Technische Universität München
Department of Civil, Geo and Environmental Engineering
Chair of Cartography
Prof. Dr.-Ing. Liqiu Meng

Detection and correction of mover identity problems in movement datasets

Nevil Jose Kodiyan

Master's Thesis

Duration:	01.05.2017 - 22.12.2017
Study Course:	Cartography M.Sc.
Supervisors:	Dr.-Ing. Mathias Jahnke Barend Köbben Dr. Georg Fuchs
Cooperation:	Fraunhofer Institute, Schloss Birlinghoven

2017

CHAIR OF CARTOGRAPHY

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master's Thesis

Detection and correction of mover identity problems in movement data sets

Author

Nevil Jose Kodiyan

Supervisors

Dr.-Ing. Mathias Jahnke

Barend Köbben

Dr. Georg Fuchs

Submission Date: 22.12.2017

I hereby confirm that this Master thesis is my own work. Direct and indirect references to other sources have been noted.

Sankt Augustin, 22.12.2017 Nevil Kodiyan

Acknowledgements

I would like to thank my supervisors Dr. Georg Fuchs and Dr.-Ing. Mathias Jahnke for giving me an opportunity to work on this topic. The support rendered by them is truly appreciated. I thank Dr. Gennady Andrienko for providing access to the maritime datasets and Mr. David Knodt for the timely help rendered.

This Master’s Programme has been a positive and enriching experience for me. I thank my colleagues for making the journey interesting and rewarding. I owe gratitude to Mrs. Julian Cron for her excellent supervision of the entire course.

Finally, I thank my family for their patience and support throughout the entire programme.

Abstract

In the last decade, movements like open-source, open data have led to a proliferation in the number of publicly available datasets. Among the variety of datasets now available (publicly as well as privately to researchers) are movement datasets. Movement datasets track the motion of several moving entities through time. The moving entities can be as diverse as ships, aircraft, pedestrians and animals. Due to technical limitations, noise or errors the movement datasets may exhibit a loss in quality. One class of errors is seen in movement data sets when several movers inadvertently use the same mover identifier. The resulting trajectories appear mixed up and incomprehensible unless it is corrected.

The aim of this thesis is to develop an algorithm to detect mover identity errors, with a focus on maritime datasets. A strategy based on grouping anomalous trajectory points by spatial and temporal distance is described. Secondly, a visual-analytic strategy to parameterize the algorithm with expert human guidance is also described. The technique is applied to two maritime large trajectory datasets from the European and Wuhan regions, respectively.

Results indicate that the described technique can identify several moving entities within anomalous trajectories. As a future improvement, the algorithm could be adapted to handle more movement modalities.

Keywords: Visual Analytics, Movement Dataset, AIS, Trajectory, Maritime

Contents

List of Figures

List of Tables

Abbreviations

AIS (Automated Identification System)

GPS (Global Positioning System)

IMO (International Maritime Organization)

Maritime Mobile Service Identity (MMSI)

SOLAS (International Convention for the Safety of Life at Sea)

VA (Visual analytics)

1. INTRODUCTION

Big Data (John Walker 2014) refers to datasets which are extremely large and fast changing. Government and private institutions work with large datasets, in the hope of deriving value from it (Kim et al. 2014) . In recent years, open government initiatives have also resulted in more datasets being published online (Attard et al. 2015).

The rise of big data sets has necessitated a corresponding drive visualize, analyze and make actionable decisions using them. The allied fields of data mining, information visualization and visual analytics attempt to solve the problem of synthesizing information and deriving insights from large data sets. The aim is "detect the expected and discover the unexpected" (Keim et al. 2010b).

Movement datasets track the movement of one or more entities through time. The movement history of the entity is derived either through self-positioning or by tracking. For a personal user, self-position is needed to aid in the navigation from point A to B. Thus, nowadays almost every mobile device is fitted with an in-built Global Positioning System (GPS) chipset (Vaughan-Nichols 2009). Taxi companies equip their vehicles with GPS trackers to dispatch the closet taxi to a waiting customer (Liao 2003). Public transport is also being tracked so that commuters can obtain the most accurate arrival and departure times. Heath conscious people track their own movements through wearable technologies to gauge the calories expended. These are all rich and diverse sources of movement data. Researchers track animal movements to learn more about their breeding habits and migration patterns. The maritime and aviation sectors also generate movement datasets through the tracking of vessels and airplanes. Every form of locomotion can potentially be a source of movement data.

Errors and uncertainties in the positioning and tracking technologies impact the quality of the movement dataset. For example, GPS positioning can be degraded in urban areas and can be completely lost inside building and tunnels. Positioning equipment can malfunction or be inadvertently switched off. It is also possible that an adversary may spoof their own position to conceal illegal activities.

1.1 Problem Statement

A movement dataset contains the position histories of various moving entities with time. Each record in a movement dataset contains a timestamp, a mover identifier, information about its position and optionally additional contextual information. The movement pattern of a single entity within a movement dataset is called a trajectory. Movement datasets usually use a unique identifier for each mover. When accuracy errors occur in the mover identifier, it results incorrect zig-zag trajectories, which Andrienko et al. (Andrienko et al. 2016) refer to as *mover identity errors*. Inaccurate records can lead to situations where the same identity is used by multiple movers or multiple identifiers are used by the same mover. When such trajectories are visualized, movers appear to move at unrealistically high speeds or distances.

1.2 Background

datACRON is a research project funded by the European Union's Horizon 2020 Programme. It's aim is introduce novel methods for threat and abnormal activity detection in very large fleets of moving entities spread across large geographical areas (Vouros 2017). One of the sub-objectives of the datAcron project is to create a framework capable of detecting quality issues in movement datasets (Doulkeridis et al. 2016) originating from the maritime and aviation domains.

Since the 2002 IMO SOLAS Agreement (Chang 2004) most ships are required to use AIS (Automated Identification System) transceivers. The data relayed by AIS messages include its unique identifier, position, course and speed. The intention is to allow monitoring of maritime traffic and to take measures to prevent collisions and optimize traffic flow. Since the AIS message contains an identifier for each ship, motion trajectories of vessel can be built and analyzed.

AIS datasets which track the positions of thousands of ships can quickly swell to several GBs in size in a matter of days. Therefore, there is need for management, interpretation and analysis of such datasets both offline and online. One of the first steps in working with a new dataset is to remove obvious data quality related artifacts.

1.3 Research Objective

The goal of this Master thesis is to detect and correct mover identity problems within movement datasets, and to present the results visually. There are several existing algorithms dealing with anomaly detection, pattern mining and clustering of trajectories (Zheng 2015). However, so far there is no algorithm for specifically dealing with mover identity errors. There is also a need to develop a corresponding visual analytic methodology for addressing this problem.

Some of the cases that can be considered are:

1. Movers moving in the same direction.
2. Movers moving in opposing directions.
3. One mover is stationary, while the other is moving.
4. Duplicate or repeated rows in dataset.

This initial study will focus on developing a solution configured for the maritime domain. More complex combinations involving multiple movers are also possible. Since human beings are proficient at detecting patterns, the detection algorithm will also be coupled with an interactive exploratory approach.

1.4 Target Audience

The expected users of this developed algorithm and interface are operators working on specialized movement datasets. For example, the interface could be used in the soon after data acquisition to visualize the quality problems and determine optimal parameters for their correction.

1.5 Outline of thesis

The thesis is divided as follows:

- Chapter 1 introduces the topic and the context in which the problem of duplicate IDs is seen.
- Chapter 2 discusses the movement data quality, with an overview of related work and argues for a visual-analytic approach to solve the problem.
- Chapter 3 explains the methodology used; the specifics of the algorithm used for anomaly detection and a description of the visual-analytic method methodology used to address the problem.
- Chapter 4 describes the results of evaluating the developed algorithm against maritime dataset.

- Chapter 5 contains a description of the results achieved.
- Chapter 6 draws the final conclusions and gives recommendations for future work.

1.6 Working steps

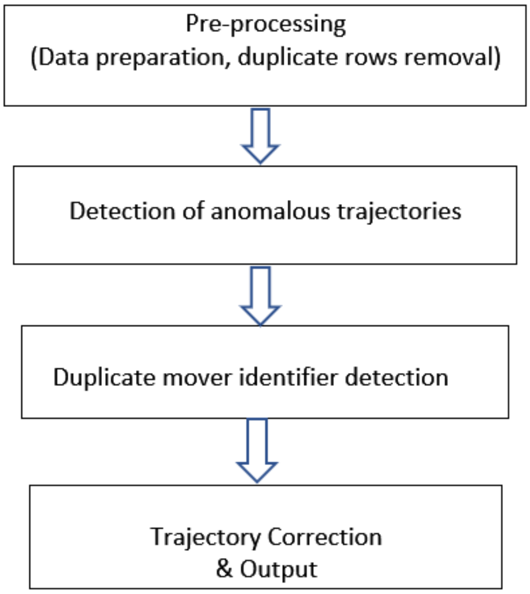

Pre-processing
(Data preparation, duplicate rows removal)
Detection of anomalous trajectories
Duplicate mover identifier detection
Trajectory Correction
& Output

Figure 1. Overall Stages of processing

Fig. 1 shows a bird's eye view of the expected working steps. It begins with pre-processing stage which involves removal of the simplest dataset errors and data conversion. In the second stage, the most likely anomalous trajectories within a dataset are identified. Next, duplicate mover identifiers are detected in the anomalous trajectories. Finally, the corrected trajectories are output to a file.

2 MOVEMENT DATA QUALITY

Movement datasets can be obtained from a variety of sources such as animal motion (Kranstauber et al. 2011), pedestrians movement (Alvarez, Leeson 2015) or urban transport. These datasets are typically delivered as collection of trajectories of movers. Since each dataset is impacted by the limitations of the underlying measurement methodology, it is important to understand the quality of the underlying data. Three types of movement data quality problems were delineated by Andrienko et al. (Andrienko et al. 2016):

- Missing data problem (which are due to spatial or temporal gaps in the position record)
- Accuracy problem (errors due to temporal, spatial, identifier or attribute level inaccuracies)
- Precision deficiency (errors due to limitations of the positioning system)

2.1 Mover identity problem

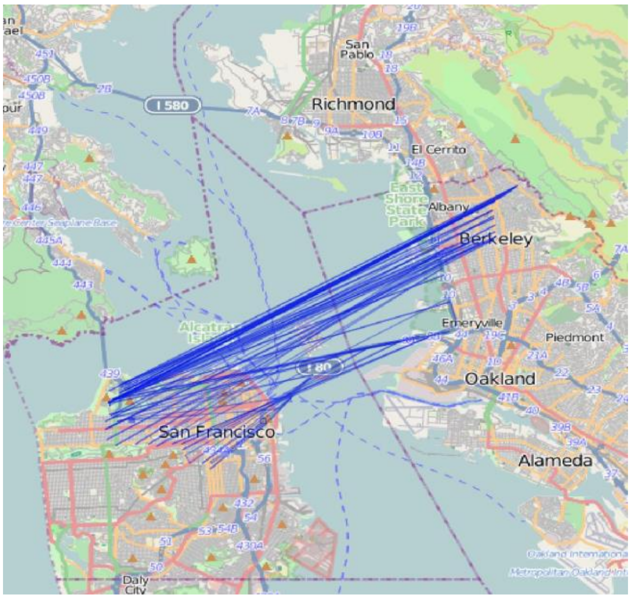


Figure 2. Example of mover identifier error where two trajectories have been mixed into one (Andrienko et al. 2016)

A mover identifier error is a movement data accuracy problem which occurs when one or more movers share the same identifier (Andrienko et al. 2016). As a result, their trajectories are joined and appear mixed up. As seen the in the Fig. 2 above, this results in trajectories with zig zag patterns. In Fig. 2 the most likely scenario is that there are two separate movers moving at the same time in San Francisco and Berkeley, respectively.

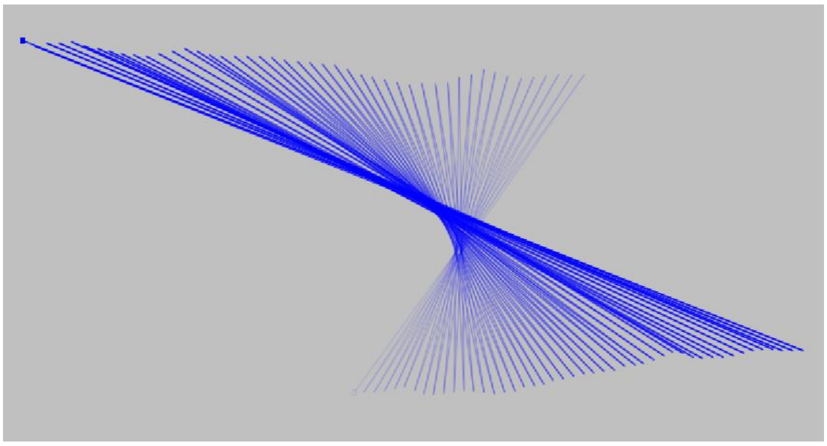


Figure 3. Simulation of two movers moving in opposite directions

Fig. 3 shows the characteristic trajectory pattern when two movers move parallel to each other but in opposite directions.

2.2 Related work

There is an extensive literature dealing with trajectory pre-processing, data mining, indexing, clustering and anomaly detection. Zheng (Zheng 2015) provides a helpful survey of the key research areas in trajectory data mining.

The anomalous trajectory detection literature is mainly concerned with identifying anomalous trajectories within a group of trajectories. Statistical, pattern mining and heuristic approaches are common. The anomaly detection method used by Laxhammar (Laxhammar 2008) is based on unsupervised clustering of sea traffic using a feature model. Two different feature models were used – one based of momentary vessel velocities and another which incorporated momentary vessel position as well. Ristic et al. (Ristic et al. 2008), uses historic vessel trajectory data to build patterns of normal motion. A motion pattern is defined by a location and velocity vector. Supplementary attributes like vessel type, season may be added. Anomaly detection is done using a Kernel Density Estimator which is applied to incoming AIS data. Deng et al. (Deng et al. 2013), divides the region in consideration into a grid of 0.1° x 0.1°. Vessel positions are discretized into a sequence of grid cell movements. Similarly, vessel speed is also discretized. The translated historical vessel status messages are used to develop a Markov Model with probabilities of transition between states. An anomaly is said to have occurred when a low probability transition is detected on the live data. Zhang et al. (Zhang et al. 2011), detect anomalous taxi trajectories by assuming that they are "few and different". In practice this is done by dividing the space into grid cells and then mapping the sequence of cells traversed by a trajectory. Isolation Forest technique intrinsically partitions the trajectories based on the “few and different” property. Anomalous trajectories tend to have shorter path lengths in the tree. Shahir et al. (Shahir et al. 2015), represents vessel interaction patterns as Hidden Markov Models and classifies such patterns using Support Vector Machines. Anomaly detection is done by using human assistance by presenting the initially detected scenario as well as contextual information. Patroumpas et al. (Patroumpas et al. 2016) detects complex motion patterns by using heuristics based on the instantaneous and mean motion of a vessel. From the stream of incoming AIS messages, “critical points” (stop, turn, slow motion) along a trajectory are determined. Events are classified as instantaneous or long-lasting. Instantaneous events like Pause, Speed Change are detected by a simple thresholding operation on the instantaneous velocity vector. Long-lasting events such as Gap, Slow Motion are detected by buffering instantaneous events within a sliding window.

Most of the existing anomaly detection techniques assumes that anomalous trajectory deviates from the norm but is otherwise a faithful representation of the vessel motion. To the best of the author’s knowledge, no technique was found which can be applied to trajectories that exhibit mover identity problems.

2.3 Related data quality issues

Data quality errors rarely occur in isolation. Usually a combination of errors manifests in the same dataset. Hence, it is instructive to understand the closely related classes of data quality errors.

2.3.1 Cut-off errors

Cut off errors are a missing data problem that occurs when a cropped subset of the movement dataset is visualized (Andrienko et al. 2016). Trajectories which were continuous are broken up as a side effect of data subsetting. Visualization of such a dataset reveals several large vertical or horizontal jumps over unlikely distances, as shown in Fig. 4.

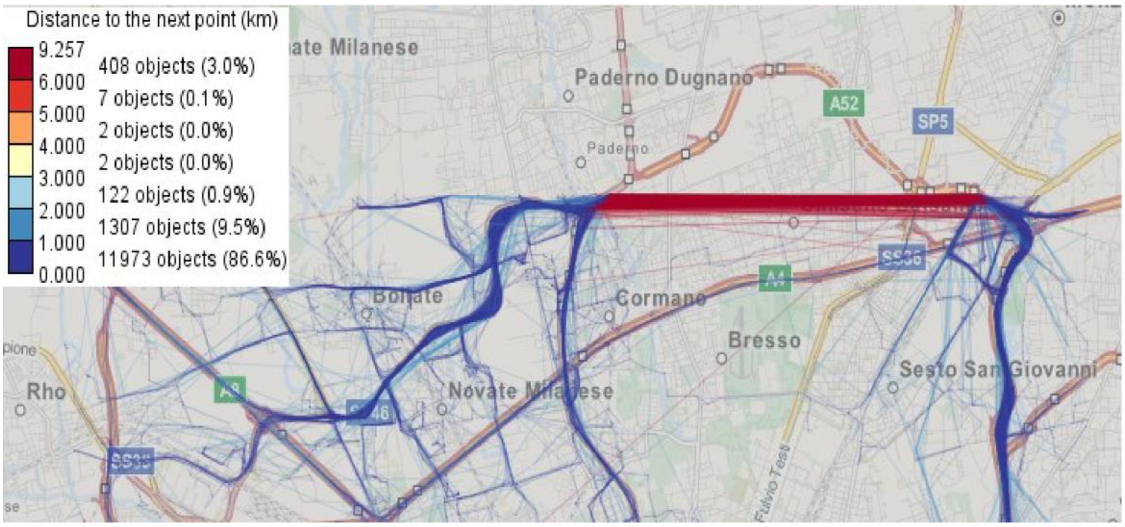


Figure 4. Example of cut-off error (Andrienko et al. 2016)

2.3.2 Jitter

Jitter occurs due to inaccuracies in the position estimation of recording devices. It is more evident at a stop point or a when a mover is moving very slowly. There is a need to distinguish jitter from the mover identity problem.

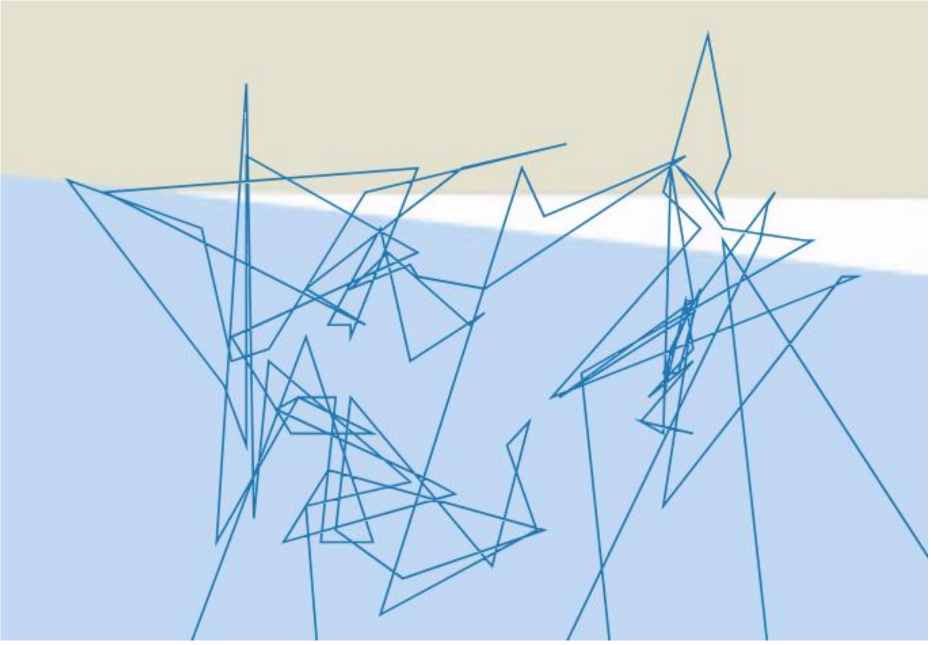


Figure 5. Example of jitter at a stop point

Jitter movements are characterized short displacements and high angular changes.

2.4 Need for visual analytics based approach

Due the potentially large amount of trajectory data being visualized automated techniques are needed for exploring the data (Andrienko, Andrienko 2008). At the same time, contentious decisions based on the data are best left to humans who are have better visual perception skills and domain expertise. Visual analytic (VA) techniques aim to assist the decision making through relevant interactive visualization techniques.

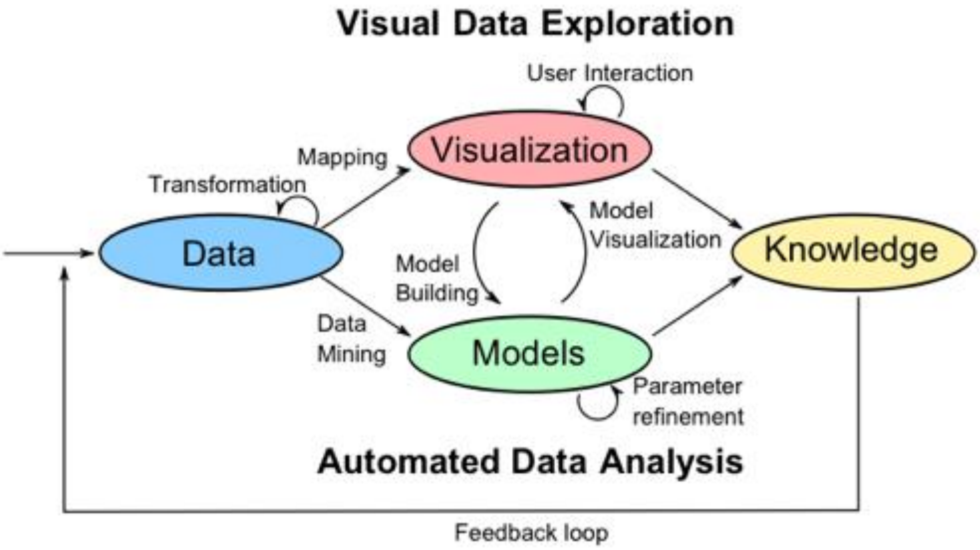


Figure 6. Visual Analytics Process Steps (Keim et al. 2010a)

Fig. 6 shows the steps of the visual analytics process (Keim et al. 2010a). The process begins with preprocessing of the data to extract the data relevant to the analysis. The next step can be either automatic (model driven) or user driven (visualization). Both the procedures complement each other, and help to refine the model parameters. The result is an increase in knowledge or a new insight into the dataset.

In the case of mover identity detection, a visual analytic strategy is needed to ensure that the selected algorithm parameters can adequately remove the error class. Secondly, it is also useful to get a visual continuous feedback in cases where it is not easy to differentiate between the different error classes.

3 METHODOLOGY

This chapter describes the details of the algorithm to detect potentially anomalous trajectories and the duplicate mover within them. As mentioned Chapter 1, the overall problem can be logically divided into pre-processing, detection followed by correction stages.

3.1 Pre-processing

A typical AIS message contains information such as the timestamp, latitude, longitude, Maritime Mobile Service Identity number (MMSI), latitude, longitude, heading and bearing among other fields (Silveira et al. 2013). The MMSI uniquely identifies a vessel. In practice, however it is possible that MMSI could be incorrectly configured, leading to mixed up trajectories. Due to errors in the AIS data acquisition process, AIS messages can be also duplicated.

As a part of the data pre-processing phase, consecutive duplicate AIS records are discarded. Latitude and longitude fields are converted to Web Mercator projection (EPSG:3857), to facilitate ease of plotting in subsequent stages. The data originally delivered in .csv format is pre-processed and stored in HDF format (Group 2014). This format has several advantages. Firstly, the original uncompressed .csv file can be compressed, leading to lower on-disk file sizes. Secondly, HDF supports column indexing. The MMSI column is indexed, which allows quick loading of a desired trajectory from the file. Finally, the HDF file can be read in chunks, which allows multi-gigabyte files to be processed without loading the entire file in-core.

3.2 Heuristic for detection of anomalous trajectories

During offline processing, a dataset containing thousands a of trajectories may need to be analyzed. A quick heuristic to detect which of these trajectories are anomalous is needed. Trajectories containing duplicate IDs are characterized by many high velocity jumps. This property is therefore used to detect anomalous trajectories. The figure below shows the speed histogram of an anomalous trajectory. Typically, marine vehicles do not exceed 100 km/h. A simple threshold value for mean speed is used to filter possible anomalous trajectories.

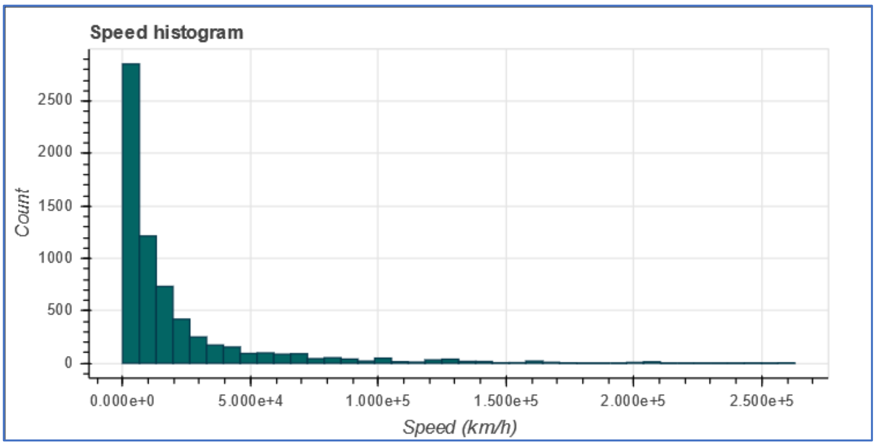


Figure 7. Speed histogram of an anomalous trajectory

3.3 Duplicate mover ID detection algorithm

A manual grouping technique for trajectory points is proposed. The algorithm works as follows: Each trajectory point is processed sequentially, and allocated to the group that best fits it based on the properties of the group and the trajectory point under consideration.

Table 1. Group properties

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

The first point of the anomalous trajectory is allocated to a new group, since there are no other groups. The next trajectory point will be allocated to the closest group which satisfies distance and time thresholds. The last inserted point in a group is used as a reference point for distance and time difference calculations. If none of the existing groups would satisfy the distance and time thresholds, a new group is created, and the point is inserted into it. After all the points are classified, groups that exhibit high mean angular changes and low total distance covered are considered as stop points. A group direction property could also theoretically be used for finding the best group, but in practice this was not found necessary when working with the available maritime datasets.

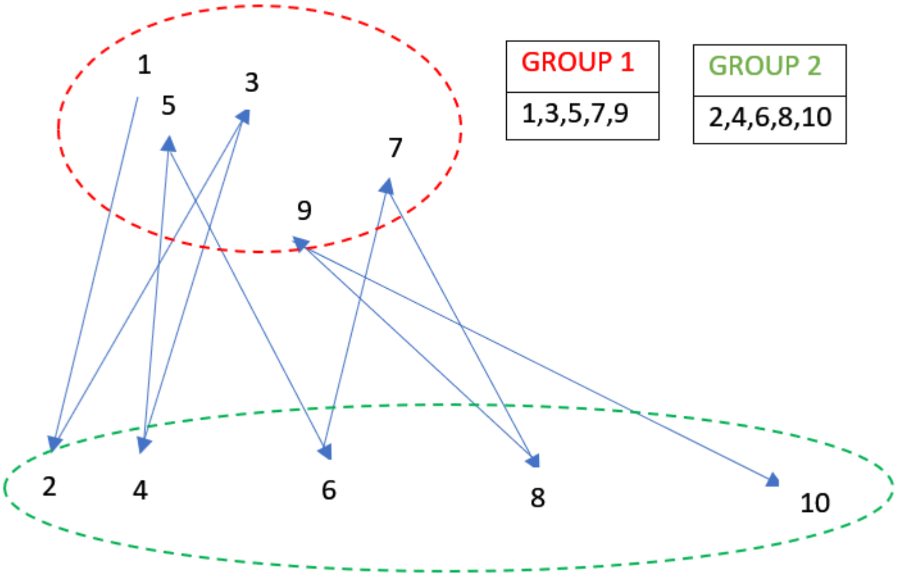


Figure 8. Trajectory points are placed into the closest group which satisfies the time and distance thresholds

A trajectory can have thousands of points, and computing the distance to each group can quickly become non-trivial. An R-Tree spatial index is therefore used to optimize the time complexity of the search for the closest matching group. The R-Tree (Guttman 1984), is a height-balanced hierarchical index data structure and it is typically used to quickly find objects falling within a rectangular query region. It is also possible to query the data structure to find the closest object from any given position. The data structure is used to keep track of the location of the last inserted trajectory point in each group. When a new trajectory point is to be categorized, a nearest neighbor search is initiated on the R-Tree to find the closest group. If a group is found that satisfies the spatial distance and time thresholds, the trajectory point is inserted into that group. The R-Tree leaf object for that group is then updated.

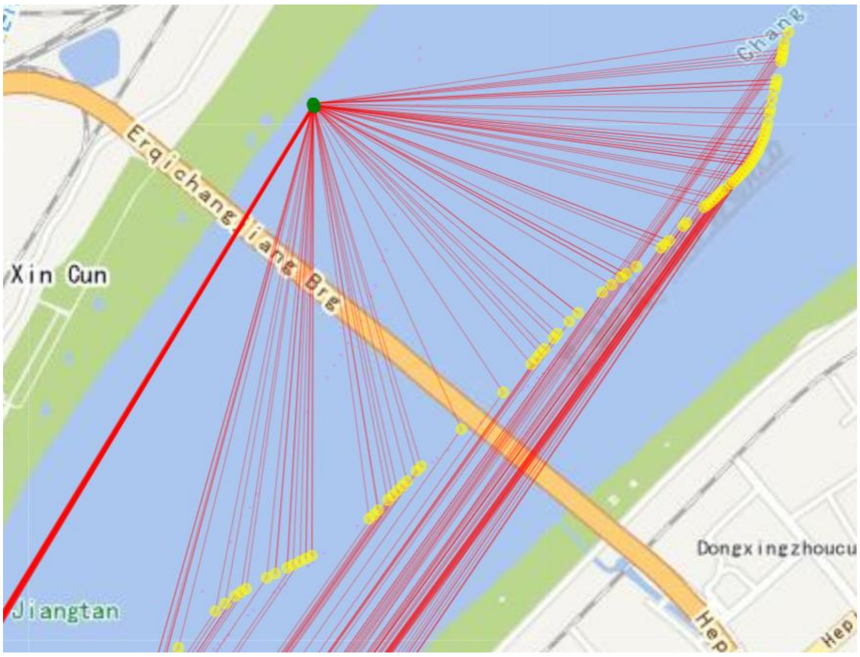


Figure 9. Duplicate mover detection algorithm applied to a trajectory in the Wuhan dataset, showing one stop point (green) and one mover (yellow). The red lines show the input trajectory

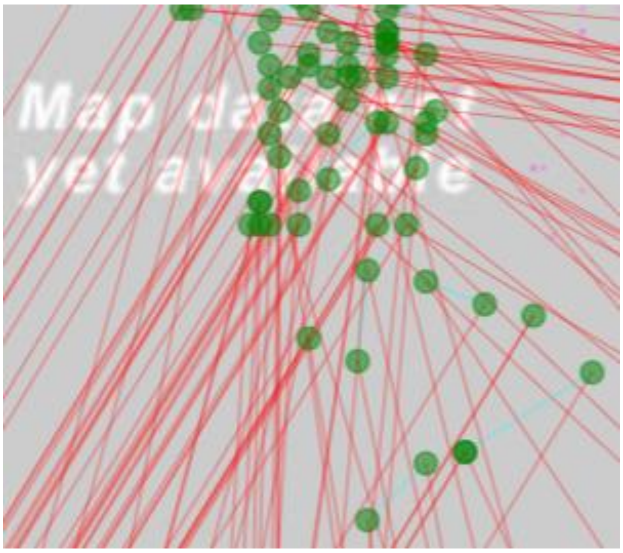


Figure 10. Close up view of a stop location showing jitter

Fig. 9 above, shows an anomalous trajectory in red from the Wuhan dataset. The duplicate mover identifier algorithm detected two groups of trajectory points. A green cluster around a stop point and a yellow cluster of a mover.

After the completion of the grouping, the corrected trajectory is trivial to obtain. Each group contains a list of points belonging to it. These points are used to generate a corrected sub-trajectory. These are output to a file with a renamed trajectory identifier. The new trajectory identifier is obtained by appending the group number to the old trajectory identifier.

3.3 Visual-interactive component

As explained in Chapter 2, a visual analytical approach is needed in approaching the problem of duplicate movers. The operator expertise will be needed to determine the ideal detection algorithm parameters. As shown in Fig. 11, The task oriented view of information visualization (Maletic et al. 2002) describes visualization as a pipelined process that takes one from the raw data to different views of the same data. Since it is essentially exploratory in nature, the user can interact with the visualization at each step to get the desired insights or to effect an outcome.

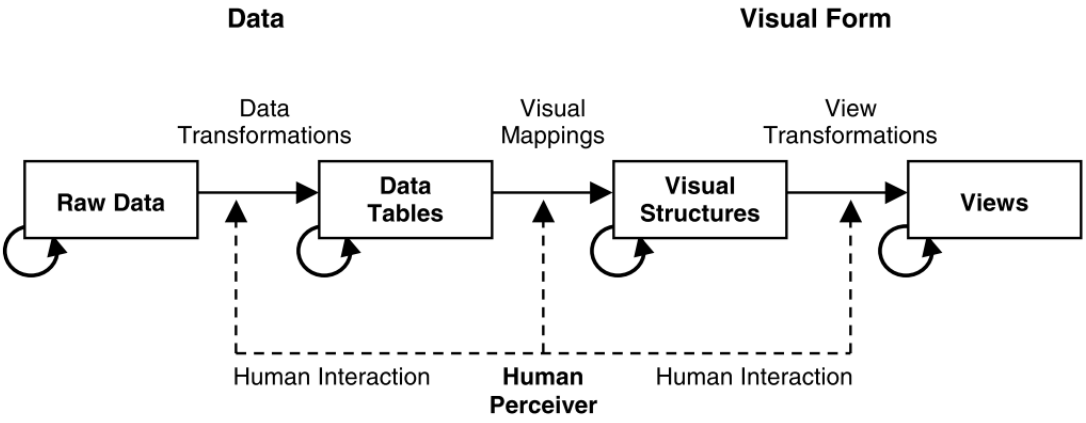


Figure 11. Task oriented view of Information Visualization (Maletic et al. 2002)

As a first step, a visualization method is needed to display the uncorrected anomalous trajectories.

The anomalous trajectories have a large number of overlapping segments, which leads to overplotting and loss of details when visualized. It is still possible to discern some details such as stop positions and long jumps if the trajectories segments are colored based on the count of overlapping segments. The most frequently overlapping areas appear brighter, as seen in Fig. 12 and Fig. 13.

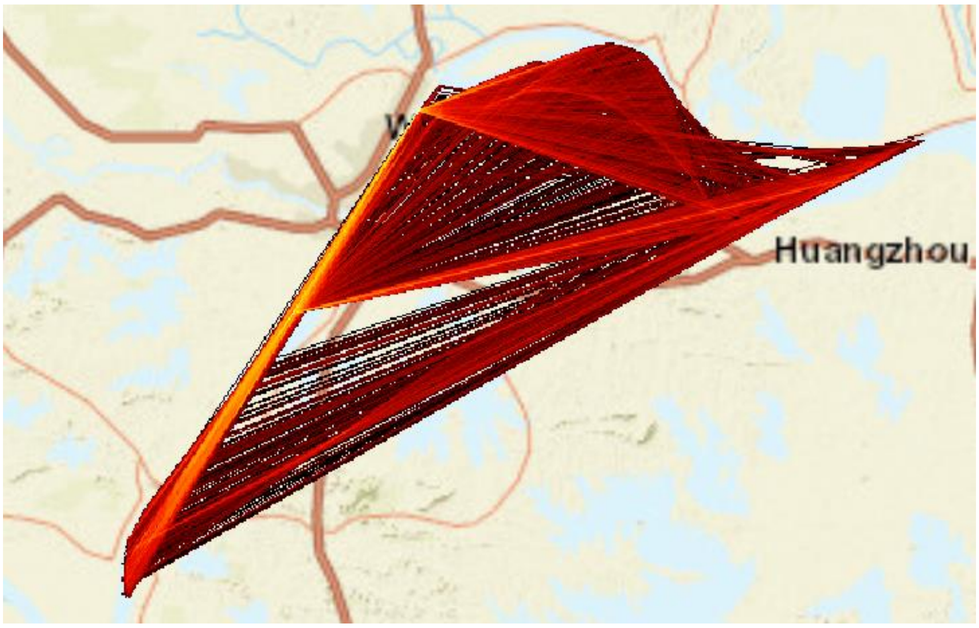


Figure 12. Visualization of an uncorrected anomalous trajectory

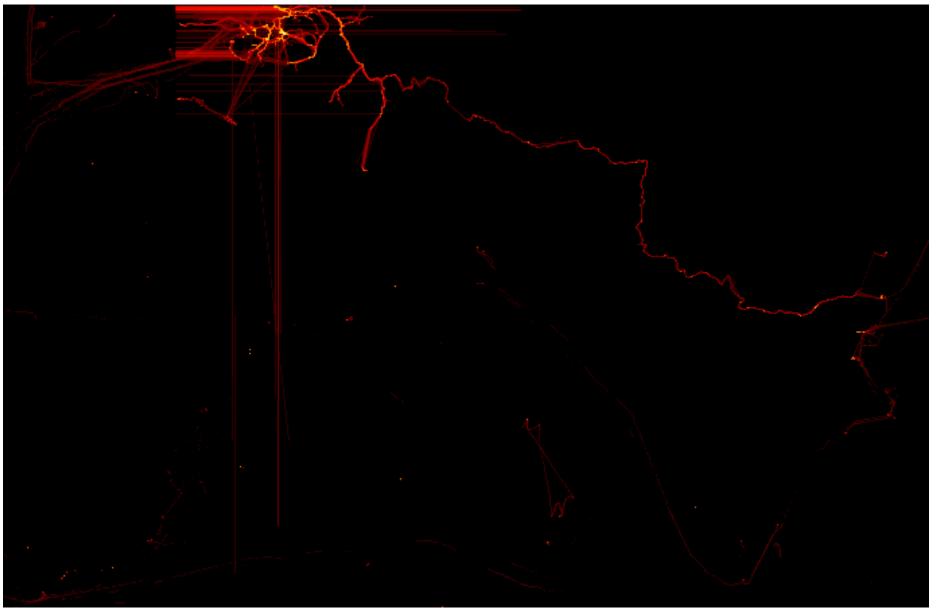
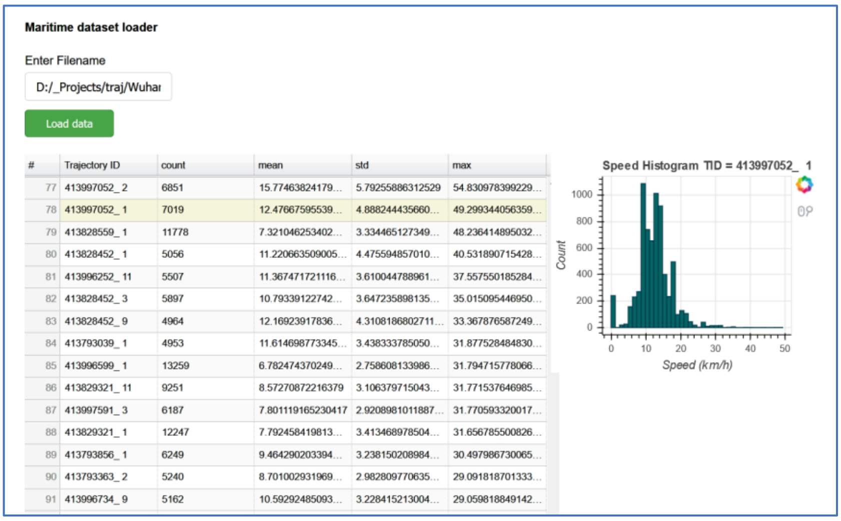


Figure 13. Visualization of 3 million records from the European maritime dataset. The frequently occurring long distance jumps are clearly visible.

The maritime dataset loader interface allows an operator to load a dataset, and view a table of statistical information of each trajectory. By interacting with the table, a linked speed histogram. Thus, a decision can be made to further investigate a trajectory, if needed.



Maritime dataset loader

Enter Filename

D:/_Projects/traj/Wuhar

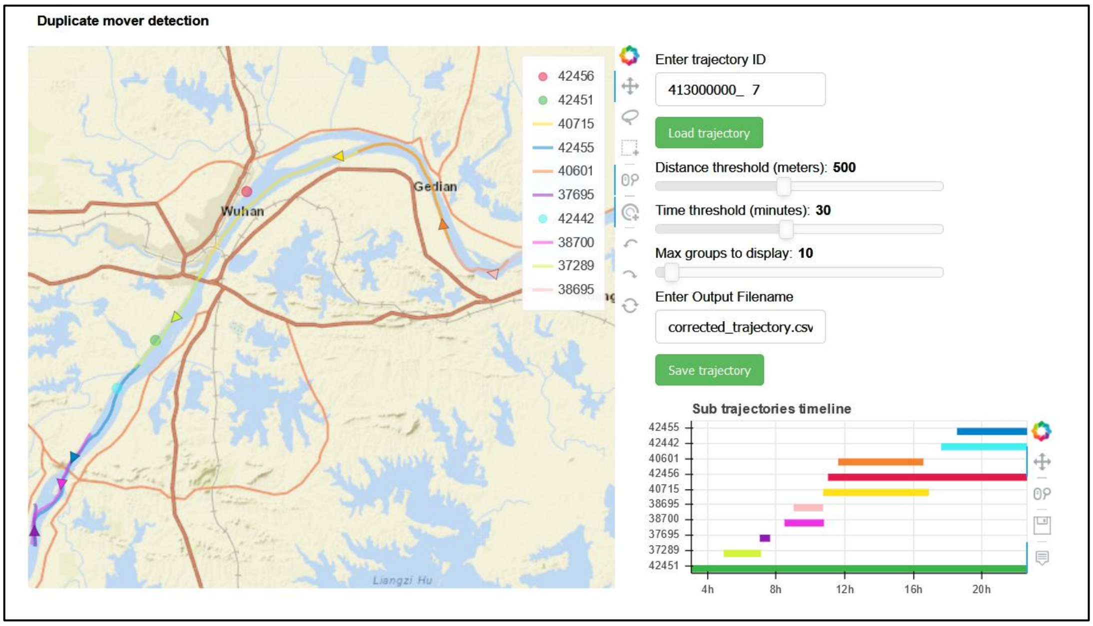
Load data

#	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...

Figure 14. Screenshot of the data loader interface

The duplicate mover detection interface (Fig. 15) consists of a large slippy map panel allied with widgets on the right side. This is coupled with a linked timeline display sub-trajectories timeline. The widgets allow the operator to load a selected trajectory by the trajectory identifier. Furthermore, the parameters of the mover identifier detection algorithm such as distance and time can be configured.

The basic interaction primitives being used are pan and zoom, filter and details on demand (Shneiderman 2003). In order to reduce clutter, the top n largest groups can be displayed using the slider widget.



Duplicate mover detection

Enter trajectory ID

413000000_ 7

Load trajectory

Distance threshold (meters): **500**

Time threshold (minutes): **30**

Max groups to display: **10**

Enter Output Filename

corrected_trajectory.csv

Save trajectory

Figure 15. Visual Interface for duplicate mover detection

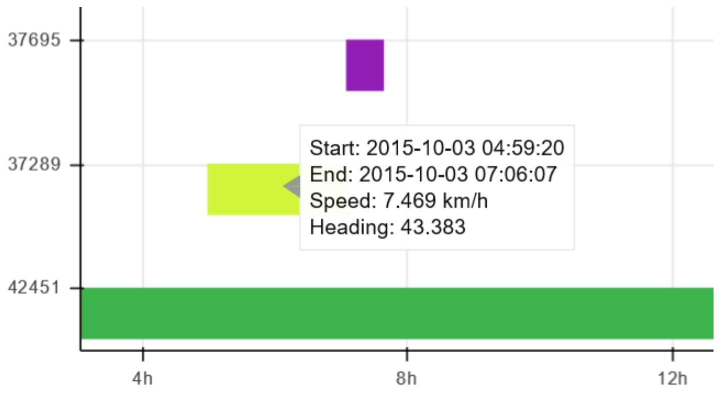


Figure 16. Sub-trajectory details are revealed on demand

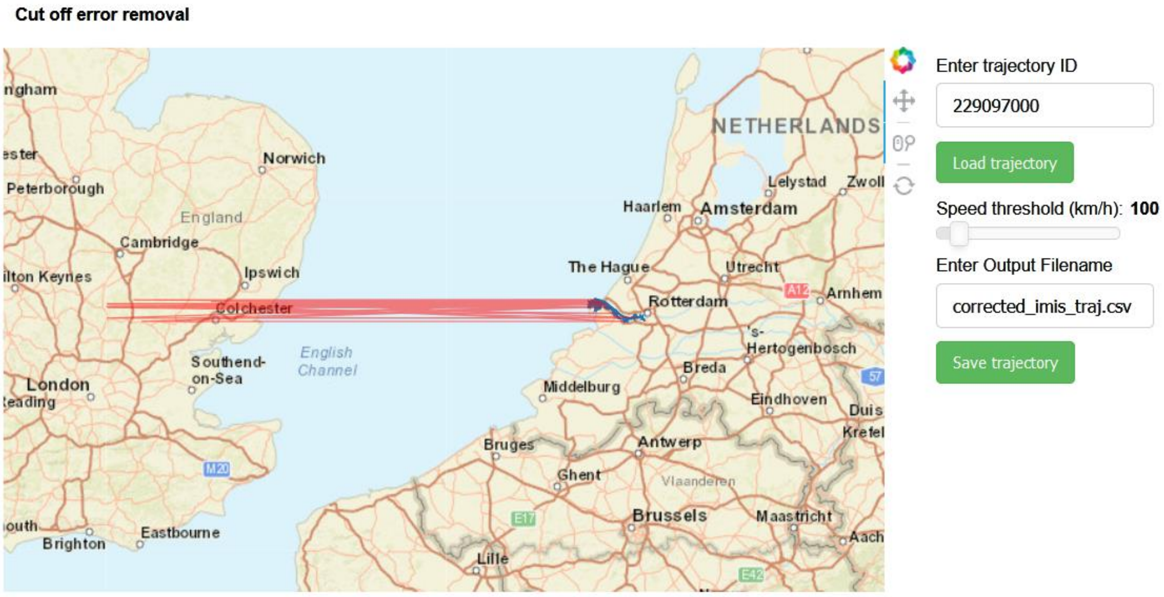
The slippy map allows pan and drag as well as zoom interactions. The sub-trajectories are displayed as line segments with an arrow indicating the direction of travel. The detected stop points are represented by a single circle, instead of a cluster of very closely spaced points. This improves legibility of the map.

A typical maritime trajectory can have thousands of points. However, the trajectory can be fairly represented by a smaller number of points. The Douglas-Pecker algorithm (Douglas, Peucker 1973) is utilized to simplify the displayed trajectory line segments. Reducing the number of points in the visualized trajectory not only reduces clutter but also improves the responsiveness of the visualization.

The timeline view gives a quick overview of the relative temporal position of each sub-trajectory. It is basically a stacked horizontal bar graph linked to the currently displayed data on the slippy map. Thus, the interface allows and operator to compare the relative start and end timings of the sub-trajectories. Stop points are seen as a continuous line in the timeline view. If two or more sub-trajectories follow each other without overlap, there is a possibility that they may be a single trajectory. The timeline view gives hints on the temporal properties of the identified movers. Finally, the corrected trajectory can be written to file using the interface.

3.4 Handling of cut-off errors and sudden jumps

As discussed in Chapter 2, cut-off errors can occur when trajectory datasets are cropped from a larger source dataset. Misconfiguration and transient failures in position acquisition can also cause sudden sporadic jumps to distant or spurious locations such as 0°N 0°E. A straightforward way to filter these sudden jumps is to compute the inter point speeds and then delete trajectory segments that fall above a given speed threshold. The visual interface for handling this class of errors is shown in Fig. 17.



Cut off error removal

Enter trajectory ID

229097000

Speed threshold (km/h): **100**

Enter Output Filename

corrected_imis_traj.csv

Figure 17. Cut off error removal interface

The speed threshold slider can be adjusted to a desired value. All trajectory segments above the speed threshold are displayed in red with the rest in blue. An operator can quickly get feel for the desired threshold values since the impact is immediately visible in the slippy map. Finally, the corrected trajectory can be exported to an output file.

3.5 Batch Processing

The primary purpose of the visual interface is to aid in data exploration and to determine the correct parameter settings for the datasets. Following this, the correct algorithms will be applied to the rest of the data in batch mode.

4 Evaluation of the maritime use case

The threshold based methodology for detection and correction of anomalous trajectories can conceivably be applied to any form of locomotion. However, within the context of this study, focus was given to the maritime domain. Some assumptions related to the maritime domain are made. For example, ships generally execute turns slowly and do not make sudden high angular movements. Therefore, a series of consecutive high angular movements with low displacement is considered to the side effect of jitter and hence taken to be a stop point.

Two AIS datasets were analyzed. The datasets capture a month's worth of vessel traffic in the Chinese Wuhan region and the European coasts respectively. The European dataset covers large parts of the Western European coastline in addition to the Mediterranean Sea and the Black Sea.

Table 2. Details of the datasets used

	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

The Wuhan datasets exhibits several trajectories having the mover identifier problem, whereas in the European dataset, cut-off errors are more frequently seen.

4.1 Mover identity problem in the Wuhan dataset

The visual data loader interface was used to quickly narrow down problematic trajectories. Initial inspection of these trajectories showed that they had the distinctive criss-cross trajectory segments caused by the mover identifier problem.

As an example, a specific trajectory from the Wuhan dataset is considered. Details of this trajectory are given in the following table.

Table 3. Details of the anomalous trajectory under analysis

Trajectory ID	413000000_ 7
Npoints	6559
Duration	19 hours
Total distance	125880 km
Mean Speed	6625 km/h

Loading the trajectory in the visual interface with the default thresholds of 500m for distance and 30 minutes for time indicates that there are at least three stop locations. It's likely that some of the vessels were docked at these points.

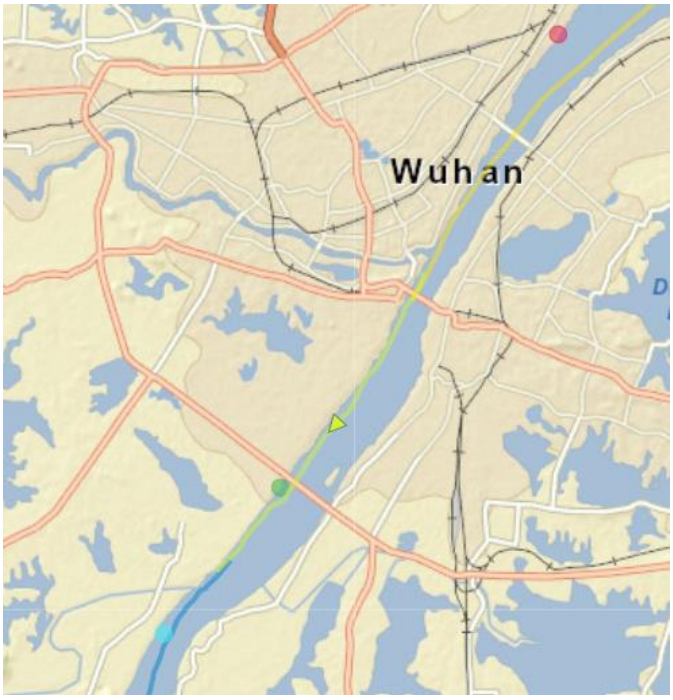


Figure 18. Stop locations identified by the colored circles

The largest identified mover groups are displayed as colored lines with a glyph indicating the direction of movement. The general direction of movement of the movers is from east to west from Wuhan towards Jinkou. A domain expert operator would be able to make a more informed decision by adjusting the algorithm parameters and observing the changes in the map and the timeline view.

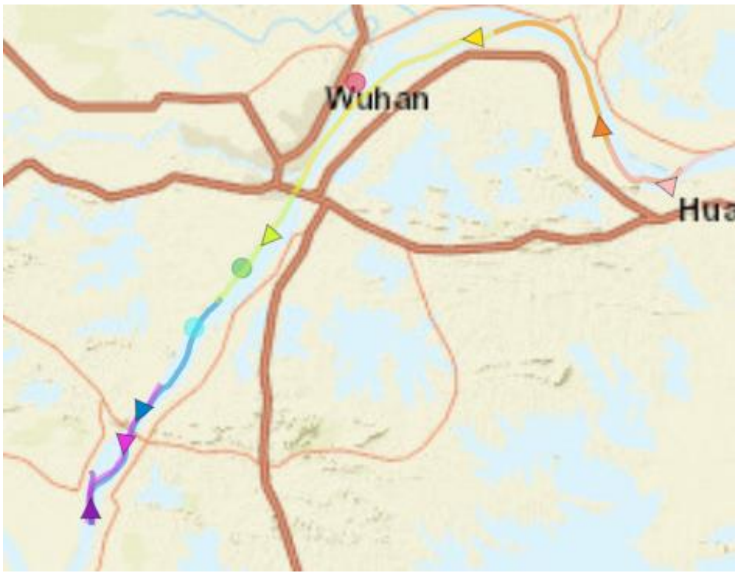
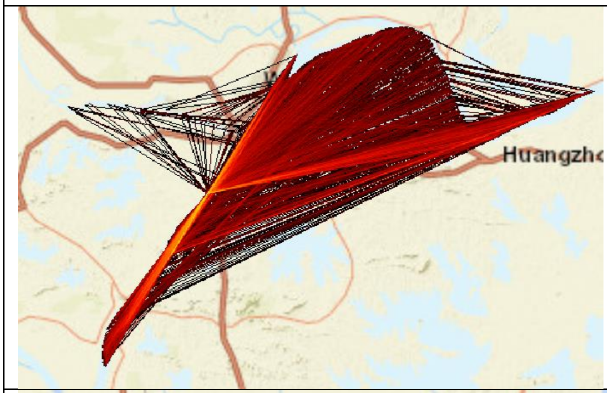
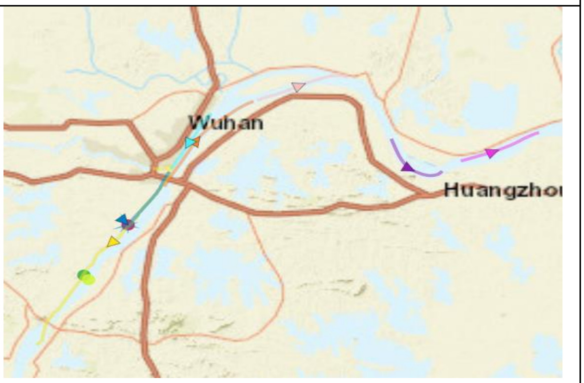
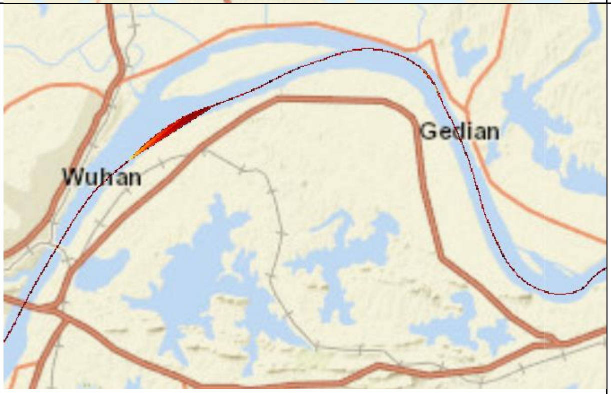
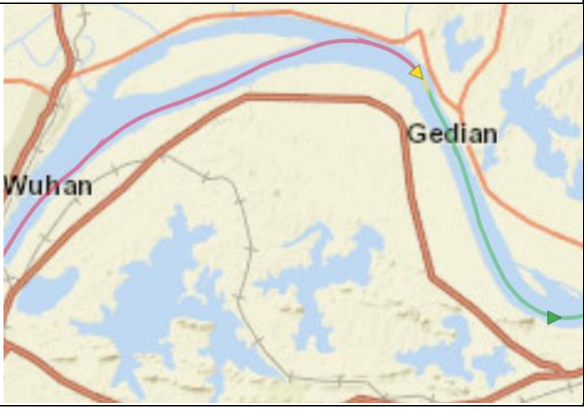


Figure 19. Identified movers shown as colored sub-trajectories

Some of the processing outputs with other trajectories from the Wuhan dataset are shown in the Fig. 20 below.

Before correction	After mover ID correction
	
	

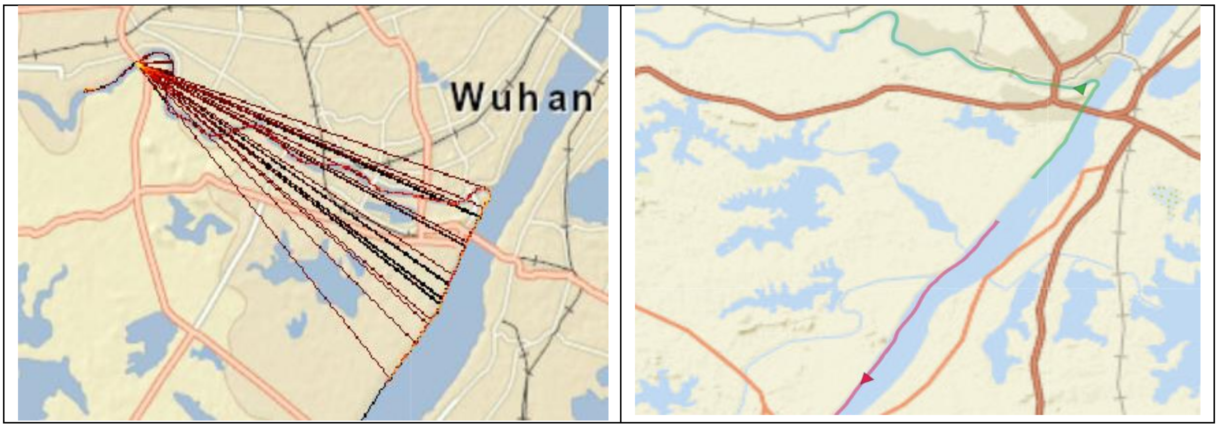


Figure 20. Examples of mover detection (before and after correction)

4.2 Cut-off problem in the European dataset

In order to visualize the problems in the dataset, several trajectories from the dataset were first visualized. The cut-off problem is then immediately apparent since large jumps are seen from river bodies to the middle of far-flung land masses.

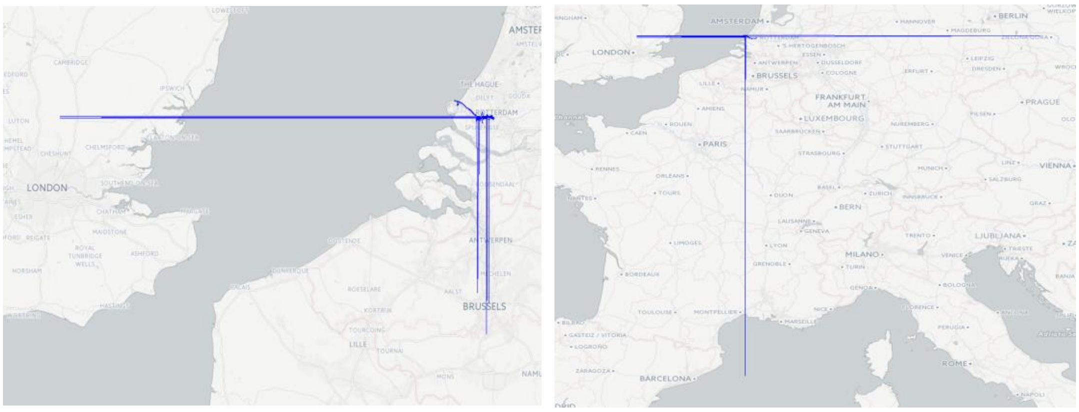


Figure 21. Trajectories in the European dataset showing cut-off errors

The visual interface was used to load several trajectories and test out various thresholds. The selected threshold value of 110 km/h was found to be remove most of the cut-off errors. This was also cross checked by looking at the histogram of speed distribution within this dataset.

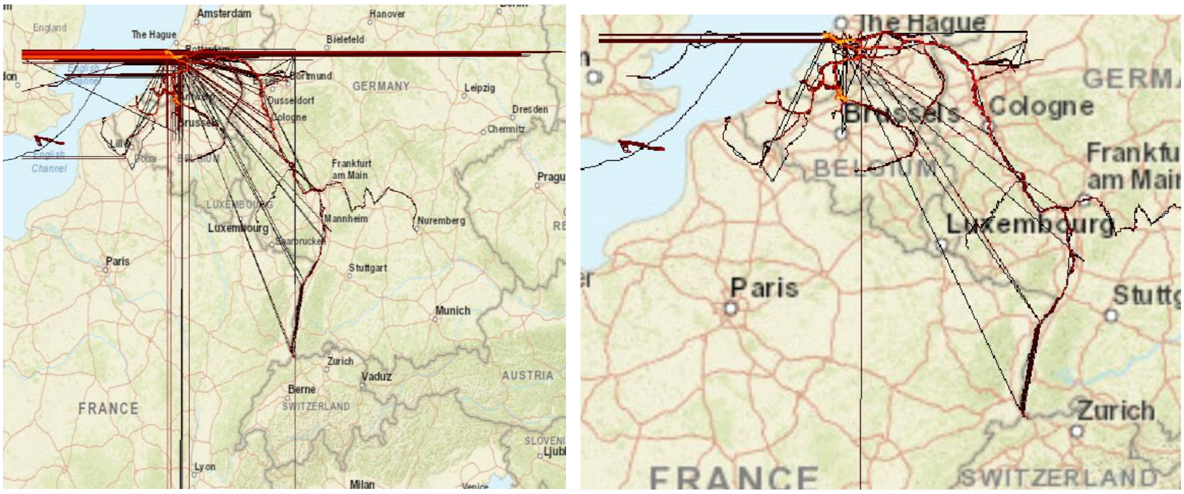


Figure 22. Visualization of the European dataset before and after applying speed based correction

Although there is a visible reduction in the number of cut-off errors (Fig. 22), there are still several long distance segments remaining. These remaining long jump errors occur due to large temporal gaps between trajectory points. It could be solved splitting the trajectory into sub-trajectories at points where the large temporal gaps are seen. It can also be feasibly approached by applying mover ID correction after the removal of cut-off errors.

In summary, the algorithms developed can target the mover ID problem as well as cut-off errors. However, an expert human operator is needed firstly to recognize which method is to be applied and secondly to determine the correct parameter settings for batch operation.

5 RESULTS

The previous chapters described the technique used to detect duplicate movers and to correct the cut-errors. To illustrate the results numerically, sample trajectories from each dataset will be chosen.

5.1 Wuhan dataset

The Wuhan dataset has 45,318 trajectories. A test trajectory with 6559 points is chosen for analysis. Prior to correction, the test trajectory had a total distance of 125880 km. After duplicate mover id correction, the total distance covered by the sub-trajectories is 300 km with an average speed of 9.3 km/h. Visualization of the corrected sub-trajectories indicate that criss-crossing line segments indicative of the mover identity problem are no longer present. Points with consistent jitter were classified as stop points. The general pattern of motion of the movers was determined.

5.2 European dataset

A test trajectory with 8965 points was chosen for analysis. The cut-off error algorithm detected 131 segments above the chosen threshold of 110 km/h. The average speed of the trajectory reduced from 193 km/h to 3.2 km/h after correction. The standard deviation of speeds reduced from 3421 to 5.7. This indicates that the corrected dataset has more uniform speed, which matches our expectations. Visual inspection also indicates a reduction in the number of anomalous jumps.

5.3 Performance

Since the developed technique are applied to extremely large datasets, it is vital that the efficient algorithms are used. In the case of the duplicate mover identification algorithm, R-Tree, which has highly efficient nearest neighbor search implementations (Roussopoulos et al. 1995) were used.

The compute performance of the algorithm was tested on a laptop with following specifications.

Table 4. Specification of system used for performance test

CPU	Intel core i5 5200U
RAM	8 GB
OS	Windows 10 64 bit
Hard Drive	Seagate Momentus 5400 rpm

In the Wuhan dataset average length of a trajectory is 297. The test trajectory with 6559 points was processed by the mover id detection algorithm in 1.87s. It should be noted that 75% of the trajectories were less than 338 points in length, and so processing of those will be much faster.

The time complexity of the cut-off error detection is much better since the algorithm is quite simple. The test trajectory of 8965 points was processed in 7.12ms.

5.4 Limitations

One of the potential issues with the mover detection algorithm is the creation of large number of sub-trajectories. In order to mitigate this problem from a visual perspective, only the largest sub-trajectories were displayed in the user-interface by default.

The cut-off detection algorithm does not deal with the case where there are large time deltas between successive points. As mentioned, this can be dealt with by splitting the trajectory into sub-trajectories.

6 CONCLUSION AND RECOMMENDATIONS

The aim of the thesis was to address develop an algorithm to address a frequent class of data quality problem seen in movement datasets. Mover identity problems can occur in any movement dataset where one or more movers share the same identifier within a trajectory. Cut-off errors are another related class of errors. It presents itself as anomalous jumps within a movement dataset and is caused due to cropping of the original source of the data. The technique to detect the mover identity problem is based on grouping trajectory points to their closest groups by time and distance thresholds. Cut-off errors on the other hand are detected by looking for and deleting segments that have very high velocities.

The solution to both problems requires a visual-analytic approach. A visual-interactive interface using slippy maps and configurable widgets was developed to address both classes of movement errors. The techniques developed were applied to two datasets, one from the Wuhan region and another covering the European coastlines. The performance of both algorithms was tested.

This thesis presented a new way of visualizing and correcting trajectories which exhibit the mover identity problem. Understanding the data quality errors in a movement dataset requires a combination of detection algorithms, visualization methods and human expertise. The work done in this thesis further strengthens that idea.

6.1 Future work

An initial study in the domain of mover identity problems was conducted as a part of this thesis. As a part of future research effort, it is possible to improve the accuracy of the duplicate mover detection by incorporating more motion characteristics such as acceleration and turning angle. The contextual information regarding the detected movers can be expanded to include more types of information such as vessel class, origin and destination.

The focus in this study was on maritime datasets. However, the data quality problems can also occur in other movement datasets such as flight trajectories or urban transport. Since the technique is general in nature, it would be promising to customize and apply them to other modes of transport.

It is also worth exploring the possibility of creating a movement data quality correction algorithm that handles multiple classes of errors in parallel. This is would especially useful as a detection and correction component of a distributed and scalable movement data quality analysis platform.

Publication bibliography

Alvarez, Pablo; Leeson, Andrew (2015): Big data helps pedestrian planning take a big step forward. In *Transportation Professional*.

Andrienko, Gennady; Andrienko, Natalia (2008): A Visual Analytics Approach to Exploration of Large, pp. 1–4.

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.

Attard, Judie; Orlandi, Fabrizio; Scerri, Simon; Auer, Sören (2015): A systematic review of open government data initiatives. In *Government Information Quarterly* 32 (4), pp. 399–418.

Chang, S. J. (2004): Development and analysis of AIS applications as an efficient tool for vessel traffic service. In : Oceans '04 MTS/IEEE Techno-Ocean '04 (IEEE Cat. No.04CH37600). Oceans '04 MTS/IEEE Techno-Ocean '04. Kobe, Japan, 9-12 Nov. 2004: IEEE, pp. 2249–2253.

Deng, Feng; Guo, Sitong; Deng, Yong; Chu, Hanyue; Zhu, Qingmeng; Sun, Fuchun (2013): Vessel track information mining using AIS data.

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).

John Walker, Saint (2014): Big data. A revolution that will transform how we live, work, and think: Taylor & Francis.

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.

Keim, Edited Daniel; Kohlhammer, Jörn; Ellis, Geoffrey (2010b): Mastering the Information Age. Solving Problems with Visual Analytics, Eurographics Association.

Kim, Gang-Hoon; Trimi, Silvana; Chung, Ji-Hyong (2014): Big-data applications in the government sector. In *Commun. ACM* 57 (3), pp. 78–85. DOI: 10.1145/2500873.

Kranstauber, Bart; Cameron, Alison; Weinzerl, R.; Fountain, Tony; Tilak, Sameer; Wikelski, Martin; Kays, Roland (2011): The Movebank data model for animal tracking. In *Environmental Modelling & Software* 26 (6), pp. 834–835.

Laxhammar, Rikard (2008): Anomaly detection for sea surveillance, pp. 55–62.

Liao, Ziqi (2003): Real-time taxi dispatching using global positioning systems. In *Commun. ACM* 46 (5), pp. 81–83.

Maletic, J. I.; Marcus, A.; Collard, M. L. (2002): A task oriented view of software visualization. In : First international workshop on visualizing software for understanding and analysis. First International Workshop on Visualizing Software for Understanding and Analysis. Paris, France, 26 June 2002. Institute of Electrical and Electronics Engineers: IEEE Comput. Soc, pp. 32–40.

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.

Ristic, B.; La Scala, B.; Morelande, M.; Gordon, N. (2008): Statistical Analysis of Motion Patterns in AIS Data. Anomaly Detection and Motion Prediction, pp. 40–46.

Roussopoulos, Nick; Kelley, Stephen; Vincent, Frédéric (1995): Nearest neighbor queries. In Michael Carey, Donovan Schneider (Eds.): Proceedings of the 1995 ACM SIGMOD international conference on Management of data - SIGMOD '95. the 1995 ACM SIGMOD international conference. San Jose, California, United States, 22-May-95 - 25-May-95. New York, New York, USA: ACM Press, pp. 71–79.

Shahir, Hamed Yaghoubi; Glasser, Uwe; Shahir, Amir Yaghoubi; Wehn, Hans (2015): Maritime situation analysis framework. Vessel interaction classification and anomaly detection. In *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pp. 1279–1289. DOI: 10.1109/BigData.2015.7363883.

Shneiderman, Ben (2003): The eyes have it. A task by data type taxonomy for information visualizations. In : The Craft of Information Visualization: Elsevier, pp. 364–371.

Silveira, P. A.M.; Teixeira, A. P.; Soares, C. Guedes (2013): Use of AIS data to characterise marine traffic patterns and ship collision risk off the coast of Portugal. In *The Journal of Navigation* 66 (6), pp. 879–898.

Vaughan-Nichols, Steven J. (2009): Will mobile computing's future be location, location, location? In *Computer* 42 (2), pp. 14–17.

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.

Zhang, Daqing; Li, Nan; Zhou, Zhi-hua; Chen, Chao; Sun, Lin; Li, Shijian (2011): iBAT - Detecting Anomalous Taxi Trajectories from GPS Traces. Detecting Anomalous Taxi Trajectories from GPS Traces. In *(None)*.

Zheng, Y. U. (2015): Trajectory Data Mining. An Overview 6 (3), pp. 1–41.