Ship Movement Anomaly Detection Using Specialized Distance Measures

Bo Liu , Erico N. de Souza , Cassey Hilliard and Stan Matwin

Presented by: Bo Liu

July 8th, 2015

DALHOUSIE UNIVERSITY
INSTITUTE FOR BIG DATA ANALYTICS

Outline

← Research Background

← Overview of the Framework

← Review of the Clustering Phase

← Abnormal Trajectory Detection

← Experiments

← Discussion & Future Work

Research Background

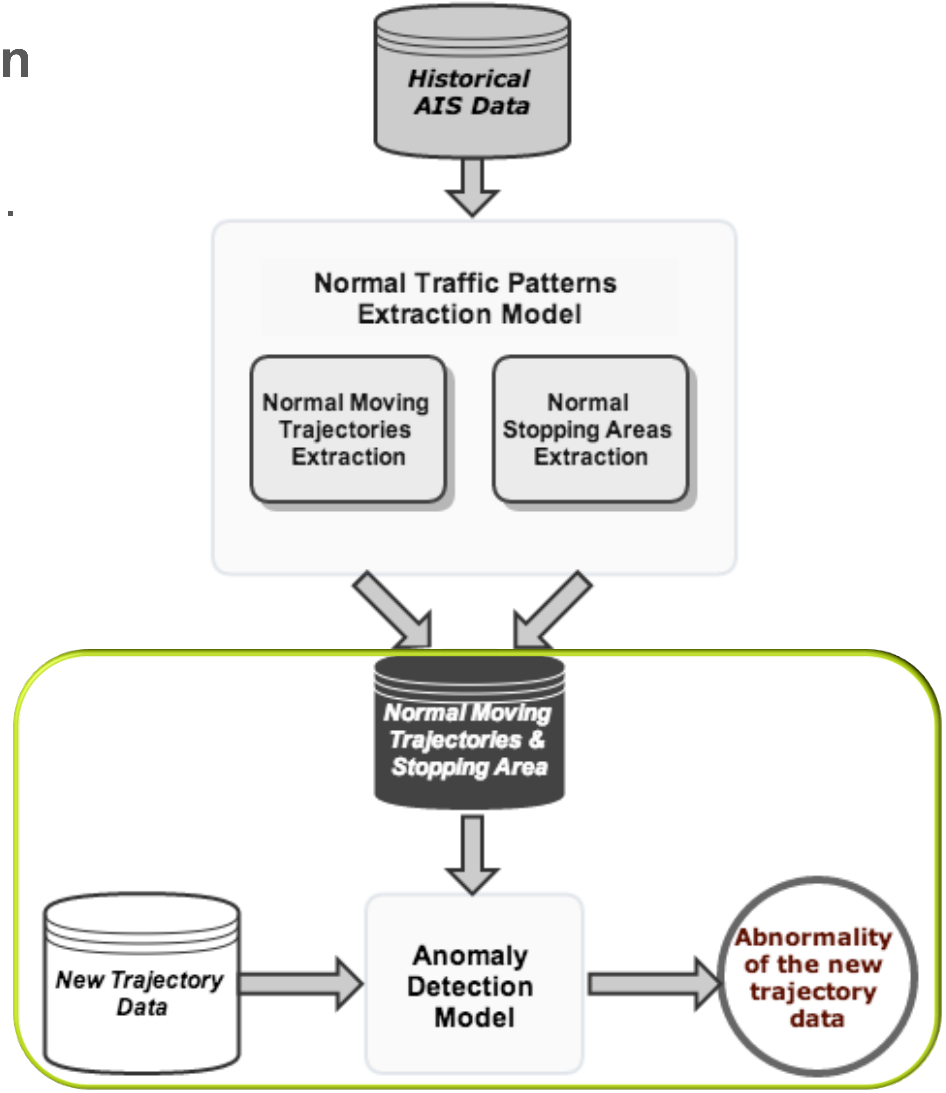
- **AIS** (Automatic Identification System)
  - Automatic tracking system for identifying and locating vessels
  - The use of AIS has been required by **IMO** (International Maritime Organization) since 2004
    - **IMO** is the global standard-setting authority for the safety, security and environmental performance of international shipping [1]
  - **Over 400,000** ships worldwide have installed the AIS transponders [1]
  - At least **100M** records/day
- AIS data from **near-port regions**
  - The traffic is **highly variable,** the vessels always change their **directions and speeds**
  - IMO controls the navigational lanes (**TSS**) that vessels must use when approaching or exiting some ports around the world.
    - TSS is Traffic Separation Schemes **without regulations on speed**
- Anomaly detection in near-port areas is challenging but worthy

[1] http://www.imo.org/en/About/Pages/Default.aspx

Overview of the Framework

- **Our previous work presented in** [2]
  - Based on **stop-and-move** model. [3]
  - For **moving trajectory points:**
    - **DBSCANSD + GV**
  - For **stopping trajectory points:**
    - **DBSCAN** [4] **+ SSP**

- **Work done in this paper** ⟶
  - Based on the above extracted GVs and SSPs
  - Propose three specialized Division Distances for anomaly detection



Historical AIS Data
Normal Traffic Patterns Extraction Model
Normal Moving Trajectories Extraction
Normal Stopping Areas Extraction
Normal Moving Trajectories & Stopping Area
New Trajectory Data
Anomaly Detection Model
Abnormality of the new trajectory data

Fig 1. Overview of the anomaly detection framework

Review of the Clustering Phase

- **DBSCANSD** (**D**ensity **B**ased **S**patial **C**lustering of **A**pplication with **N**oise considering **S**peed and **D**irection)
- **Key idea of DBSCANSD**
  - For each point of a cluster the neighborhood of a given radius has to contain at least a minimum number of points **that have similar speed and direction**
- **Advantages**
  - No need to specify the number of clusters
  - Capable of finding arbitrary shape of clusters
  - Robust to outliers
  - Capable of taking speed and direction into account
  - Associated with IMO rules for parameter selection

Review of the Clustering Phase

- **Moving** clusters from DBSCANSD are represented using
    - **Gravity Vector (GV)** – centroid and envelope [5]
        - A GV is a vector of five features: average COG, average SOG, average Latitude, average Longitude and Median Distance.
        - Cluster is partitioned and each cell will have a particular GV
- **Stopping** clusters from DBSCAN are represented using
    - **Sampled Stopping Point (SSP)**
        - SSPs are sampled from the clustering results based on a modified random selection method
        - SSP depends only on the geographic shape of the stopping area

Review of the clustering phase

- An example of the clustering process in Los Angeles Long Beach port area

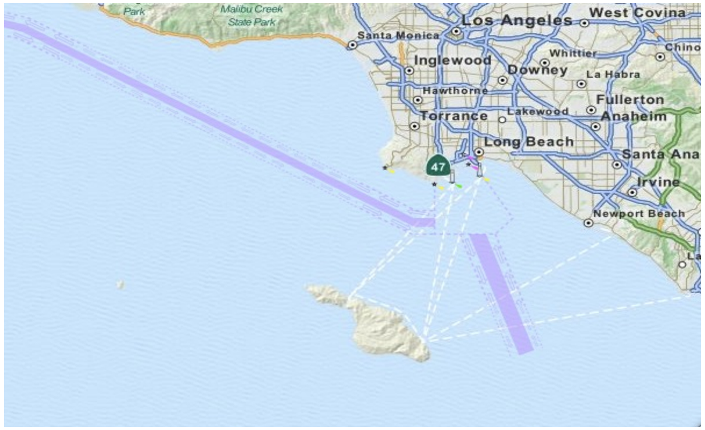


Fig 2. TSS Boundaries (IMO rules) defined in Los Angeles Long Beach area [2]

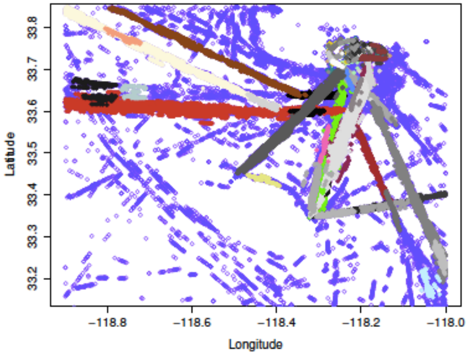


Fig 3. Clustering results after applying DBSCANSD on moving points

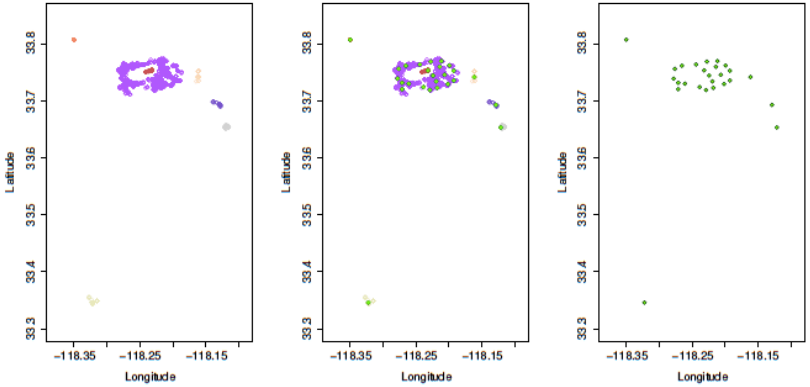


Fig 4. Extract SSPs from stopping clustering results using DBSCAN [2]

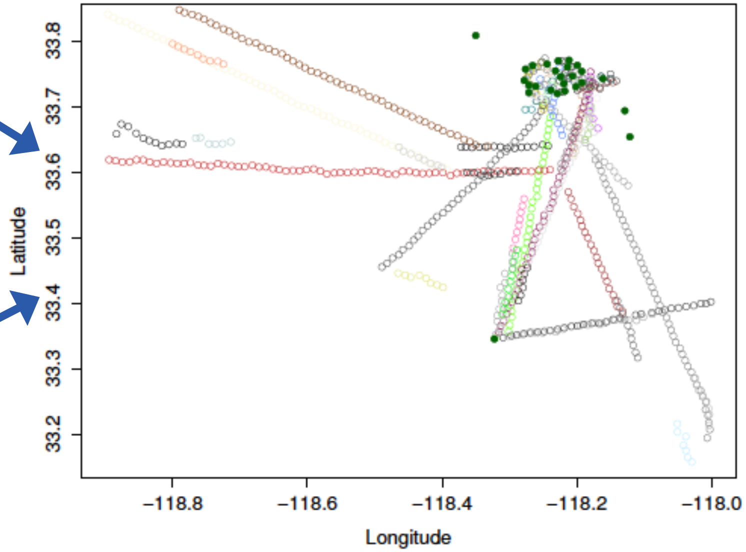


Fig 5. GVs (open circles) and SSPs (solid dark green circles) extracted from the clusters.

[2] http://www.openseamap.org

Abnormal Trajectory Detection

- **Two Division Distances in relation to location**
  - **Stopping Points Abnormality Detection**
    - **ADD** (Absolute Division Distance):

    The ADD between a target point $\boldsymbol{P_t}$ and a SSP $\boldsymbol{P_s}$ is defined as:

$$D_{absolute} = Distance((P_t.Lat, P_t.Lon), (P_s.Lat, P_s.Lon))$$

  - **Moving Points Abnormality Detection**
    - **RDD** (Relative Division Distance)

    The RDD between a target point point $\boldsymbol{P_t}$ and a GV ***GV*** is defined as:

$$D_{relative} = \frac{Distance((P_t.Lat,P_t.Lon),(GV.Lat,GV.Lon))}{GV.MedianDistance}$$

Abnormal Trajectory Detection

- **Why do we need a third division distance?**
  - **Two abnormal cases in relation to speed or direction**
    - *If we only consider ADD and RDD, the target points will be both considered as normal because they are close enough to the GVs in both cases.*

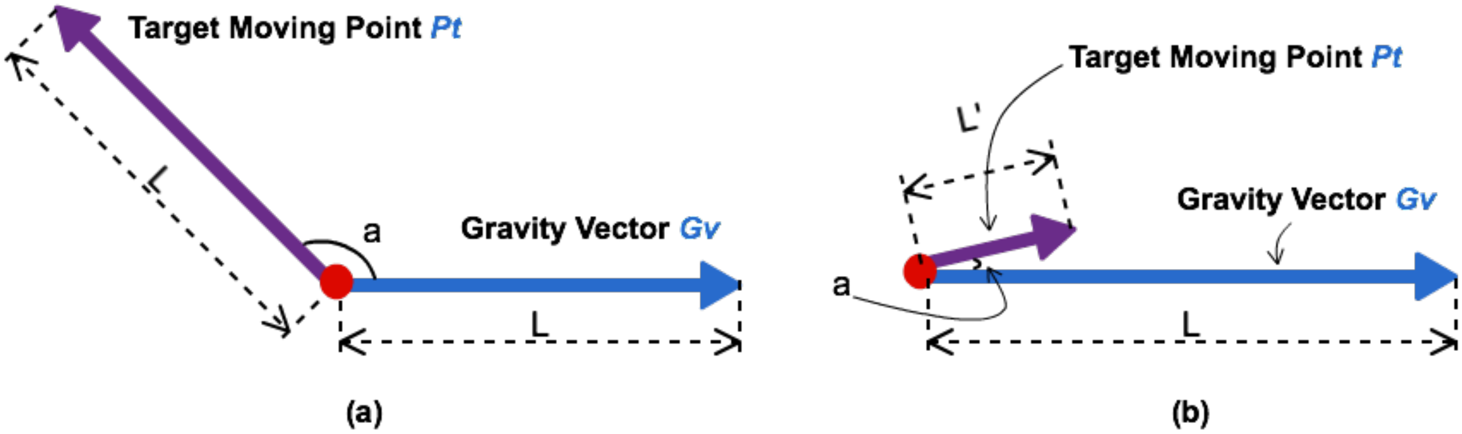


Fig 6. Two abnormal cases

Abnormal Trajectory Detection

- **The Third Division Distance in relation to SPEED and DIRECTION**
  - **Moving Points Abnormality Detection** *(continued)*
    - **CDD**(Cosine Division Distance):

    The CDD between a target point $\boldsymbol{P_t}$ and a GV $\boldsymbol{GV}$ is defined as:

$$D_{cosine} = \cos\alpha \times \frac{min(P_t.SOG,GV.SOG)}{max(P_t.SOG,GV.SOG)}$$

    where α is the angle between the two directions (the difference between $\boldsymbol{P_t}$'s COG and ***GV***'s COG).

    - **Why not calculate the distances of speeds and directions separately and then check the abnormality?**

Abnormal Trajectory Detection

- **Why not calculate the distances of speeds and directions separately and then check the abnormality?** *(continued)*
  - The alternative is not capable of handling the following situation.

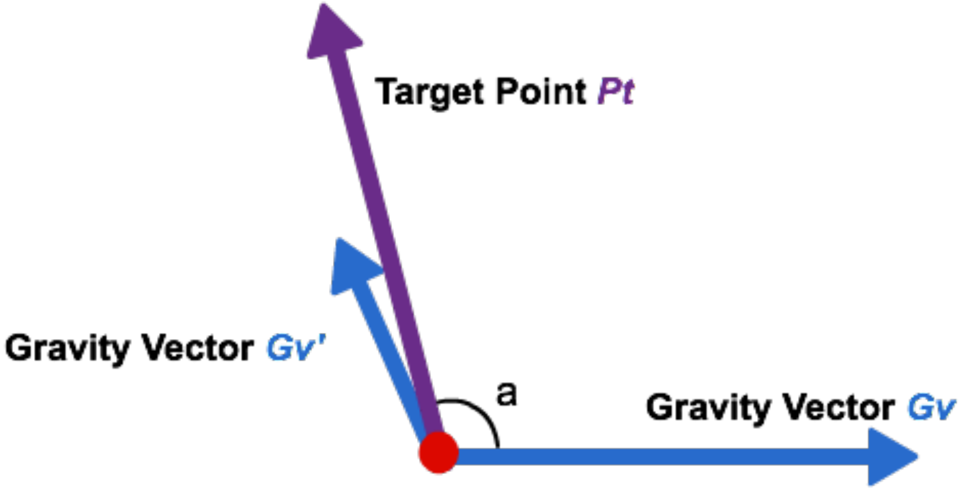


Fig 7. Case with multiple GVs around the target trajectory point

Abnormal Trajectory Detection

- The model treats stopping and moving points accordingly:
  - Initialize all points as **normal**
  - For stopping point $\boldsymbol{p_s}$

    Calculate the minimum ADD between $\boldsymbol{p_s}$ and all the SSPs

    If the distance is not less than ***add_threshold***, label it as **abnormal**
  - For moving point $\boldsymbol{p_m}$

    Calculate the minimum RDD between $\boldsymbol{p_m}$ and all the GVs

    IF the distance is not less than ***rdd_threshold,*** label it as **abnormal**

    ELSE

      Calculate the maximum CDD between $\boldsymbol{p_m}$ and all the GVs

      IF the distance is less than ***cdd_threshold,*** label it as **abnormal**

Experiments

- **Data Set**
  - **Two months** (November 1st – December 31st ,2012) of trajectory data in the area of *Juan de Fuca Strait* (**67,850 trajectory points**).
    - 46,000 trajectory points (40,000 moving points + 6,000 stopping points) selected for clustering
    - The rest of the data is used for estimating the three thresholds
  - The next half month (January 1st-15th , 2013) data is used as target data set.
    - 17,431 trajectory points
    - 284 unique trajectories (284 different MMSIs)
- **Two experiments**
  - Experiment On Unlabeled Data Set
  - Experiment On labeled Data Set

Experiments

- The IMO rules (TSS) of Juan de Fuca Strait and the clustering result (GVs and SSP)

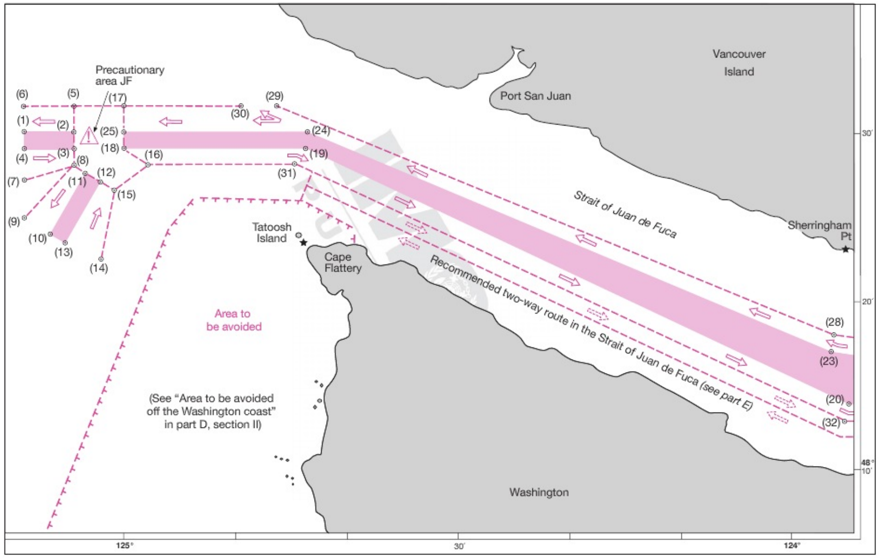


Fig. 8. Juan de Fuca Strait and its approaches (west) [4]

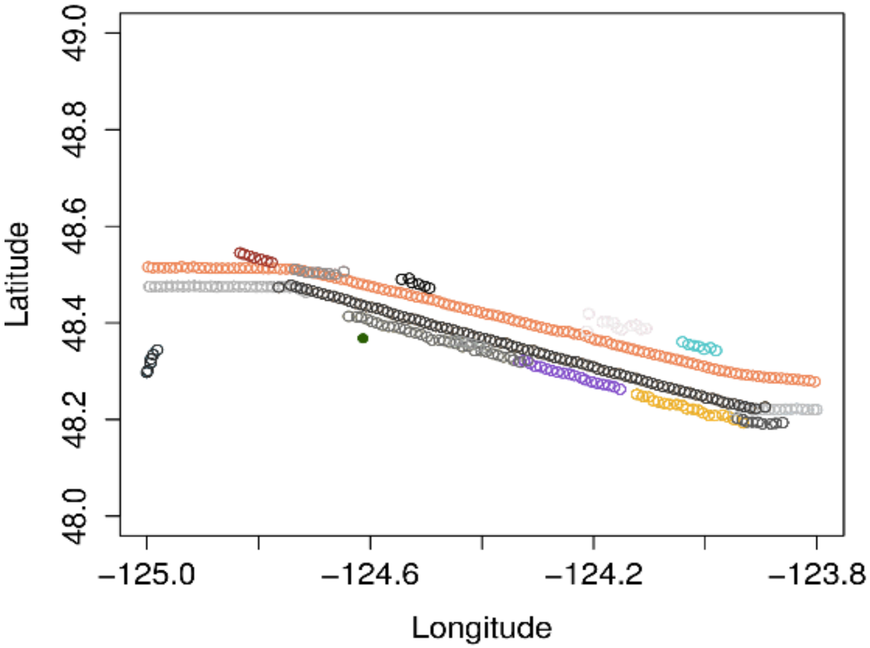
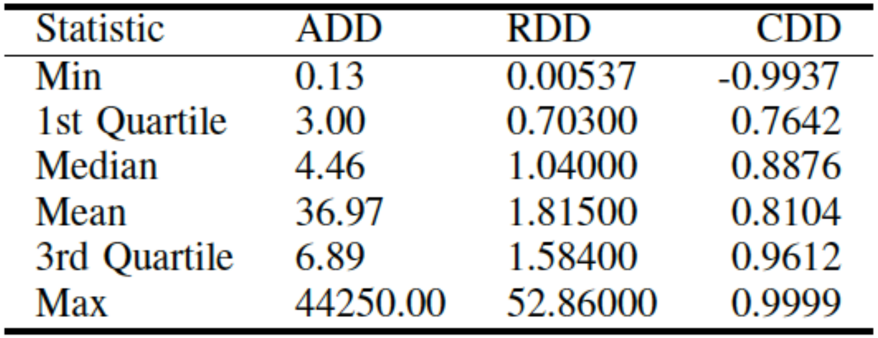


Fig. 9. The GVs and SSP extracted from the clusters in Juan de Fuca Strait area.

Experiment

- Estimate the three thresholds



Statistic	ADD	RDD	CDD
Min	0.13	0.00537	-0.9937
1st Quartile	3.00	0.70300	0.7642
Median	4.46	1.04000	0.8876
Mean	36.97	1.81500	0.8104
3rd Quartile	6.89	1.58400	0.9612
Max	44250.00	52.86000	0.9999

Table I Quatile Statistics of the three division distances

- We choose the sample quantiles of 0.95 for both ADD and RDD thresholds
  - ADD_threshold = 97.290
  - RDD_threshold = 5.938
- We select 0.05 as the possibility to decide the CDD threshold
  - CDD_threshold = 0.485

Experiment On Unlabeled Data

- Result of the target data set
  - Six abnormal trajectories examples→
  - Labeling results (below)

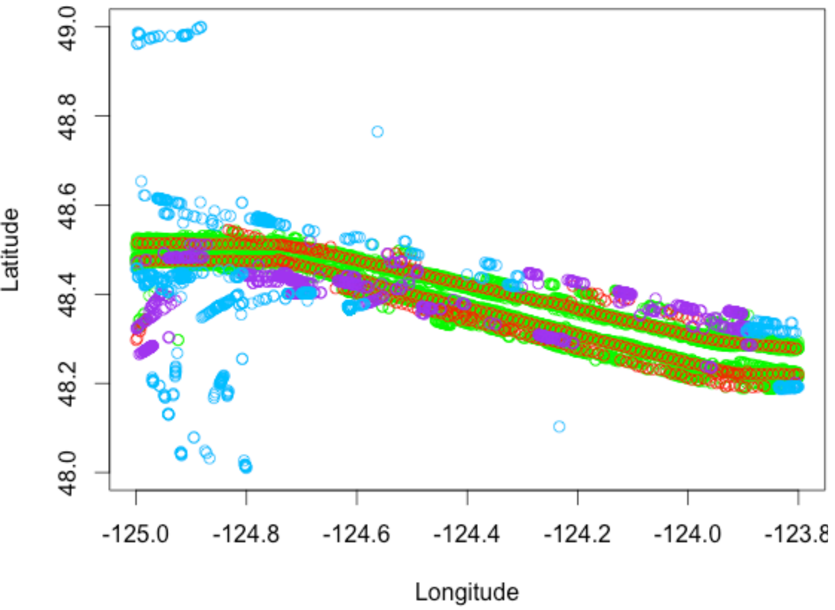
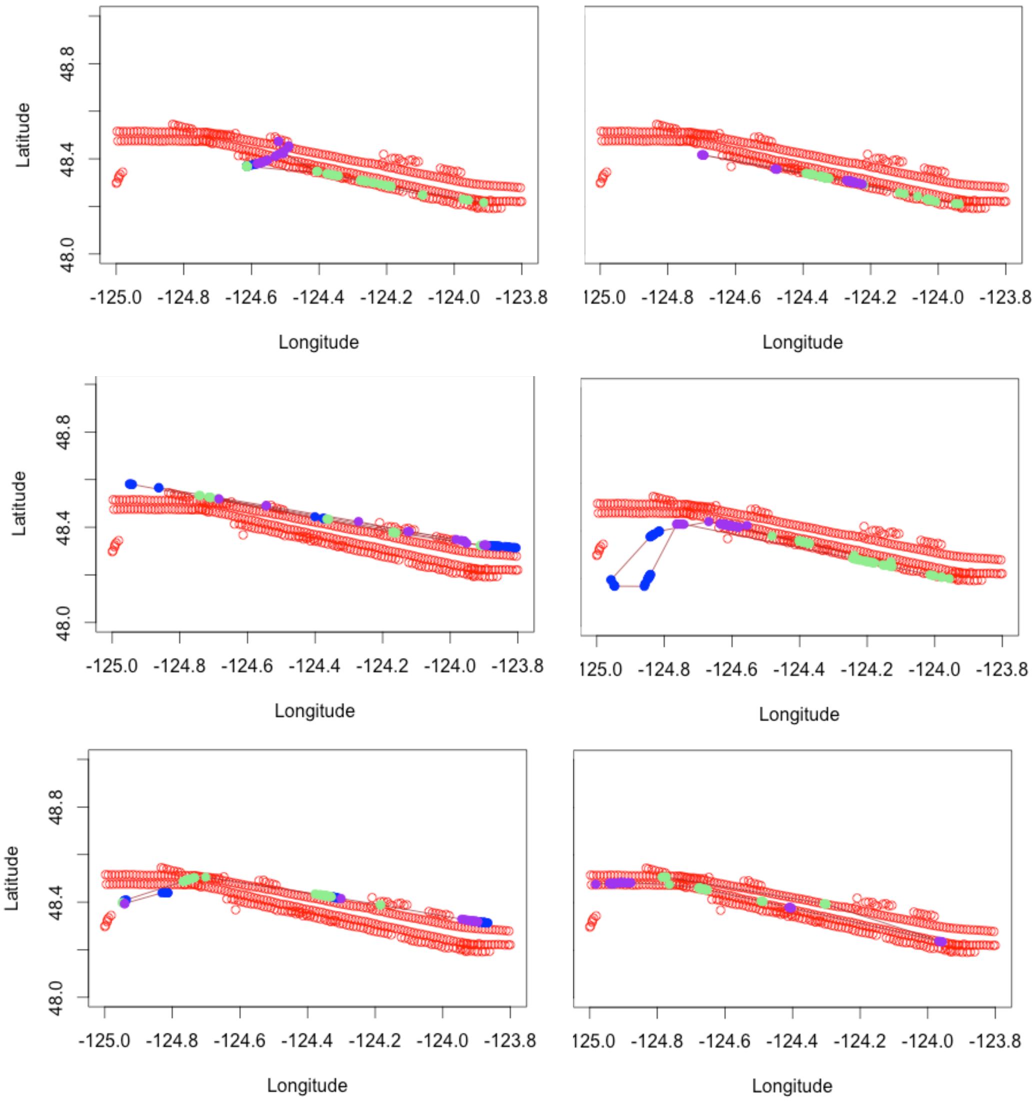
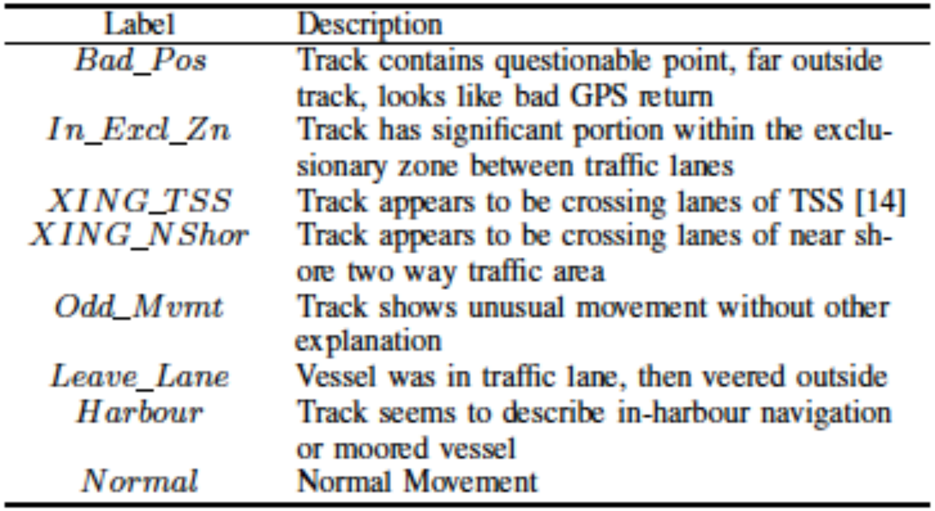


Fig.10. The anomaly labeling results. GVs and the SSP are in red and the normal points are in green. The two types of abnormal points are in blue (abnormal in relation to ADD or RDD) and purple (abnormal in relation to CDD).



Experiment On Labeled Data

- The Expert’s Labeling Process
  - **The same AIS data set** is labeled by expert
    - First, the points are divided into distinct tracks based on MMSIs
    - Then, temporally sequential points are separated if there is more than 4.5 minutes time gap between the two points
    - For the one-point track case, the length of the track is 0 nautical mile and the track is not assigned with any labels
  - After dividing **284 tracks (284 unique MMSIs), 2,122 sub-tracks** were generated **(imbalanced**)
    - **680** sub-tracks only contain **one point**
    - **14** sub-tracks are classified as **abnormal**
    - **1,428** sub-tracks are classified as **normal**

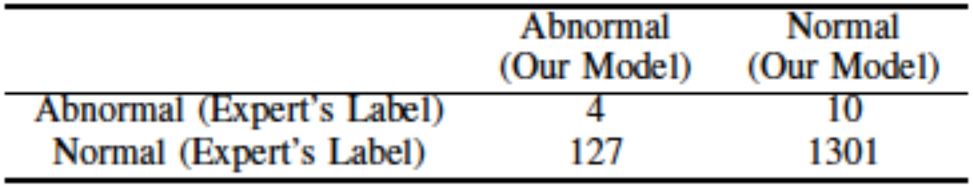


Label	Description
Bad_Pos	Track contains questionable point, far outside track, looks like bad GPS return
In_Excl_Zn	Track has significant portion within the exclusionary zone between traffic lanes
XING_TSS	Track appears to be crossing lanes of TSS [14]
XING_NShor	Track appears to be crossing lanes of near shore two way traffic area
Odd_Mvmt	Track shows unusual movement without other explanation
Leave_Lane	Vessel was in traffic lane, then veered outside
Harbour	Track seems to describe in-harbour navigation or moored vessel
Normal	Normal Movement

Table II The labels and their descriptions provided by the expert

Experiment On Labeled Data

- Comparison between our model’s labels with the expert’s
    - The expert’s division method is firstly applied to separate the tracks
  - Experiment A
    - All abnormal points caused by ADD, RDD or CDD count
    - We employ 60% as the anomaly ratio threshold
      - E.g. when the track’s anomaly ratio is not less than 60%, it will be labeled as abnormal



	Abnormal (Our Model)	Normal (Our Model)
Abnormal (Expert’s Label)	4	10
Normal (Expert’s Label)	127	1301

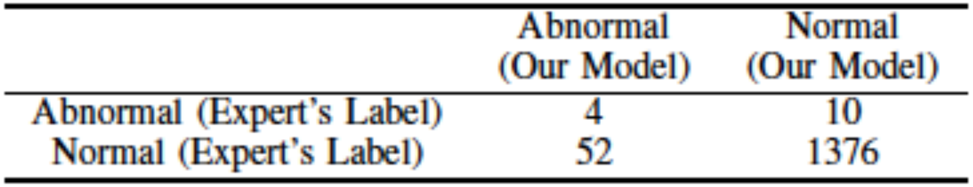
Table III Confusion Matrix I

The result’s overall accuracy is 90.49%

The recall for abnormal class is 28.57%

Experiment On Labeled Data

- Comparison between our model's labels with the expert's
  - Experiment B
    - Only abnormal points caused by CDD count
    - We employ 10% as the anomaly ratio threshold
      - Because abnormal points caused by ADD or RDD no longer count, a lower threshold is needed.



	Abnormal (Our Model)	Normal (Our Model)
Abnormal (Expert's Label)	4	10
Normal (Expert's Label)	52	1376

Table III Confusion Matrix after only considering CDD

The result's overall accuracy is 95.70% (90.49% in the previous)

The recall for abnormal class is 28.57% (same as the previous)

Discussion & Future Work

1. To consider **speed** (COG) during the expert's labeling process
   - This will also help **reduce the false alarm rate**.
2. More experiments are to be done in **other near-port regions**
   - e.g. Los Angeles Long Beach Port area
3. Other **classification algorithms** which are designed for **imbalanced data** are to be tried
   - The proposed **division distances (ADD, RDD and CDD)** can be used as the models' **features**.
4. An **ensemble model** incorporates **predictive models** and **our own framework** is to be developed.
   - Predictive models usually predict future status information of a vessel and then compare the real data with the prediction to decide the abnormality.

References

[1] H.Ball, “Satellite AIS for Dummies.” Mississauga, ON: Wiley, 2013

[2] B. Liu, E. N. de Souza, S. Matwin, and M. Sydow, “Knowledge-based clustering of ship trajectories using density-based approach,” in Big Data (Big Data), 2014 IEEE International Conference on. IEEE, 2014, pp. 603–608.

[3] S. Spaccapietra, C. Parent, M. L. Damiani, J. A. de Macedo, F. Porto, and C. Vangenot, “ A conceptual view on trajectories,”” Data Knowl. Eng., vol. 65, no. 1, pp. 126-146, Apr. 2008

[4] M. Ester, H. peter Kriegel, J. S, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise.” AAAI Press, 1996, pp. 226–231.

[5] B. T. Morris and M. M. Trivedi, “A survey of vision-based trajectory learning and analysis for surveillance.”” IEEE trans. Circuits Syst. Video Techn., no. 8, pp. 1114-1127

Questions ?

Thank you for your attention.

Bo Liu
Dalhousie University
boliu@dal.ca

July 8th, 2015

DALHOUSIE
UNIVERSITY
INSTITUTE FOR
BIG DATA ANALYTICS