Detection and correction of mover identity problems in movement datasets

Nevil Jose Kodiyan

Supervisors:

Dr.-Ing. Mathias Jahnke

Barend Köbben

Dr. Georg Fuchs

Outline

- Background
- Related Work
- Movement data quality
- Methodology
  - Algorithms
  - Visual Analytic strategy
- Evaluation
- Results
- Conclusion
- Future work

Background

- Big Data – Four v's
  - Volume, Variety, Velocity, Veracity
- Movement data
  - Pedestrians
  - Animals
  - Aviation
  - Maritime
- datACRON project

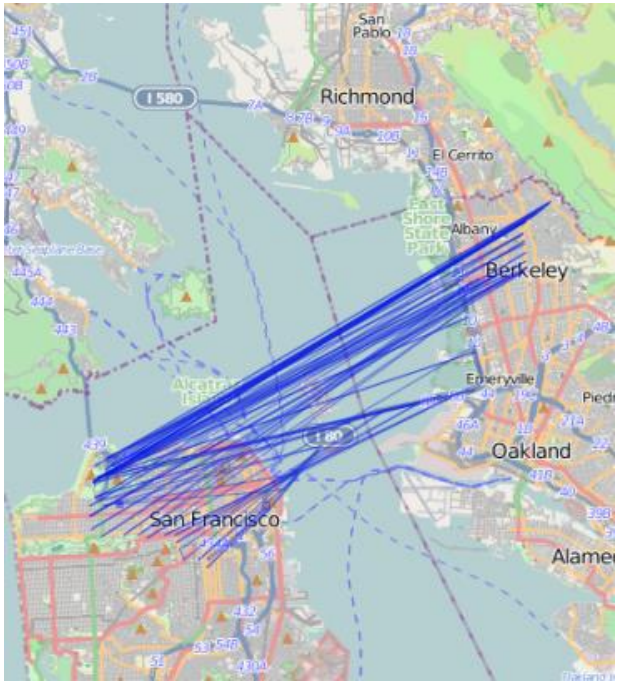
Background

- Example of movement data

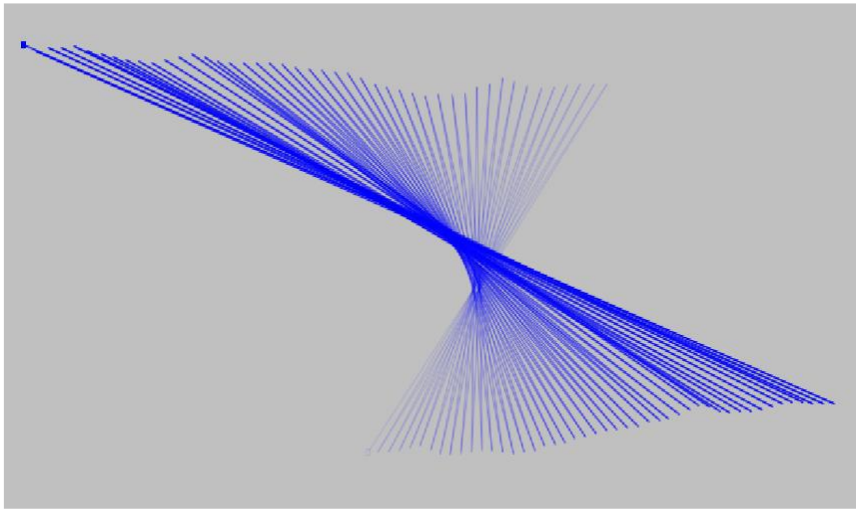
MOVER_ID	TIMESTAMP	LON	LAT	SPEED	HEADING
0_ 1	01-10-15 1:58	114.75152	30.6056	7.9	251.9
0_ 1	01-10-15 2:00	114.7494	30.60493	7.7	246.9
0_ 1	01-10-15 2:08	114.74012	30.60043	8.1	239.7
0_ 1	01-10-15 2:08	114.73996	30.60034333	8.1	239.3
0_ 1	01-10-15 2:08	114.739173	30.59998667	6.9	244.5
0_ 1	01-10-15 2:09	114.739013	30.59991667	8.0	243.8
0_ 1	01-10-15 2:09	114.738707	30.59977667	9.1	241.7
0_ 1	01-10-15 2:09	114.738547	30.59969667	7.9	241.9
0_ 1	01-10-15 2:09	114.737907	30.59943	8.9	245.8
0_ 6	07-10-15 17:21	114.251783	30.50421333	0.9	0.0
0_ 6	07-10-15 17:22	114.251793	30.50428333	60.5	0.0
0_ 6	07-10-15 17:48	114.416587	30.69275167	7.0	250.0
0_ 6	07-10-15 17:48	114.41637	30.6927	4.5	249.8
0_ 6	07-10-15 17:48	114.416332	30.692695	6.9	249.9
0_ 6	07-10-15 17:48	114.416105	30.69262333	4.3	249.7
0_ 6	07-10-15 17:48	114.416058	30.69260833	6.0	249.6
0_ 6	07-10-15 17:48	114.415965	30.69256833	6.2	249.4
0_ 6	07-10-15 17:48	114.415933	30.69255333	8.8	249.3
0_ 6	07-10-15 17:48	114.415862	30.69253	4.9	249.3

Background

- Ramification of mover identity inaccuracy



Example of two movers in different cities

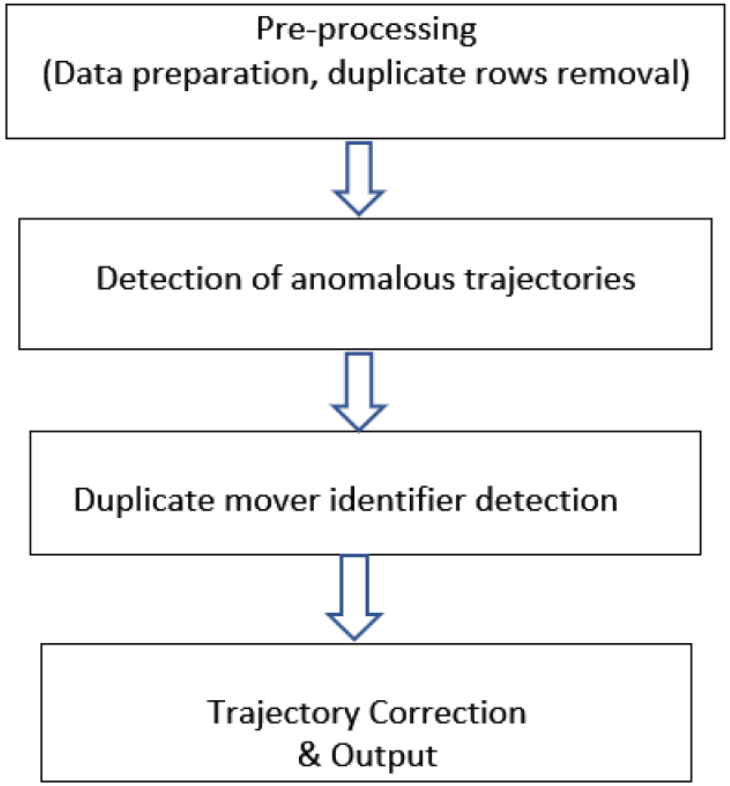


Simulation of two movers moving in opposite directions

Objectives

- Develop an algorithm to detect mover identity errors.
- Develop a visual-analytic strategy to facilitate parameter discovery.
- Correct erroneous trajectories.

Working Steps



Related Work

- Anomalous trajectory detection approaches
  - Survey of trajectory data mining (Zheng, 2015)
  - Pattern mining approach
    - Building up patterns of motion from historical data
  - Statistical approach
    - Deviation from models of normal motion
  - Heuristic approach
    - Critical points (stop, turn, slow motion)

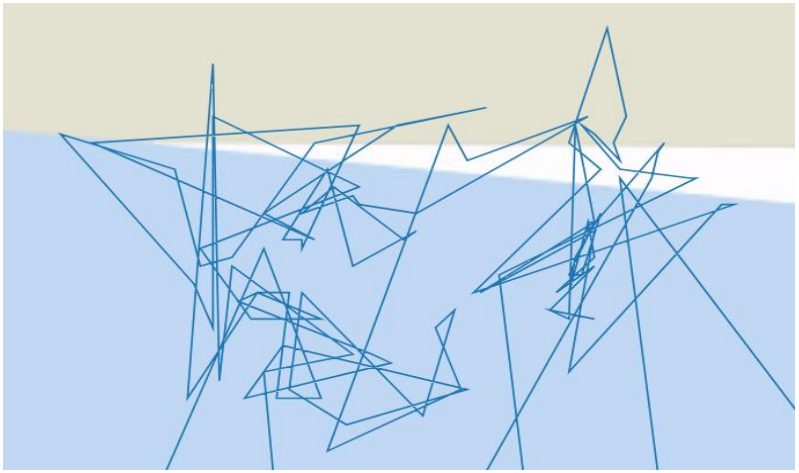
Movement data quality

- Missing data problem (spatial/temporal gaps)
- Accuracy problems (inaccuracies in identifier, position, timestamp, or other attributes)
  - Mover identity error
  - Spatial errors
  - Temporal errors
  - Attribute errors
- Precision deficiency (limitations of the positioning system)

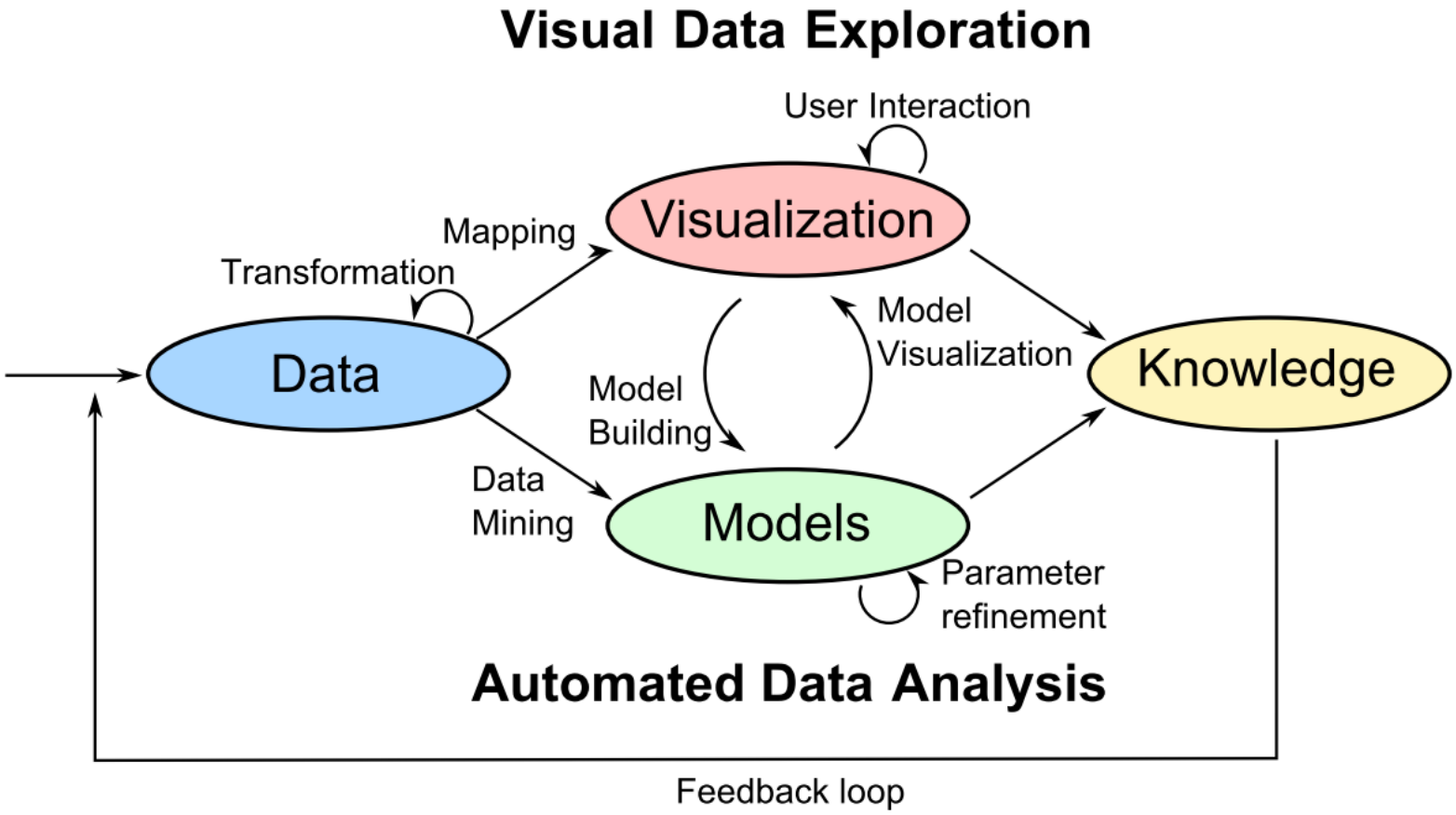
Related classes of errors

Cut-off error

Jitter



Need for visual analytics approach



Methodology (algorithmic component)

1. Pre-processing
2. Detection of anomalous trajectories
3. Correction of anomalous trajectories

Methodology (algorithmic component)

Pre-processing steps

- Removal of duplicate rows
- Conversion to Web Mercator projection
- Conversion to Hierarchical Data Format

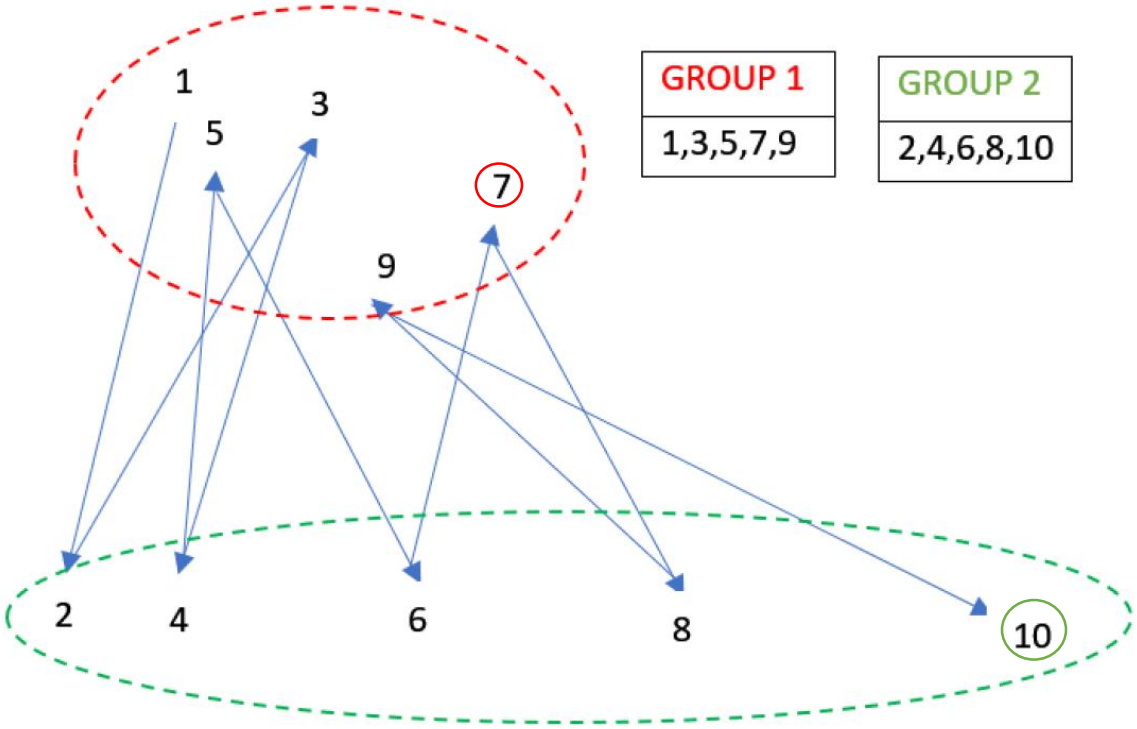
Methodology (algorithmic component)

Detection of anomalous trajectories

- Simple thresholding based on mean velocity
- Other properties like total distance covered can also be used

Methodology (algorithmic component)

Correction of anomalous trajectory



Property	Description
Last point	Latitude, longitude of the last inserted point
Last point timestamp	Timestamp of the last inserted point
Points	List of all points belonging to this group
Angles	List of angle changes between subsequent points in this group

Methodology (algorithmic component)

R-Tree spatial index data structure for fast lookups

- Number of trajectories and points in trajectories is non-trivial.
- R-tree data structure is utilized to quickly find the nearest matching candidate group to a candidate point.

Methodology (visual analytic component)

1. Data loader interface
2. Duplicate mover detection interface
3. Cut-off error removal interface

Methodology (visual analytic component)

Data loader interface

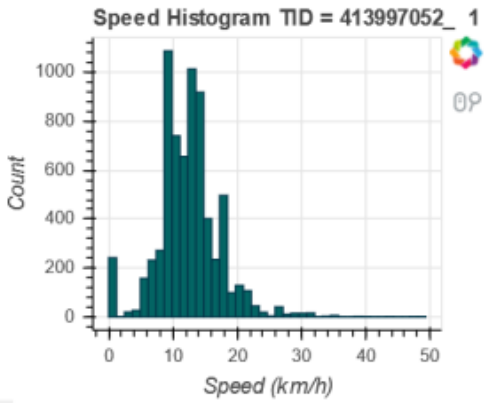
Maritime dataset loader

Enter Filename

D:/_Projects/traj/Wuhar

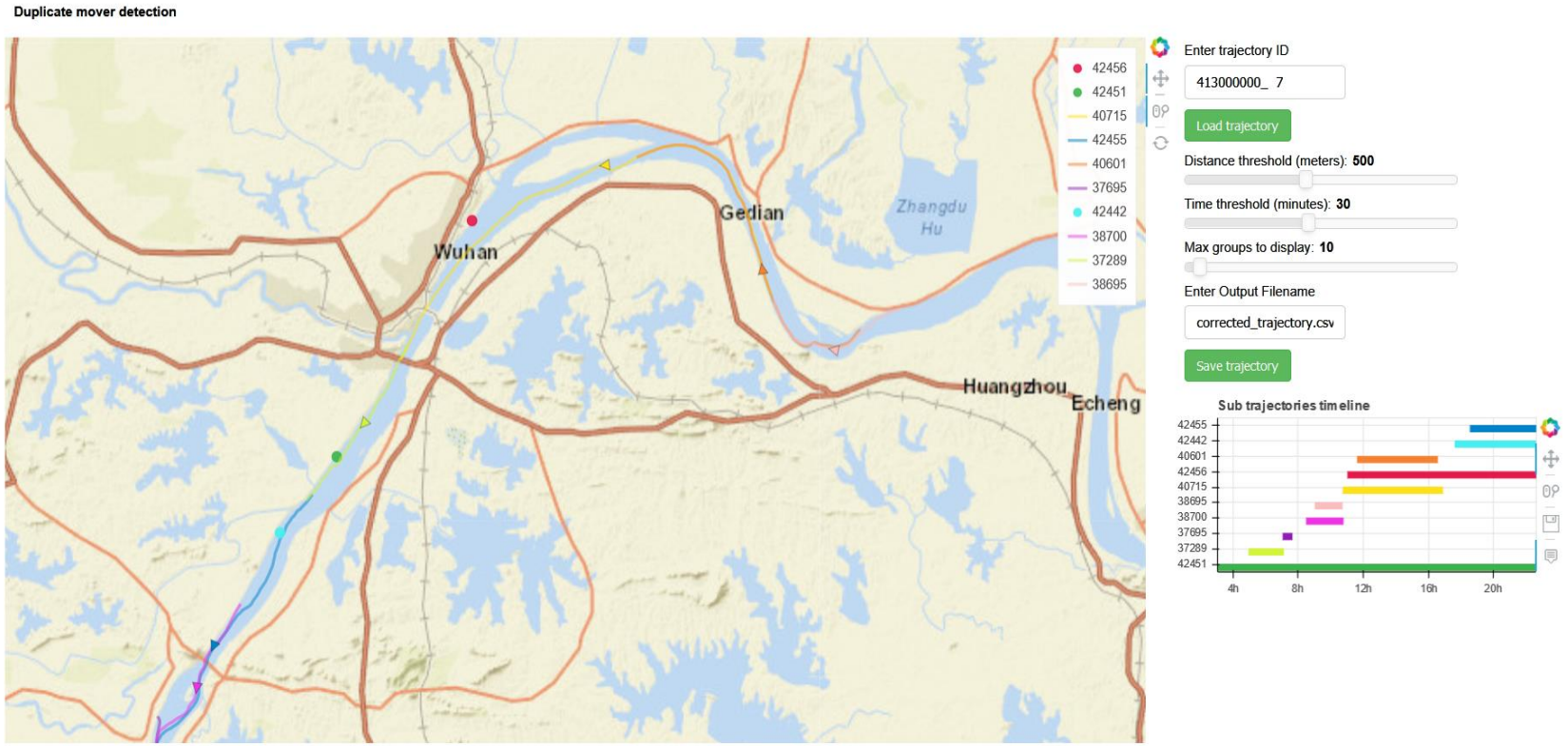
Load data

#	Trajectory ID	count	mean	std	max
77	413997052_2	6851	15.77463824179...	5.79255886312529	54.830978399229...
78	413997052_1	7019	12.47667595539...	4.888244435660...	49.299344056359...
79	413828559_1	11778	7.321046253402...	3.334465127349...	48.236414895032...
80	413828452_1	5056	11.220663509005...	4.475594857010...	40.531890715428...
81	413996252_11	5507	11.367471721116...	3.610044788961...	37.557550185284...
82	413828452_3	5897	10.79339122742...	3.647235898135...	35.015095446950...
83	413828452_9	4964	12.16923917836...	4.3108186802711...	33.367876587249...
84	413793039_1	4953	11.614698773345...	3.438333785050...	31.877528484830...
85	413996599_1	13259	6.782474370249...	2.758608133986...	31.794715778066...
86	413829321_11	9251	8.57270872216379	3.106379715043...	31.771537646985...
87	413997591_3	6187	7.801119165230417	2.9208981011887...	31.770593320017...
88	413829321_1	12247	7.792458419813...	3.413468978504...	31.656785500826...
89	413793856_1	6249	9.464290203394...	3.238150208984...	30.497986730065...
90	413793363_2	5240	8.701002931969...	2.982809770635...	29.091818701333...
91	413996734_9	5162	10.59292485093...	3.228415213004...	29.059818849142...



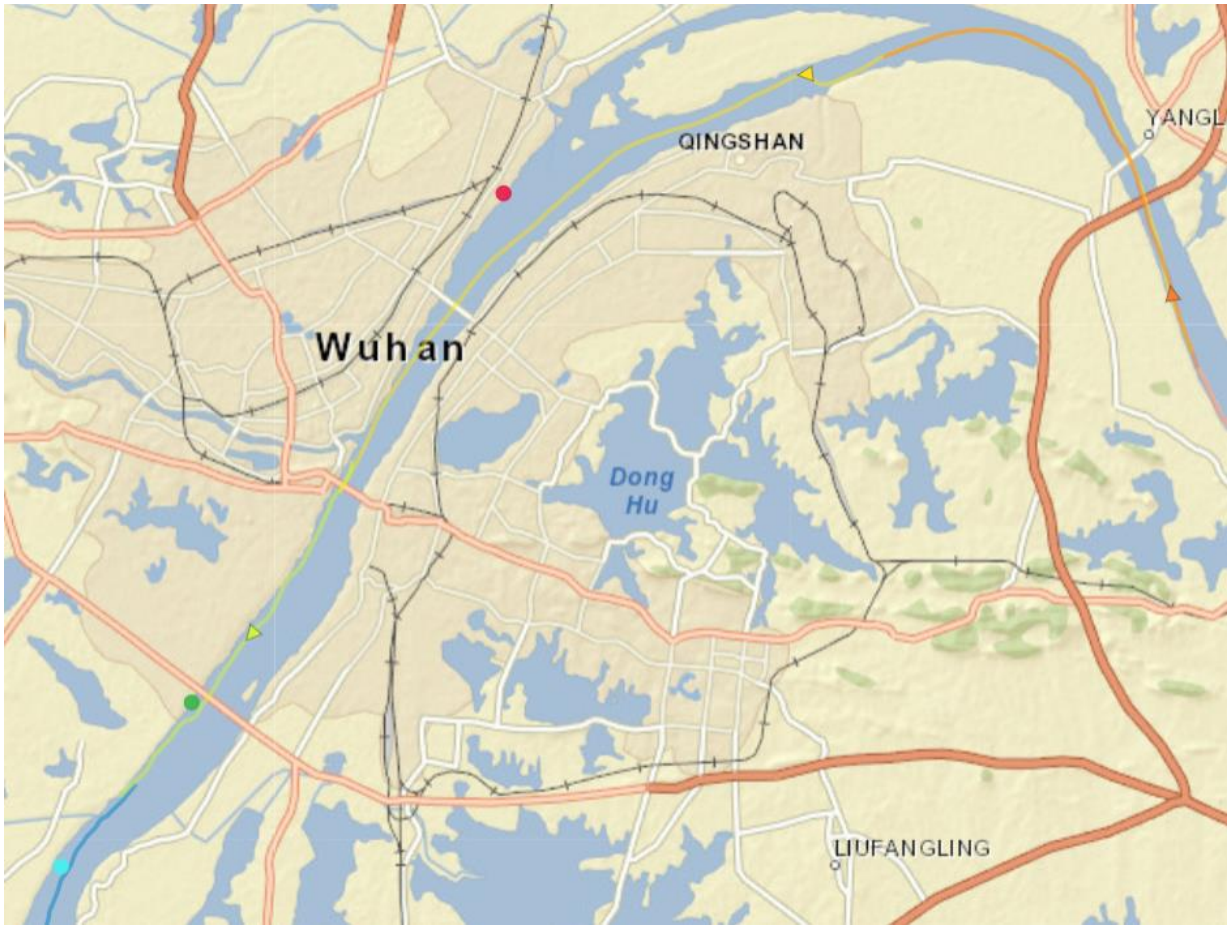
Methodology (visual analytic component)

Duplicate mover detection interface



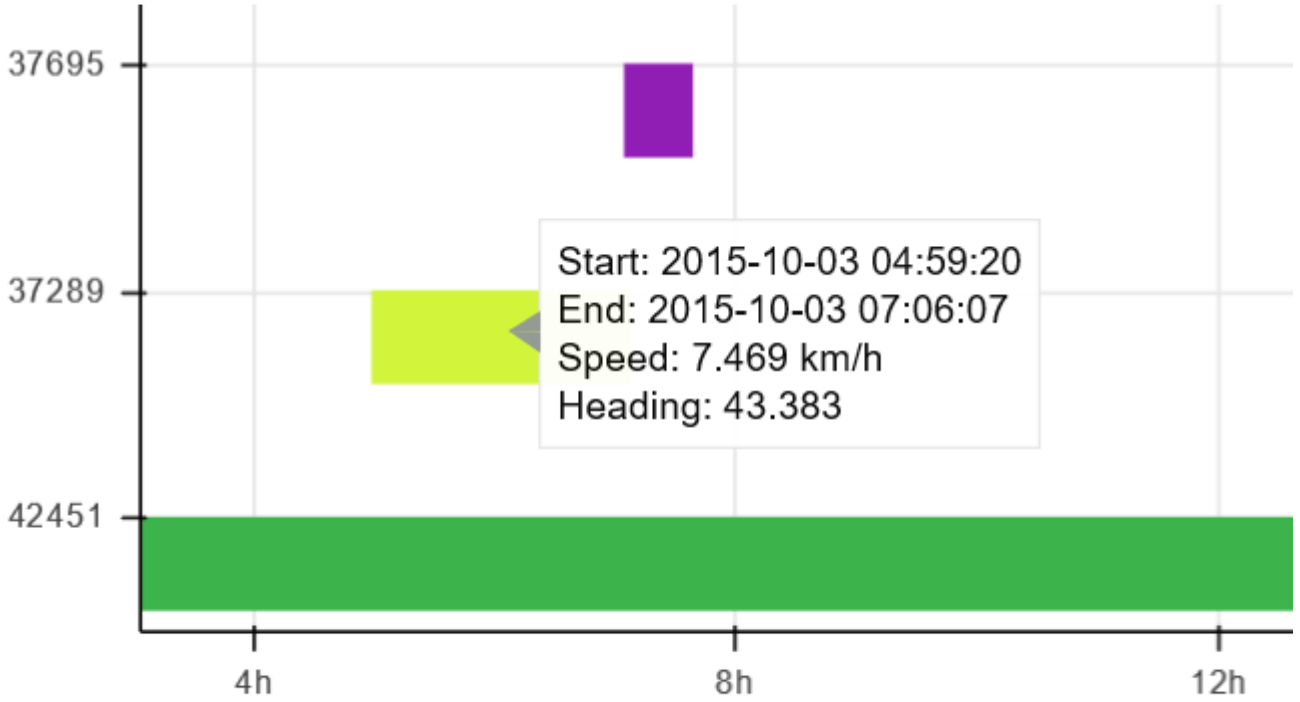
Methodology (visual analytic component)

Duplicate mover detection interface
(slippy map with glyphs for movers and stop points)



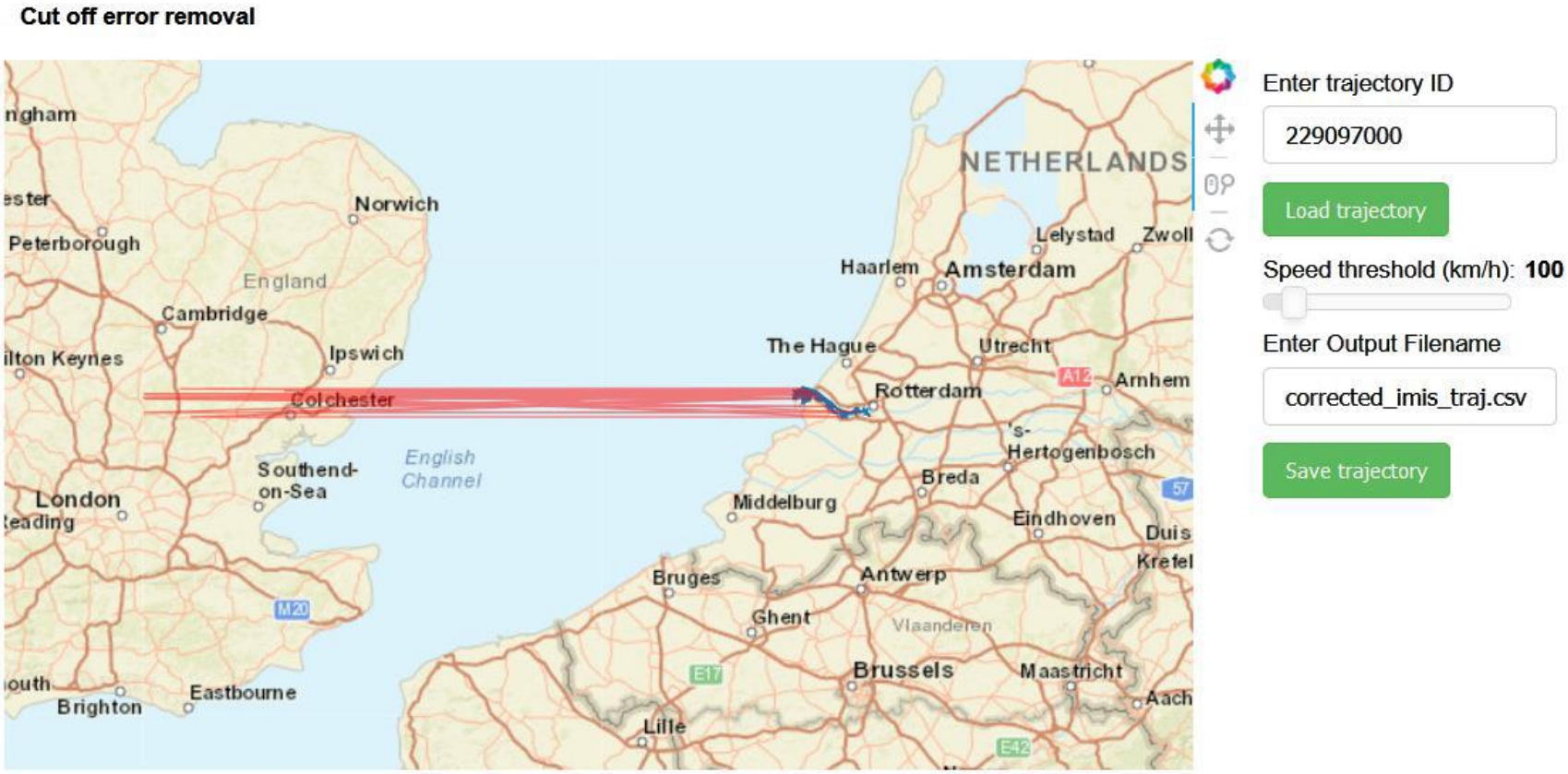
Methodology (visual analytic component)

Duplicate mover detection interface (timeline view)



Methodology (visual analytic component)

Cut-off error removal interface



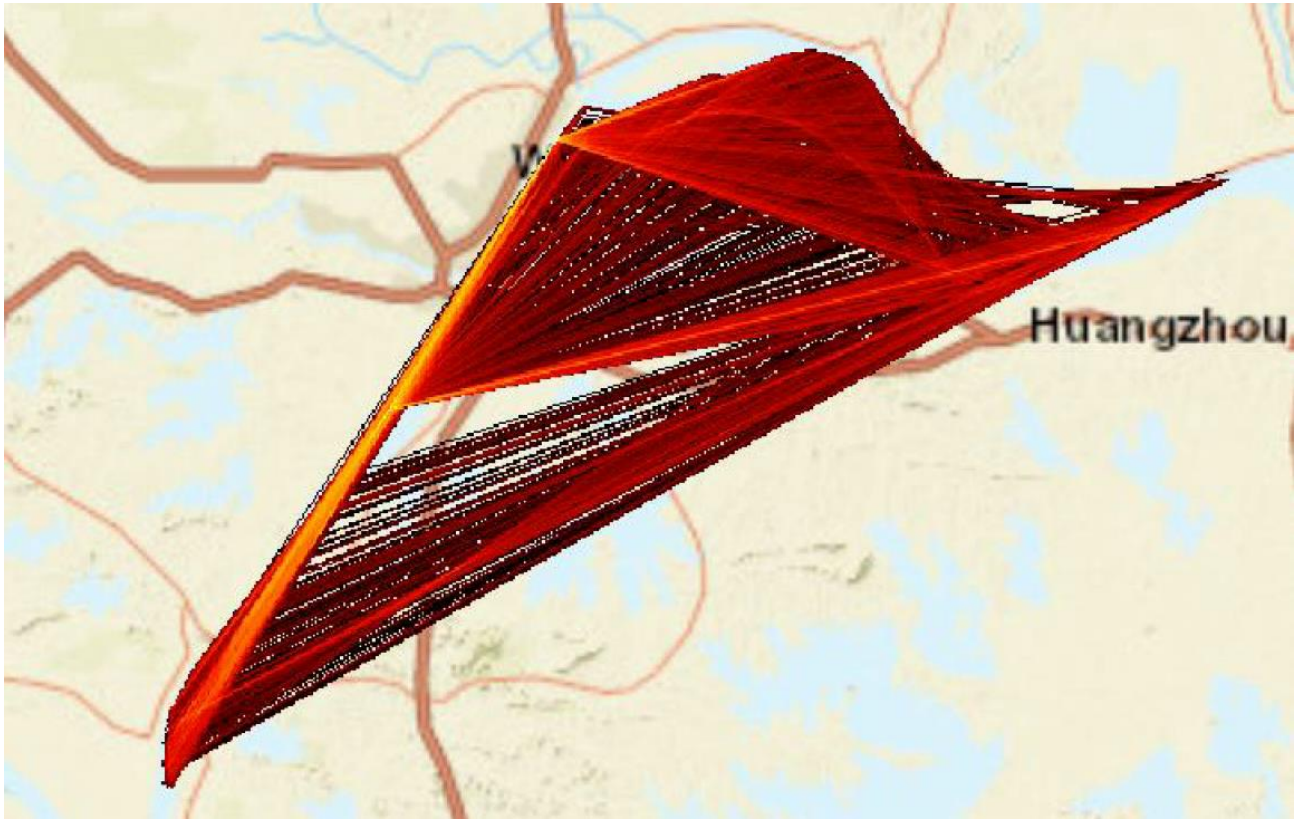
Evaluation

2 maritime datasets analyzed

	Wuhan Dataset	European Dataset
Region	China	Europe
Total number of records	~13 million	~61 million
Total number of trajectories	45,318	118,000
Duration	1 month (2015-10-01 to 2015-10-31)	1 month (2016-01-01 to 2016-01-31)
Original format	CSV	CSV
File size	6 GB	3.7 GB

Evaluation

Visualization of an anomalous trajectory (Wuhan dataset)



Evaluation

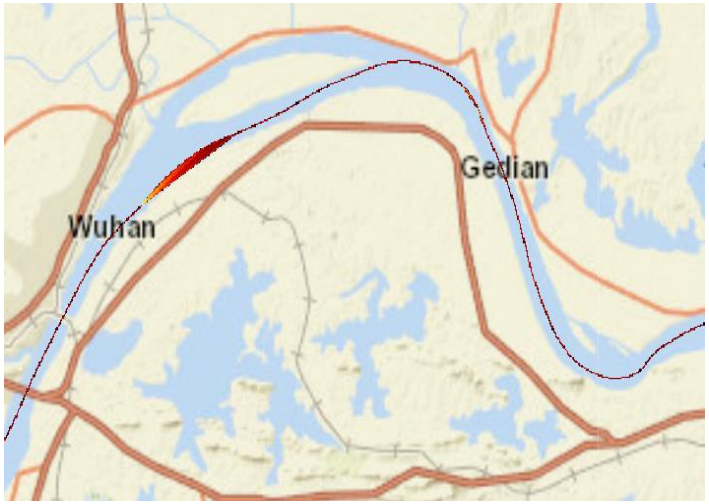
Sample mover identity correction result (Wuhan dataset)

Original Trajectory

Corrected

Evaluation

Sample Mover identity correction result (Wuhan dataset)

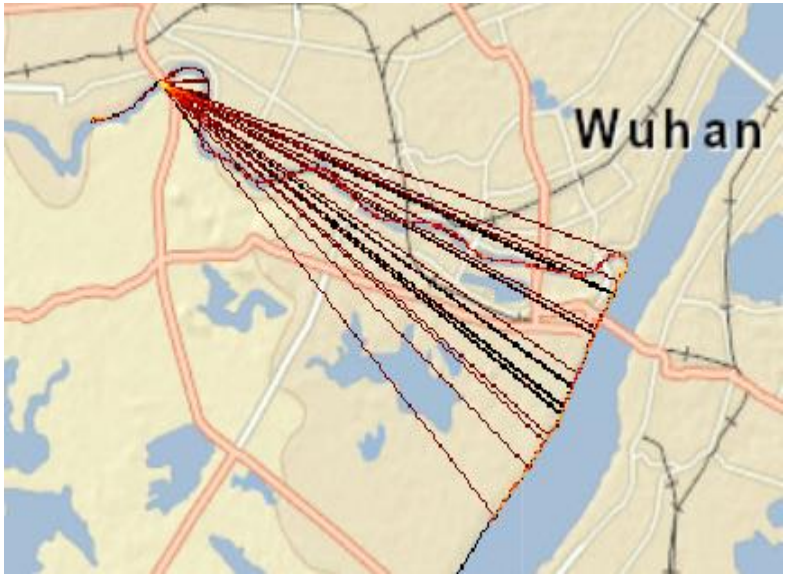


Original Trajectory

Corrected

Evaluation

Sample Mover identity correction result (Wuhan dataset)

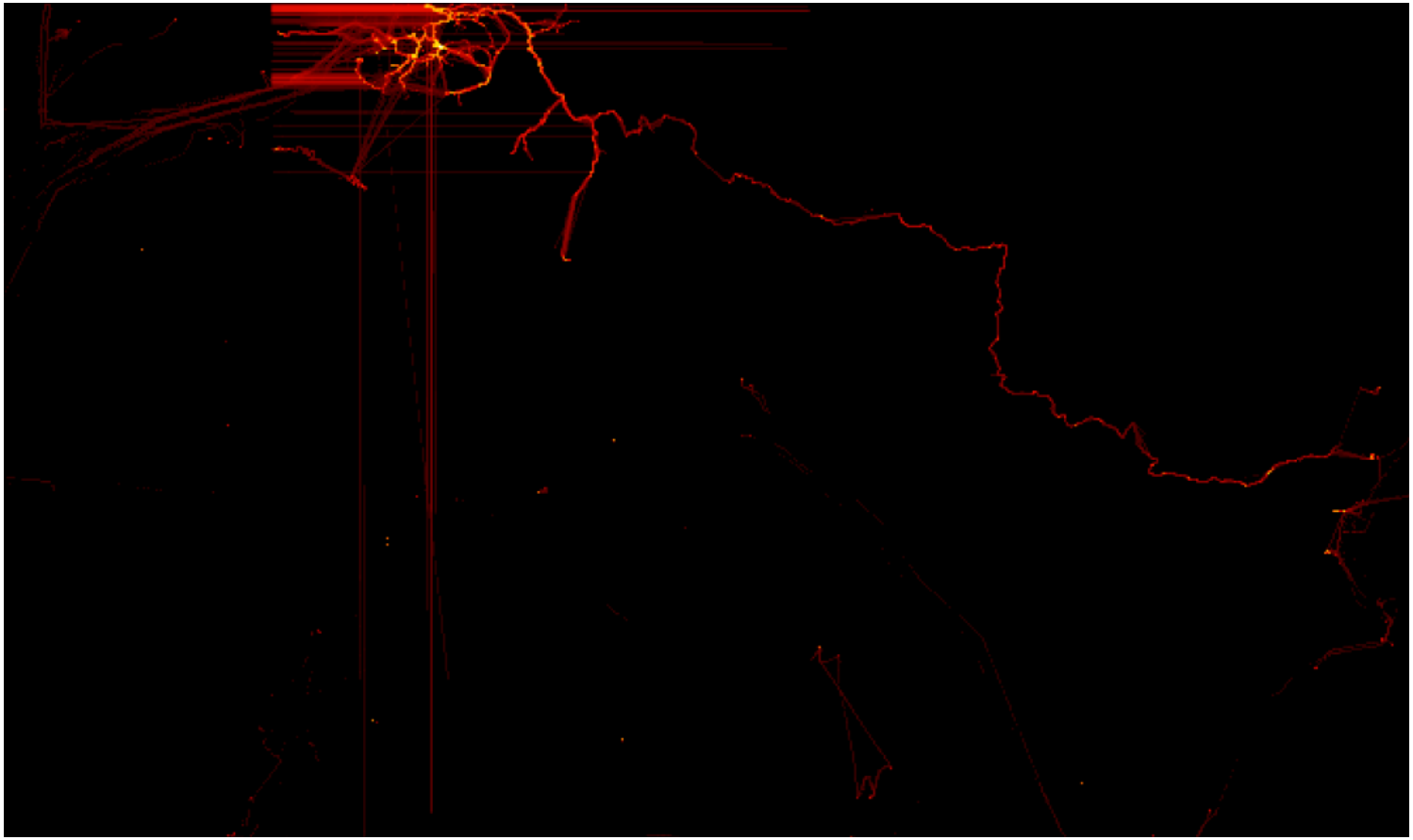


Original Trajectory

Corrected

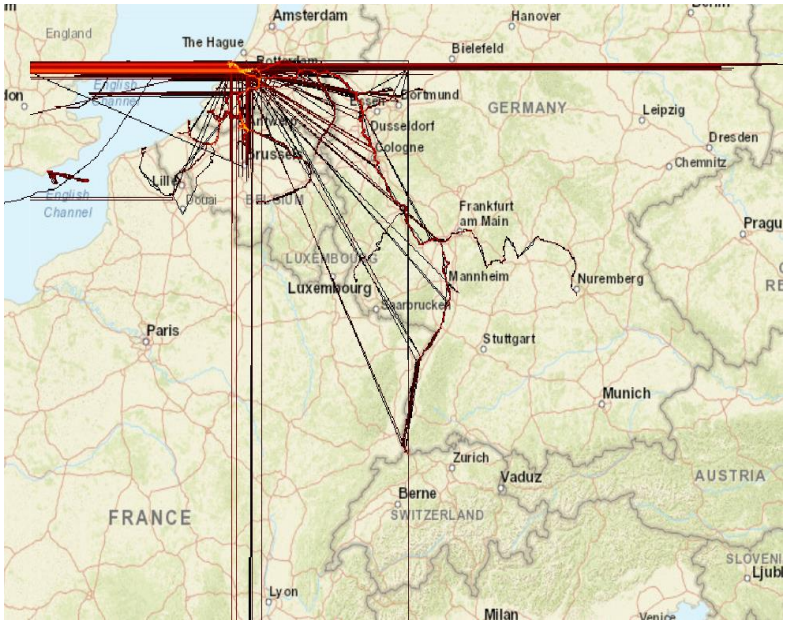
Evaluation

Visualization of European dataset

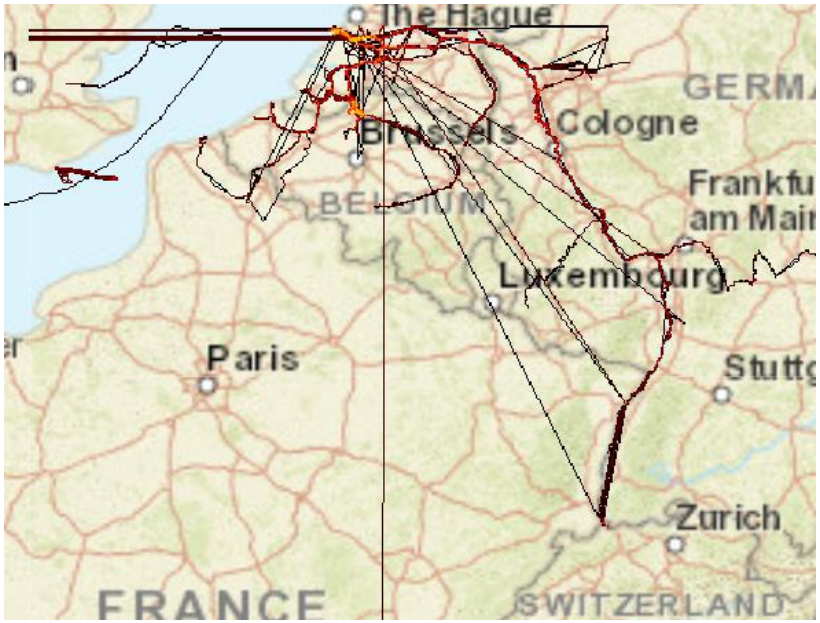


Evaluation

Cut-off error reduction (European dataset)



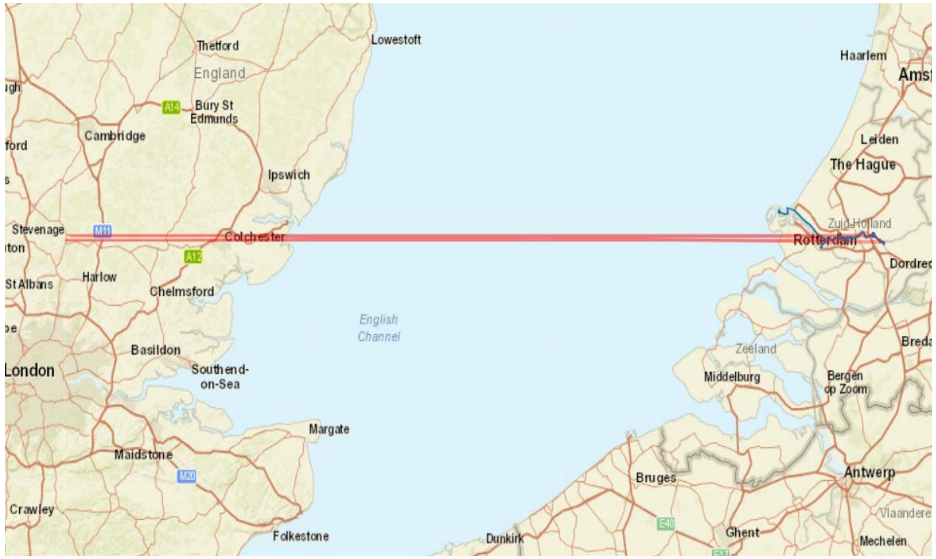
Original Trajectory



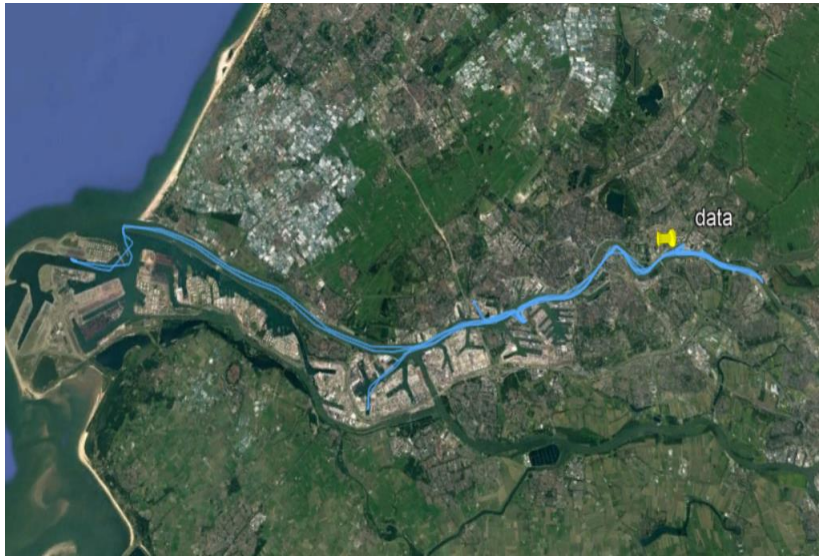
Reduced number of cut-off errors

Evaluation

Cut-off error reduction (European dataset)



Original Trajectory



Reduced number of cut-off errors

Results

Mover ID correction results from a sample trajectory (Wuhan dataset)

	Before	After
Distance	125880 km	300 km
Mean speed	6625 km/h	9.3 km/h
Visual	Many criss-crossing segments	None
Performance	6559 points processed in 1.87 s	

Cut-off error correction results from a sample trajectory (European dataset)

	Before	After
Number of high-speed segments removed	131	
Mean speed	193 km/h	3.2 km/h
Visual	Many long distance segments	Reduced
Performance	8965 points processed in 7.12 ms.	

Results

Wuhan dataset batch mode correction results (20 anomalous trajectories)

Original Trajectories

Corrected trajectories

Conclusions

- Method to detect groups of movers from an anomalous trajectory was developed
- Visual analytic method to alter algorithm parameters was developed
- Methods were applied to two maritime datasets

Future work

- Additional modes of transport
- Parallel detection of other multiple classes of errors

References

- Andrienko, Gennady; Andrienko, Natalia; Fuchs, Georg (2016): Understanding movement data quality. In *Http://Dx.Doi.Org/10.1080/17489725.2016.1169322* 10 (1), pp. 31–46. DOI: 10.1080/17489725.2016.1169322.
- Douglas, David H.; Peucker, Thomas K. (1973): Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. In Cartographica: The International Journal for Geographic Information and Geovisualization 10 (2), pp. 112–122.
- Doulkeridis, Christos; Vlachou, Akrivi; Santipantakis, Giorgos; Glenis, Apostolos; Vouros, George (2016): Big Data Analytics for Time Critical Mobility Forecasting. Available online at http://ai-group.ds.unipi.gr/datacron/node/40.
- Group, H. D.F. (2014): Hierarchical data format, version 5.
- Guttman, Antonin (1984): R-trees. A dynamic index structure for spatial searching: ACM (2).
- Keim, Daniel A.; Bak, Peter; Bertini, Enrico; Oelke, Daniela; Spretke, David; Ziegler, Hartmut (Eds.) (2010a): Advanced visual analytics interfaces. Proceedings of the International Conference on Advanced Visual Interfaces: ACM. Available online at http://dl.acm.org/citation.cfm?id=1842995.

References

- Patroumpas, Kostas; Alevizos, Elias; Artikis, Alexander; Vodas, Marios; Pelekis, Nikos; Theodoridis, Yannis (2016): Online event recognition from moving vessel trajectories. In Geoinformatica. DOI: 10.1007/s10707-016-0266-x.
- Vouros, George (2017): dataACRON, Big Data Analytics for Time Critical Mobility Forecasting, H2020. In impact 2017 (5), pp. 75–77. DOI: 10.21820/23987073.2017.5.75.
- Ward, Jonathan Stuart; Barker, Adam (2013): Undefined by data. A survey of big data definitions. In *arXiv preprint arXiv:1309.5821*.
- Zheng, Y. U. (2015): Trajectory Data Mining. An Overview 6 (3), pp. 1–41.

Thank You