: <{o1,o2,o3},[t1 ,t4]> (taken from [10])

Experiments: detecting evolving convoys in the Suez Canal

- gap between 2 timestamps: 10 minutes
- minpts $= 5$, epsilon $= 8$ kilometers
- m $= 6$, k $= 4$, w $= 6$

two shortcomings:

- many anchored points are included in results
- south and north convoys are mixed with each other

Experiments: two minor improvements

- At each timestamp, only ships with an average speed larger than 1 knot/h (in the last 10 minutes) are clustered
- The definition of neighborhood is changed such that a point p_1's neighbor p_2 should have a heading difference smaller than the threshold $angleTH$

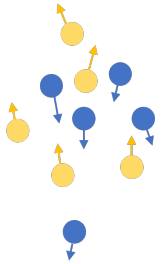


Figure: heading-aware DBSCAN

1. **49 convoys are returned with parameters:**
   - gap between 2 timestamps: 10 minutes
   - minpts = 5, epsilon = 15 kilometers, angle = 30°
   - m = 6, k = 4, w = 6

2. remaining issue: undesired cluster split leads to redundancy in the results

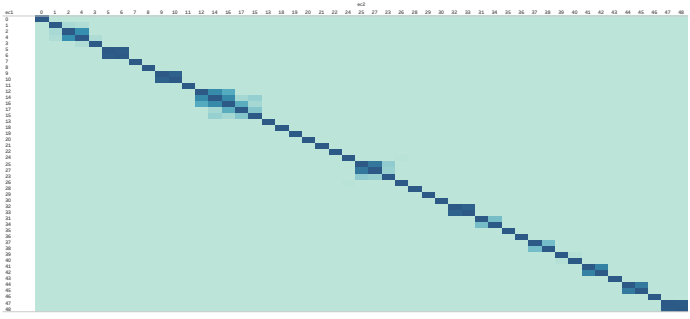


Figure: pairwise similarity of evolving convoys

3. **Goal: a sound definition of 'convoy'**

thank you! feel free to ask questions

References I


A. T. Palma, V. Bogorny, B. Kuijpers, and L. O. Alvares, "A clustering-based approach for discovering interesting places in trajectories," in *Proceedings of the 2008 ACM SAC*, p. 863–868, ACM, 2008.


J. A. M. R. Rocha, V. C. Times, G. Oliveira, L. O. Alvares, and V. Bogorny, "DB-SMoT: A direction-based spatio-temporal clustering method," in *2010 5th IEEE International Conference Intelligent Systems*, pp. 114–119, 2010.


A. Soares Júnior, B. N. Moreno, V. C. Times, S. Matwin, and L. d. A. F. Cabral, "Grasp-uts: an algorithm for unsupervised trajectory segmentation," *International Journal of Geographical Information Science*, vol. 29, no. 1, pp. 46–68, 2015.


L. A. Leiva and E. Vidal, "Warped k-means: An algorithm to cluster sequentially-distributed data," *Information Sciences*, vol. 237, pp. 196–210, 2013.


M. Etemad, A. Soares, E. Etemad, J. Rose, L. Torgo, and S. Matwin, "Sws: an unsupervised trajectory segmentation algorithm based on change detection with interpolation kernels," *GeoInformatica*, vol. 25, no. 2, pp. 269–289, 2021.

References II


M. Etemad, Z. Etemad, A. Soares, V. Bogorny, S. Matwin, and L. Torgo, "Wise sliding window segmentation: A classification-aided approach for trajectory segmentation," in *Canadian Conference on Artificial Intelligence*, pp. 208–219, Springer, 2020.


Y. Liu, H. Dai, B. Li, J. Li, G. Yang, and J. Wang, *ECMA: An Efficient Convoy Mining Algorithm for Moving Objects*, p. 1089–1098.
New York, NY, USA: Association for Computing Machinery, 2021.


F. Orakzai, T. Calders, and T. B. Pedersen, "K/2-hop: Fast mining of convoy patterns with effective pruning," *Proc. VLDB Endow.*, vol. 12, p. 948–960, may 2019.


H. Jeung, M. L. Yiu, X. Zhou, C. S. Jensen, and H. T. Shen, "Discovery of convoys in trajectory databases," *Proc. VLDB Endow.*, vol. 1, p. 1068–1080, aug 2008.


H. Yoon and C. Shahabi, "Accurate discovery of valid convoys from moving object trajectories," in *2009 IEEE International Conference on Data Mining Workshops*, pp. 636–643, 2009.


H. H. Aung and K.-L. Tan, "Discovery of evolving convoys," in *Proceedings of the 22nd International Conference on Scientific and Statistical Database Management*, SSDBM'10, (Berlin, Heidelberg), p. 196–213, Springer-Verlag, 2010.

The following 34 features have been tried in this work:

- *smoothTurn*: Turn Frequency
- *speedChange*: Speed-Change Frequency
- *the remaining eight speed-related events*:
  - Average Duration = (dur(e^1) + dur(e^2)) / 2
  - Duration Ratio = (dur(e^1) + dur(e^2)) / (t_{end}-t_{start})
  - Spanning Ratio = (e^2_{end} - e^1_{start}) / (t_{end}-t_{start})
  - Frequency of Events = 2 / (t_{end}-t_{start})

start/end of a (sub-)trajectory

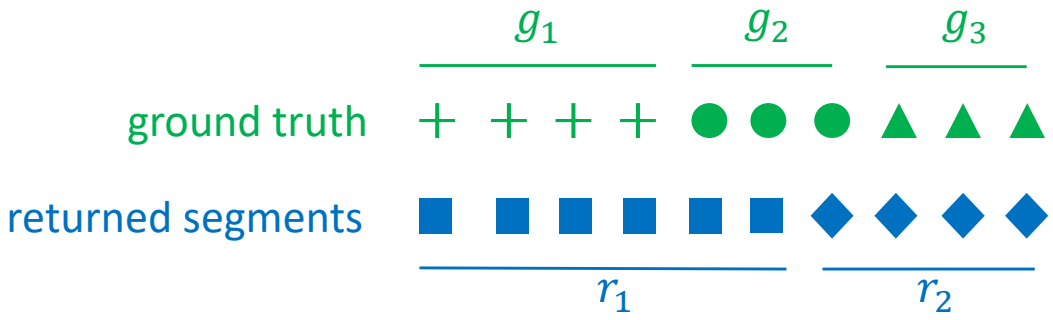
start/end of 1st occurrence of an event

start/end of 2nd occurrence of an event

three state-of-the-art for comparison: CB-SMoT, W-Kmeans, SWS

Four metrics are used:

- Purity (introduced in [3])
- Coverage (introduced in [3])
- Harmonic Mean of Purity and Coverage
- # of returned segments



Purity = (4/6 + 3/4) / 2
Coverage = (4/4 + 2/3 + 3/3) / 3

Estimation of Waiting Time by Linear Regression (**LRall** and **LR40**)

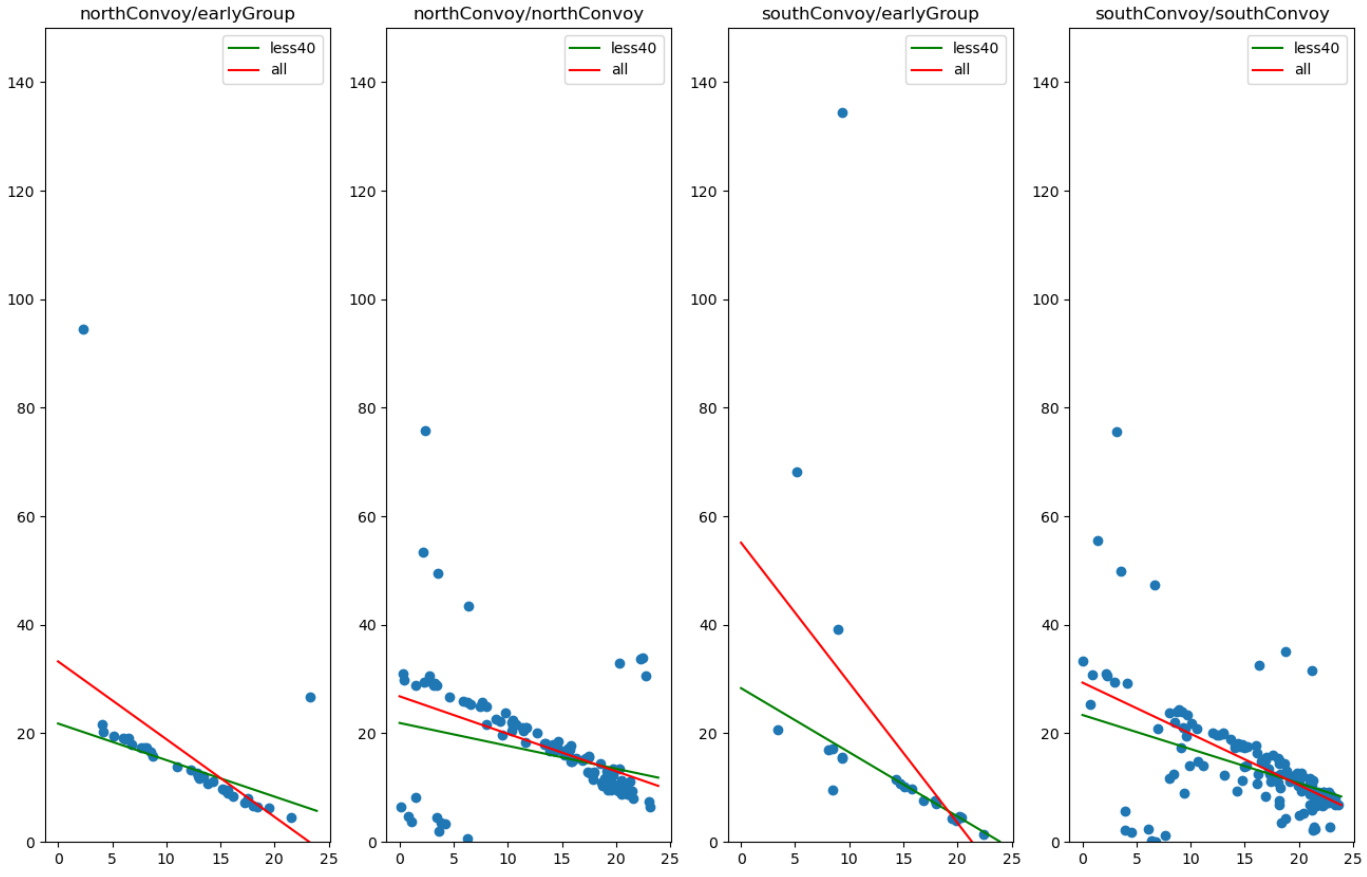


Figure: waiting time fitted by linear regression

Estimation of Waiting Time by Linear Regression (**LRall** and **LR40**)

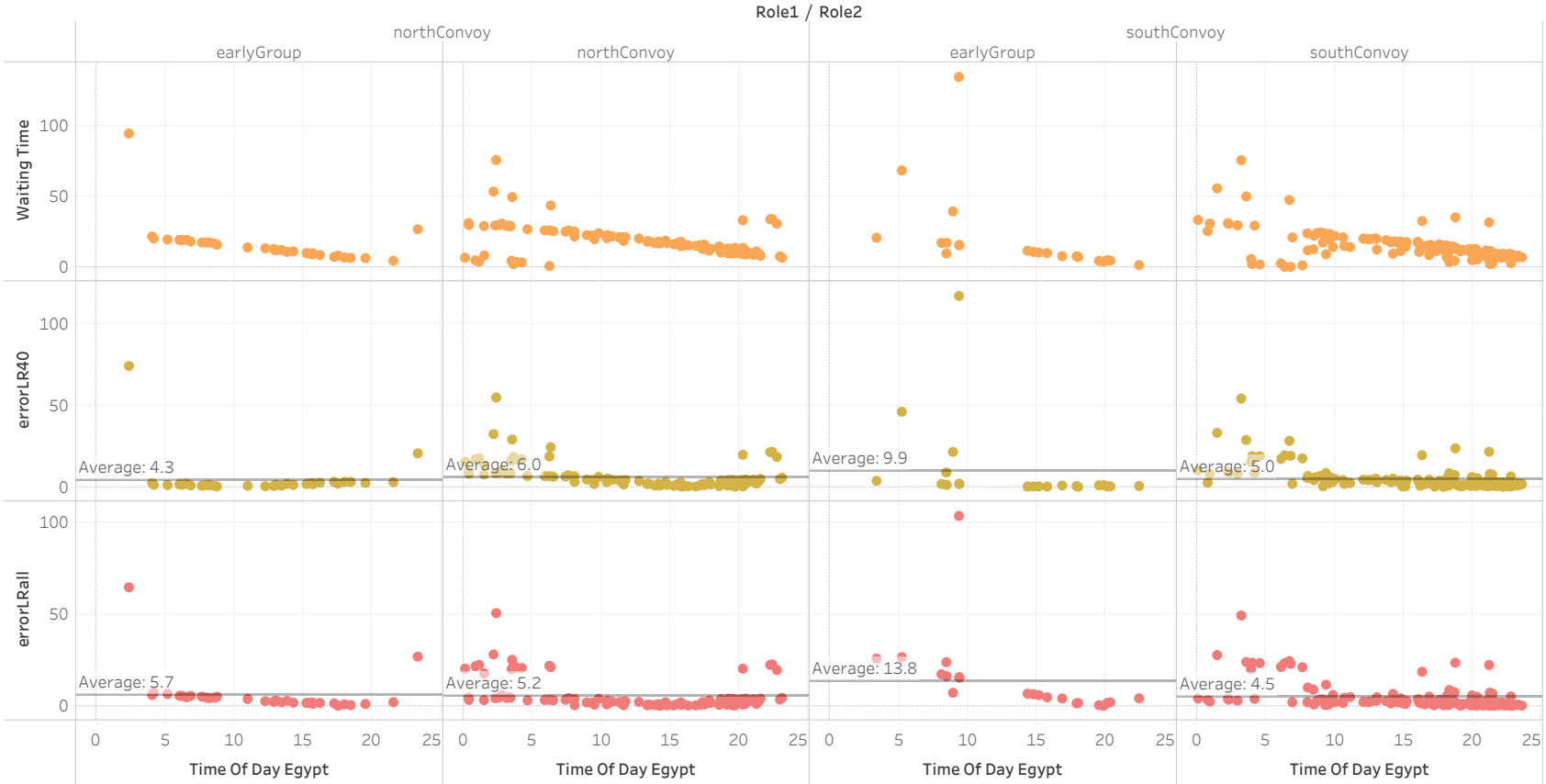


Figure: mean absolute error of waiting time using linear regression