J. Ocean Univ. China (Oceanic and Coastal Sea Research)
https://doi.org/10.1007/s11802-019-3717-9
ISSN 1672-5182, 2019 18 (2): 403-412
http://www.ouc.edu.cn/xbywb/
E-mail:xbywb@ouc.edu.cn

FVID: Fishing Vessel Type Identification Based on VMS Trajectories

HUANG Haiguang[1), 2)], HONG Feng[1), *], LIU Jing[1)], LIU Chao[1)], FENG Yuan[1)], and GUO Zhongwen[1)]

1) *College of Information Science and Engineering*, *Ocean University of China*, *Qingdao* 266100, *China*
2) *Wenzhou Ocean and Fishery Vessel Safety Rescue Information Center*, *Wenzhou* 325000, *China*

(Received October 27, 2017; revised December 27, 2017; accepted February 23, 2018)

Abstract Vessel Monitoring System (VMS) provides a new opportunity for quantified fishing research. Many approaches have been proposed to recognize fishing activities with VMS trajectories based on the types of fishing vessels. However, one research problem is still calling for solutions, how to identify the fishing vessel type based on only VMS trajectories. This problem is important because it requires the fishing vessel type as a preliminary to recognize fishing activities from VMS trajectories. This paper proposes fishing vessel type identification scheme (FVID) based only on VMS trajectories. FVID exploits feature engineering and machine learning schemes of XGBoost as its two key blocks and classifies fishing vessels into nine types. The dataset contains all the fishing vessel trajectories in the East China Sea in March 2017, including 10031 pre-registered fishing vessels and 1350 unregistered vessels of unknown types. In order to verify type identification accuracy, we first conduct a 4-fold cross-validation on the trajectories of registered fishing vessels. The classification accuracy is 95.42%. We then apply FVID to the unregistered fishing vessels to identify their types. After classifying the unregistered fishing vessel types, their fishing activities are further recognized based upon their types. At last, we calculate and compare the fishing density distribution in the East China Sea before and after applying the unregistered fishing vessels, confirming the importance of type identification of unregistered fishing vessels.

Key words VMS; vessel type identification; fishing density; trajectory analysis; classification

1 Introduction

The original target of Vessel Monitoring System (VMS) is to enforce sailing security. It records vessel sailing information, including ID, timestamp, location, instant speed and heading, *etc*. When deployed satellite devices on fishing vessels, VMS generates a large amount of trajectory data. Meanwhile, with the improvement on the trajectory data processing (Wang *et al*., 2012; Wang and Dai, 2017), these VMS data provide a new opportunity for quantified fishing research.

Previous fishing research on VMS trajectory has two phases. The first is to recognize the fishing segments from all VMS trajectories. Different classification methods are exploited in this step, including thresholds on speed and heading (Deng *et al*., 2005; Lee, 2010), statistical inference (Vermard *et al*., 2010; Walker and Bez, 2010), machine learning (Joo *et al*., 2011; Russo *et al*., 2011; Perera *et al*., 2012; Joo *et al*., 2015; Kim and Jeong, 2015; Russo *et al*., 2016) and image processing (Zong *et al*., 2016). The second calculates fishing related metrics, covering fishing density (Witt and Godley, 2007; Bertrand *et al*., 2008; Castro *et al*., 2010; Chang *et al*., 2010; Chang, 2011; Lambert *et al*., 2012; Murray *et al*., 2013; Wang *et al*., 2015), and fishing efforts (Bastardie *et al*., 2010; Coro *et al*., 2013; Zong *et al*., 2016; Russo *et al*., 2016).

Previous research often depends on the knowledge of fishing vessel type as the preliminary, because the fishing activity recognition requires fishing vessel types to apply different conditions of vessel speed, direction, *etc*. However, the fishing vessel types are often unknown in practice, for not all vessels have registered to the fishing management bureaus related to the certain fishing area. For example, Zhejiang Ocean and Fishery Bureau records 10031 active fishing vessels in the East China Sea under its surveillance in March 2017. Meanwhile, there are 1350 unregistered fishing vessels appearing in the same region, which requires identifying their types before further analysis on fishing metrics likes density or efforts. Therefore, the research question here is how to identify fishing vessel type based on VMS trajectories. In Campanis (2008) and Coro *et al*. (2013), the proposed schemes try to calculate the fishing activities from unregistered fishing vessels. These schemes exploit speed, direction and bathymetry information directly to build rule-based classifiers, which may be error-prone for different types of fishing vessels.

* Corresponding author. E-mail: hongfeng@ouc.edu.cn

There are two major challenges in the vessel type identification problem. Firstly, the fishing vessels among different types are often of similar tonnage and engine, so their sailing and fishing activities may not be much different. Secondly, the data of the VMS traces are huge, calling for an effective computing method. In this paper, we propose a fishing vessel type identification scheme (FVID) based on only VMS trajectories. FVID exploits feature engineering and machine learning scheme of XGBoost (Chen and Guestrin, 2016) as its two key blocks. FVID first leverages feature engineering to extract the features from VMS trajectories, representing the differences among various fishing vessel types. As a feature vector indicates one fishing vessel's specialty in trajectory, it decreases the input size for the next block. XGBoost classifiers exploit the feature vectors from registered fishing vessels in the training phase and then identify the unregistered fishing vessel type. To solve the imbalance between the numbers among different fishing vessel types, we further apply SMOTE over-sampling method (Chawla *et al*., 2011) before classification.

The VMS dataset is recorded by Zhejiang Ocean and Fishery Bureau in March 2017, which contains a record number of 49236165 and 13.6 GB storage in total. Especially, each vessel has 4326 records on average with a maximum of 12051 and a minimum of 2032 records. Besides, these trajectories are recorded with China Beidou Satellite system, giving them a 5-minute resolution. The registered fishing vessels in this area have nine types, including shrimp trawl, otter trawl, pair trawl, gill net, canvas stow net, crab cage, square net, light seine, and transportation. Hence, the task of fishing vessel identification is to classify an unregistered vessel to one of these nine types.

To verify the classification accuracy, we first apply a 4-fold cross-validation on FVID with the trajectories of registered fishing vessels. The classification accuracy is 95.42%, confirming the performance of FVID. We then apply FVID to the unregistered fishing vessels for type classification. After the unregistered fishing vessels are identified with their types, we recognize their fishing activities. At last, we calculate the changes in fishing density distribution after applying the trajectories of both registered and unregistered fishing vessels, in order to confirm the importance of unregistered vessel type identification.

2 Dataset and Methods

This section first briefly introduces our VMS dataset that brings up the problem of fishing vessel type identification in practice. Then we describe the proposed scheme of fishing vessel identification on types (FVID) in details.

2.1 Dataset

The dataset is the VMS trajectories of active fishing vessels recorded by Zhejiang Oceanic and Fishery Bureau, China. The trajectory data contains the time, position, sailing speed and direction, *etc*. of the fishing vessels in the East China Sea in March 2017. The recorded trajectory had a temporal resolution of 5 min and the dataset has 49236165 records. There are active 10031 registered 'fishing vessels' activities recorded in this month. Moreover, the dataset contains the trajectories from 1350 unregistered fishing vessels.

The registered fishing vessels have nine types according to their fishing catches or methods, including shrimp trawl, otter trawl, pair trawl, gill net, canvas stow net, crab cage, square net, light seine, and transportation. The vessel number for each type are shown in Table 1.

Table 1 Types and numbers of registered fishing vessels

Type	Description	Quantity
T_1	Shrimp trawl	2761
T_2	Otter trawl	1581
T_3	Pair trawl	1364
T_4	Gill net	2093
T_5	Canvas stow net	727
T_6	Crab cage	292
T_7	Square net	271
T_8	Light seine	135
T_9	Transportation	807

For 1350 unregistered vessels, however, the bureau has no clues on their vessel types. It leads to that the bureau cannot distinguish their sailing or fishing behaviors from their VMS trajectories, which relies on the vessel type as a preliminary. The research problem here is to identify the types for 1350 unregistered fishing vessels based only on their VMS trajectories.

2.2 Methods

Because the registered fishing vessels have type labels, we can apply the classifiers of supervised machine learning to identify the unregistered vessel type. FVID composes of five steps, which are preprocessing, feature enumeration, feature selection, classifier training, and recognition, as shown in Fig.1.

Fig.1 Sketch of FVID.

The first step of preprocessing is to deal with the problems of value missing and calculate the speed and direction of the vessels based on their tracked positions. The second step extracts the features, representing the diffe-

rences in the trajectories among the various types of fishing vessels. We extract the features on the timestamp, speed, direction, and zones from trajectory analysis, and apply feature selection to find the best feature vector for classification. FVID exploits the XGBoost classifier for its accuracy and lightweight cost. As shown in Table 1, the distribution of different fishing types is imbalanced. We adopt the Over-sampling method (Chawla *et al*., 2011) on the types with small vessel number. After training the classifiers, we use them to recognize the types of unregistered fishing vessels. We illustrate each step of our method in details in the following.

1) Preprocessing: Due to transmission errors, data missing is very common. Once there is a null value on the items of GPS timestamp, latitude or longitude, we drop the record. Meanwhile, the recorded speed and direction are the instantaneous values observed by Beidou terminal devices, which exhibit high fluctuations due to vessel shaking and waves. We calculate the average direction and speed between two records by (Williams, 2011), which are more stable than the instantaneous ones.

2) Feature enumeration: This section describes the feature enumeration phase. There are 61 features enumerated to represent the difference among various fishing vessel types.

For each VMS record, we choose timestamp, longitude, latitude, direction, and speed as the raw fields for feature numeration after preprocessing. Pointed out by previous research (Deng *et al*., 2005; Lee, 2010), the speed of fishing activities is different from sailing speed for many types of fishing vessels. Therefore, we count the speed distribution for each type of registered vessels in our dataset.

Fig.2 shows the distributions on the speed of the fishing vessels of all nine types. The whole speed section for all types of fishing vessels is from zero to twelve. For different types of fishing vessels, there is some difference in the dense area of speed distribution. For example, the speed distributions of both Shrimp Trawl and Otter Trawl have only one clear dense section of (2.5, 5.5), while Gill nets have two dense sections of (0.5, 2.5) and (5.5, 12). When counting the dense section for each kind, three splitting lines are found and labeled on Fig.2, which are the speed of 0.5, 2.5 and 5.5. Depending on these lines, the whole speed section is divided into four sections of (0, 0.5), (0.5, 2.5), (2.5, 5.5) and (5.5, 12).

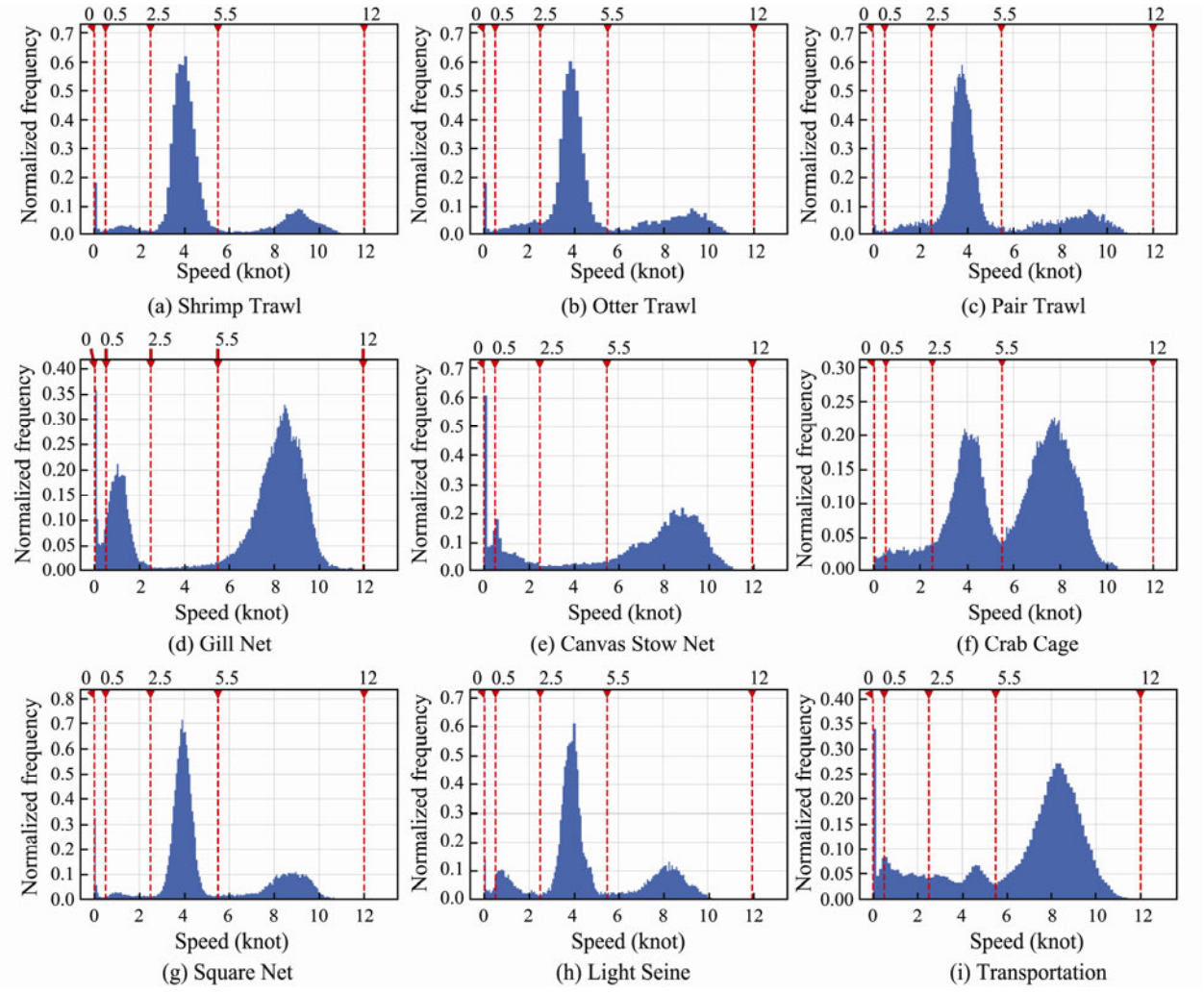


Fig.2 The speed distributions of nine fishing vessel types.

The nine types of fishing vessels have some similar or different distributions according to these four speed sections. This provides the clues to distinguish nine types into at least four groups. The first contains only one type

of fishing vessels of Gill Nets, which have two dense sections of (0.5, 2.5) and (5.5, 12). The second has only Crab Cages with two dense sections of (2.5, 5.5) and (5.5, 12). The third consists of two types of fishing vessels, which are Canvas Stow Net and Transportation. Their dense sections lie in speed (5.5, 12). The last group composes of five types, including Shrimp Trawl, Otter Trawl, Pair Trawl, Square Net and Light Seine. As shown in Fig.2, the distributions of these five types have only one typical dense section in (2.5, 5.5). These four groups reveal that it helps for classification when taking the distribution of four speed sections as the features. As a result, We quantify the first four features as the histogram of four speed sections on speed values for each vessel.

When further comparing the distributions of different types inside the last group, Fig.2 shows the density distributions in speed (2.5, 5.5) still have some difference. We calculate the mean, median, and standard deviation on speed section (0, 0.5), (0.5, 2.5) and (2.5, 5.5) for each vessel to represent such kind of difference. These features form nine features, labeled as f_5–f_{13}. The relative values on speed section (5.5, 12) are excluded because this section represents the vessel in the stage of sailing, which contains almost no differences for all types.

In (Coro *et al.*, 2013), the research pointed out that most fishery vessels are performing fishing behaviors around speed section of (0, 5.5). We depict the positon distribution of each type of fishing vessels in this speed section in Fig.3. It shows the density map of fishing position frequency. There is some difference in the spatial distributions, so we extract the features on longitude and latitude as follows: the mean, median on speed section (0, 0.5), (0.5, 2.5) and (2.5, 5.5). These accounts for feature f_{14}–f_{25}.

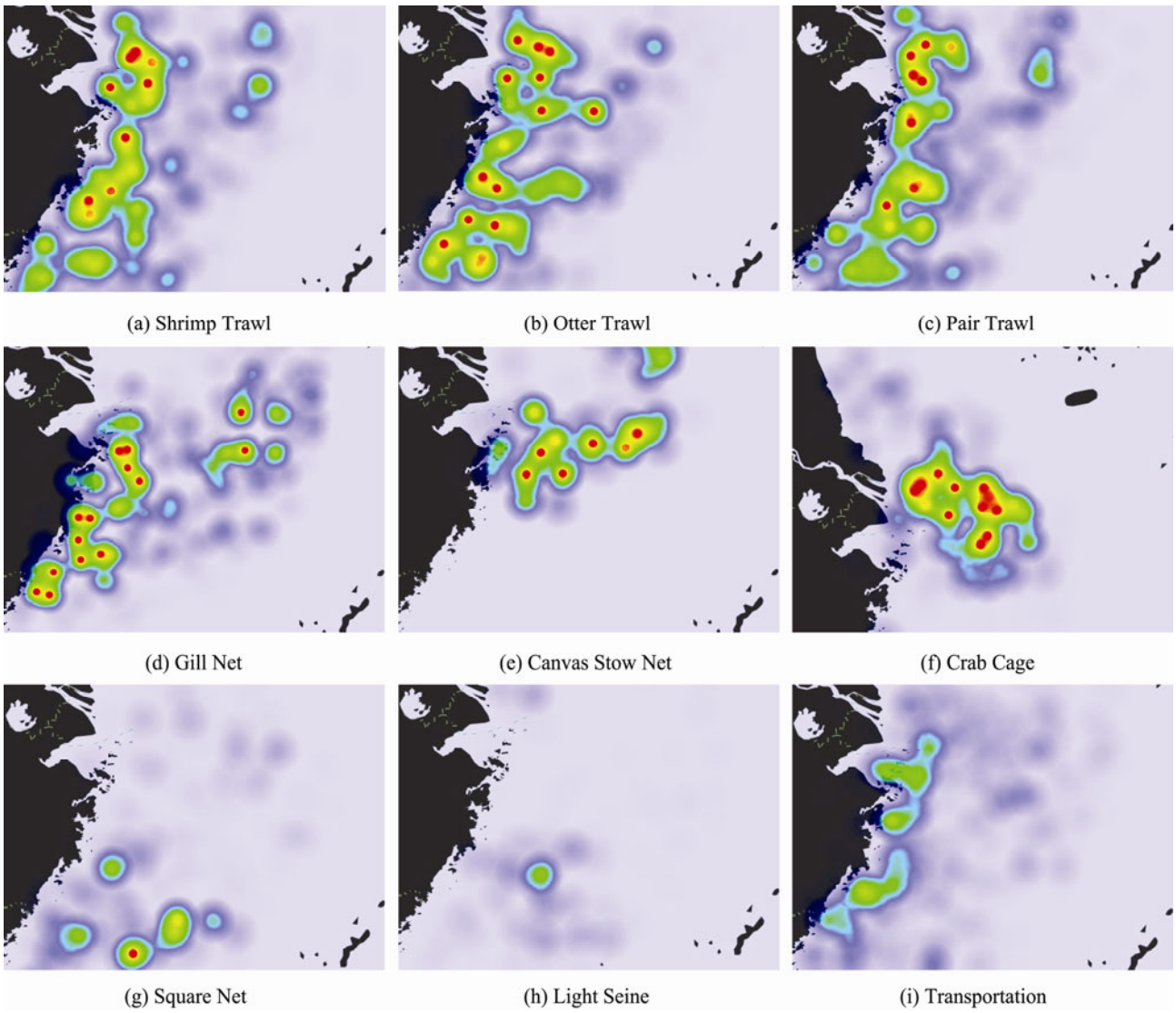


Fig.3 Spatial frequency distributions of fishing activities for each vessel type.

For the field of directions, the variation exists when vessels are fishing. This is because fishing vessels of difference types exploit different methods to catch fishes. For example, Otter Trawls drag their nets with frequent turns to catch fishes. These kinds of activities bring in the fast direction changing. On the contrary, Gill Nets put down their nets in one place, sail off, and return to pull out the nets. Thus, their single typical fishing trajectory sails along one path back and forth with only one turn. There is no clear direction variation except the only turn.

Therefore, we use the standard deviation of direction on speed section (0, 0.5), (0.5, 2.5) and (2.5, 5.5) as feature f_{26}-f_{28}.

The fishing periods also vary to some extent for different types of fishing vessels. Especially for light seines, their fishing activities happen from 4 pm to 6 am in the next day, when the sky is dark. We design the features to capture the proportion of speed section (2.5, 5.5) taking up in every hour, which accounts for feature f_{29}–f_{52}. Speed section of (2.5, 5.5) is the typical dense section for Light Seine, Shrimp Trawl, Otter Trawl, Pair Trawl, and Square Net. For example, f_{29} represents the proportion of fishing records in speed section (2.5, 5.5) takes up from 0:00 to 1:00 for a vessel. We also define feature f_{53}-f_{61} as the fishing proportion in period of (0:00–5:00), (5:00–7:00), (5:00–8:00), (6:00–11:00), (6:00–17:00), (12:00, 14:00), (12:00,17:00), (18:00–23:00), and (18:00–5:00) respectively.

Summing all the features up, we extract 61 features from the raw VMS data. All these features are related to speed section (0, 0.5), (0.5, 2.5), (2.5, 5.5) and (5.5, 12) respectively. f_1-f_4 indicate histogram on speed sections for (0, 0.5), (0.5, 2.5), (2.5, 5.5) and (5.5, 12). f_5-f_{13} label the mean, median, standard deviation on speed section (0, 0.5), (0.5, 2.5) and (2.5, 5.5). The mean, median on longitude and latitude construct the spatial features f_{14}-f_{25} of speed sections of (0, 0.5), (0.5, 2.5) and (2.5, 5.5), respectively. f_{26}–f_{28} represent the standard deviation of direction on speed section of (0, 0.5), (0.5, 2.5) and (2.5, 5.5) respectively. Timestamp feature f_{29}–f_{52} calculate the proportion of the speed section taking up in every hour. Feature f_{53}–f_{61} as the fishing proportion in some special periods.

3) Classifier and over-sampling: Because the feature selection method is of the Wrapper kind, we illustrate the classifier in this subsection and then describe our feature selection in the next subsection. With the VMS trajectories of registered fishing vessels of known types, the problem to identify the unregistered fishing vessel type can be defined as supervised multi-class classification. As our dataset is of huge amounts, it is important to choose a classifier, which is capable of parallel execution. Therefore, we choose XGBoost (Chen and Guestrin, 2016) as the classifier. Moreover, we take the One-Versus-Rest (OVR) (Boutell *et al*., 2004) method for multi-class classification problem *i.e*. nine classifiers are trained for type T_1–T_9. For an unregistered vessel, each classifier produces the probability that the instance belongs to the corresponding type. Then the instance is identified as the type with the maximum probability. The hyper-parameters of XGBoost are tuned *via* cross-validation.

As shown in Table 1, the vessel number of fishing type T_6, T_7 and T_8 are quite small comparing to another fishing type. It leads to that the samples for training are imbalanced among types. We use Synthetic Minority Oversampling Technique (SMOTE) (Chawla *et al*., 2011) to solve the problem. SMOTE is a procedure that utilizes oversampling techniques to deal with minority and synthesizes new sample.

Applying SMOTE, 9739, 9760 and 9896 positive samples are oversampled for type T_6, T_7 and T_8 respectively. We will compare the classification accuracy before and after oversampling in the evaluation section.

4) Feature selection: In the previous feature enumeration, we construct features of speed, direction, space, timestamp features. There are 61 features totally. However, not all features will contribute to the classification; perhaps some features contradict each other or have minor contributions to the classifier (Feng and Lang, 2017). We apply feature selection to evaluate the contribution of every feature to classification and find the best feature vector for each vessel type classifier.

First, we remove features with low variance through the Variance Threshold method (Clarkson *et al*., 1994). Three features of the timestamp feature of f_{54}, f_{55} and f_{58} are removed, which concerns the proportion of fishing taking up in some periods.

Second, we use the model-based ranking method. XGBoost (Chen and Guestrin, 2016) provides a feature importance function to evaluate the features, where importance indicates the appearance of a feature in constructing classification trees. Label

$$I_j = \{i \mid q(x_i) = j\},$$

as the instance set of leaf j. Assume that I_L and I_R are the instance sets of left and right nodes after the split. Letting $I = I_L \cup I_R$, the loss reduction after the split is defined as Eq. (1).

$$L_{split} = Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma . \tag{1}$$

where $G = \sum_{i\in I} g_i$, $H = \sum_{i\in I} h_i$, g_i and h_i are first and second order gradient statistics on the loss function, $\frac{G_L^2}{H_L+\lambda}$ is the score on the new left leaf, $\frac{G_R^2}{H_R+\lambda}$ is the score on the new right leaf, $\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}$ is the score on the original leaf, γ is regularization on the additional leaf. Gain records the number of times every feature causes the classification tree to split. Therefore, it is an indicator of the contribution of a feature on classification. We then sort all the features in descending order according to their Gain values.

With the feature order for each vessel type classifier, we apply the Wrap method to select the best feature vector *i.e*. we add the features to XGBoost one by one according to the Gain order and use the dataset of registered vessels to evaluate the classification accuracy. Fig.4 plots the classification results for Otter Trawl when adding the features one by one. It shows that this classifier achieves the highest accuracy with 23 features. Therefore, we use the feature vector of 23 features in XGBoost for Otter Trawl. Similarly, the feature vectors for the other eight types can be calculated. Table 2 summarized the feature vectors for every vessel type classifier.

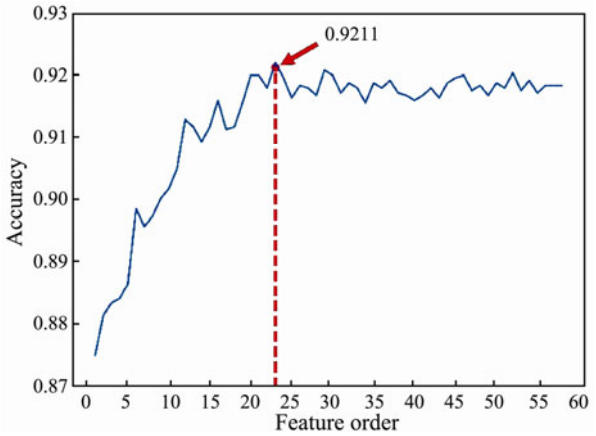


Fig.4 Feature selection of Otter Trawl.

Table 2 shows that feature f_1–f_4 are the top ten important feature for all nine vectors. This illustrates that speed related features are most important for classification, which was also confirmed by previous research (Campanis, 2008; Coro *et al*., 2013). However, the way on exploiting speed related features are different between FVID and previous research. Previous approaches only exploit average and median values of speed, while FVID exploits the speed histogram feature of f_1–f_4. Because feature f_1–f_4 are almost among the top five most important features for each classifier in Table 2, the proposed histogram feature plays an important rule for vessel type classification.

The intersection of all nine feature vectors is feature f_1–f_{12}, f_{14}–f_{16}, f_{26}–f_{28} and f_{34}. Feature f_1–f_{12} are speed relative features. Feature f_{14}–f_{15} is the mean of longitude and latitude on speed section (0.5, 2.5). Feature f_{16} is the median of longitude on speed section (0.5, 2.5). Feature f_{26}–f_{28} indicate the standard deviations of direction on speed section (0, 0.5), (0.5, 2.5) and (2.5, 5.5). These features confirm the importance of the sailing speed and direction to the classification of vessel types. Feature f_{34} labels the fishing proportion in period of (05:00–06:00), which represents the uniqueness of certain type of vessels fishing in the early morning.

Table 2 Selected features for every fishing vessel type

Type	Quantity	Selected features sorted in descending order
T_1	42	$f_2, f_1, f_3, f_4, f_9, f_{10}, f_{13}, f_5, f_9, f_{11}, f_{27}, f_{26}, f_7, f_{15}, f_{28}, f_{12}, f_6, f_{14}, f_{19}, f_{46}, f_{18}, f_{16}, f_{34}, f_{47}, f_{39}, f_{35}, f_{36}, f_{17}, f_{33}, f_{44}, f_{45}, f_{31}, f_{40}, f_{29}, f_{37}, f_{32}, f_{48}, f_{38}, f_{41}, f_{30}$
T_2	23	$f_1, f_3, f_2, f_4, f_5, f_9, f_{13}, f_{11}, f_{10}, f_{26}, f_{27}, f_9, f_{18}, f_7, f_{14}, f_{28}, f_{12}, f_6, f_{15}, f_{16}, f_{34}, f_{47}, f_{17}$
T_3	25	$f_3, f_1, f_2, f_4, f_{10}, f_8, f_{13}, f_{11}, f_9, f_5, f_{26}, f_{27}, f_7, f_{28}, f_{15}, f_{14}, f_{12}, f_{16}, f_6, f_{34}, f_{18}, f_{35}, f_{19}, f_{47}, f_{36}$
T_4	43	$f_1, f_2, f_3, f_4, f_9, f_{10}, f_5, f_{27}, f_{13}, f_{11}, f_{26}, f_7, f_6, f_9, f_{15}, f_{29}, f_{14}, f_{12}, f_{33}, f_{34}, f_{19}, f_{18}, f_{35}, f_{16}, f_{36}, f_{46}, f_{33}, f_{37}, f_{32}, f_{38}, f_{39}, f_{17}, f_{47}, f_{40}, f_{41}, f_{29}, f_{45}, f_{44}, f_{60}, f_{53}, f_{42}, f_{43}, f_{56}, f_{21}$
T_5	28	$f_1, f_3, f_{10}, f_2, f_4, f_{15}, f_9, f_{14}, f_5, f_9, f_{16}, f_{19}, f_{11}, f_{45}, f_{36}, f_6, f_{17}, f_{26}, f_{19}, f_{35}, f_{20}, f_7, f_{44}, f_{42}, f_{12}, f_{28}, f_{47}, f_{56}$
T_6	46	$f_3, f_1, f_{15}, f_{13}, f_{12}, f_{34}, f_2, f_{11}, f_4, f_{17}, f_{29}, f_{10}, f_{42}, f_9, f_{27}, f_5, f_6, f_{35}, f_{39}, f_{36}, f_{56}, f_{16}, f_{51}, f_{18}, f_{19}, f_{37}, f_{33}, f_{52}, f_{49}, f_9, f_{47}, f_{46}, f_7, f_{25}, f_{57}, f_{14}, f_{23}, f_{21}, f_{45}, f_{20}, f_{43}, f_{60}, f_{41}, f_{40}, f_{26}, f_{44}$
T_7	48	$f_1, f_2, f_3, f_8, f_4, f_{14}, f_{15}, f_{11}, f_{10}, f_{18}, f_{16}, f_{17}, f_9, f_{20}, f_{19}, f_{13}, f_7, f_5, f_{28}, f_6, f_{27}, f_{12}, f_{26}, f_{37}, f_{21}, f_{35}, f_{46}, f_{52}, f_{34}, f_{43}, f_{36}, f_{39}, f_{29}, f_{24}, f_{47}, f_{44}, f_{41}, f_{33}, f_{39}, f_{42}, f_{22}, f_{40}, f_{49}, f_{30}, f_{48}, f_{45}, f_{31}, f_{59}$
T_8	48	$f_1, f_2, f_4, f_{14}, f_3, f_{15}, f_{10}, f_{11}, f_{27}, f_{17}, f_{34}, f_{13}, f_6, f_5, f_{16}, f_{12}, f_{21}, f_{19}, f_{33}, f_9, f_7, f_{26}, f_9, f_{18}, f_{37}, f_{20}, f_{45}, f_{28}, f_{35}, f_{34}, f_{47}, f_{41}, f_{51}, f_{39}, f_{40}, f_{23}, f_{24}, f_{44}, f_{36}, f_{31}, f_{52}, f_{25}, f_{22}, f_{39}, f_{42}, f_{46}, f_{29}, f_{30}, f_{32}$
T_9	20	$f_2, f_1, f_4, f_3, f_{10}, f_5, f_{13}, f_9, f_{11}, f_{26}, f_7, f_{27}, f_{14}, f_9, f_6, f_{28}, f_{12}, f_{15}, f_{34}, f_{16}$

The union of all feature vectors covers all features except f_{50} and f_{61}. Feature f_{50} and f_{61} represents the fishing proportion in period of (21:00–22:00) and (18:00–5:00). The first is of little importance for the reason that most fishing vessels may do not fish in the evening. The second is deleted because other features may have covered its specialty on vessel types. All the features on the proportion of period locate at the ends of all the feature vectors. The union result indicates that the features on the proportion of periods are less important but still play some roles for classification.

In general, FVID calculates 56 features for every vessel from its VMS trajectory for further type classification. Nine feature vectors are constructed with the different combinations of 56 features.

5) Type identification: After feature enumeration and selection, FVID train the nine type classifiers with all the dataset on registered vessels. Then FVID can identify the vessel type when the trajectory of an unregistered vessel is fed in. We will evaluate the classification performance in next section.

3 Result

This section validates the performance of FVID. To validate the classification accuracy, we first perform testing on the dataset of registered fishing vessels, which can reveal the classification accuracy with ground truth. We then apply our recognition scheme on unregistered vessels. After type recognition, we segment the fishing activities of unregistered vessels based on their types and compare the fishing density distribution with the results only from the trajectories of registered vessels.

Our dataset is stored on an Oracle RAC cluster with two nodes, each of which is equipped with Intel Xeon(R) CPU E5-2640 v3 (8 kernels) and 64GB memory. For every vessel, nine feature vectors are constructed for each vessel type classifier. As shown in Table 2, 56 features are

calculated once and all the feature vectors can be created. We apply SQL statements on Oracle database to create the feature vectors. We use a computing node for the classifiers with Intel Xeon(R) CPU E5-2640 v3 (8 kernels) and 32GB memory. All the nine classifiers are implemented with Python on this computing node.

3.1 Classification Accuracy

We perform 4-fold cross-validation based on the dataset of registered vessels for ten times. For each time of validation, the dataset of registered vessels is randomly split into one training set (75%) and one testing set (25%). The classifiers for nine types are trained with the recognize training set. Then we apply the nine classifiers to nize the vessel types for the test set, comparing the recognition results with ground truth. The accuracy results are the average over ten times.

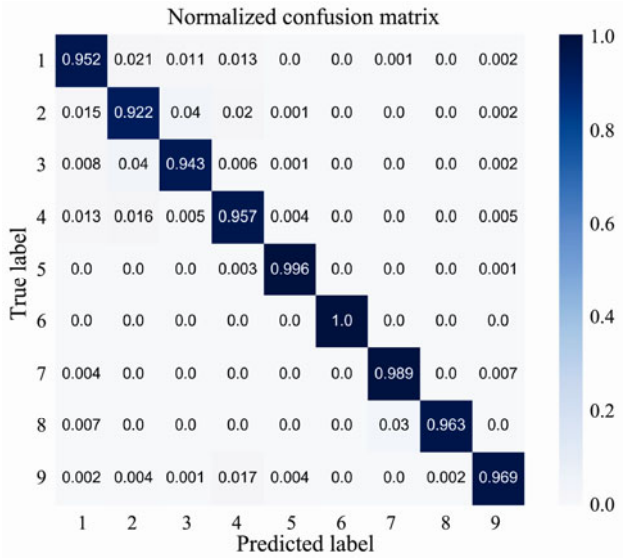


Fig.5 Normalized confusion matrix of classification for Shrimp Trawl, Otter Trawl, Pair Trawl, Gill Net, Canvas Stow Net, Crab Cage, Square Net, Light Seine, and Transportation labeled from 1 to 9.

Fig.5 shows the confusion matrix of classification. The lowest classification accuracy exists for Otter Trawler (T_2), which is 92.2%. There are 395 otter trawlers in testing, among which 7 are misclassified as Shrimp Trawl (T_1), 16 misclassified as Pair Trawl (T_3), and 8 misclassified as Gill Net (T_4). Comparing the classification error for the first four types, there is over 1% error on all the cells among four types. This indicates that the four types are sharing similar fishing behaviors to some degree *e.g.* they have similar speed distribution as discussed above.

Except for the first four types, the classification accuracies for the other types are over 96%. There is an interesting classification error on the cell that the transportations are misclassified as Gill Nets. The error is 1% with 5 transportations misclassified. As the transportation vessels do not have the fishing behaviors, we are confused why the errors happen. We trace back to the records of these vessels and visualize the trajectory with speed section (0.5, 2.5) and (2.5, 5.5) of this transportation vessel is shown in Fig.6(a). A typical trajectory of Gill Net with speed section (0.5, 2.5) and (2.5, 5.5) is depicted in Fig.6(b). Gill Nets have a typical way of fishing, which put down their nets in some place, sail off, and returning along the path back to pull out the nets. Comparing these two figures, we find that the trajectories in Fig.6(a) contains the pattern of this kind of fishing activities. The typical trajectory of transportation vessel is shown in Fig.6(c), the trajectory is significantly different from Figs.6(a) and (b). Consequently, the misclassification here is perhaps not an error, but a transportation deploys with some gill nets and tries to catch fishes illegally.

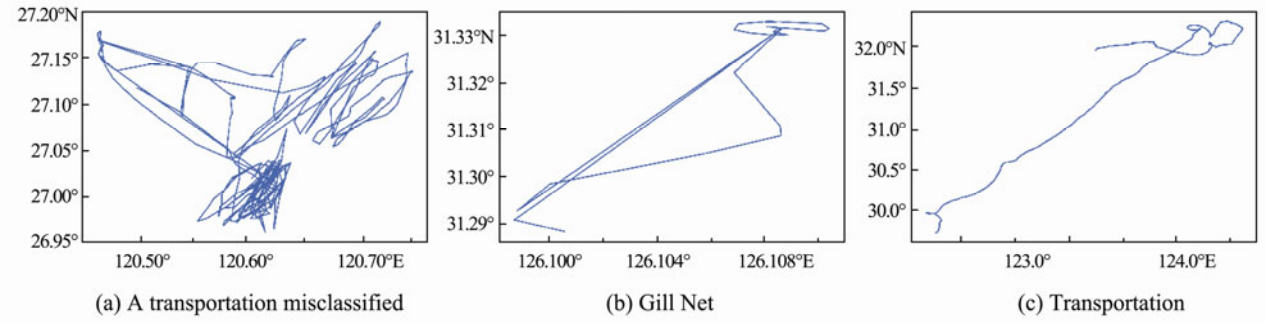


Fig.6 Typical trajectories of misclassified Transportation, Gill Net and normal Transportation.

This inspired us to build an online type identification system to track the registered transportations every day. If a transportation is misclassified as another type, a person will double-check the trajectory of the vessel and decide whether to send some administration ship to check onsite. In April, the online recognition system reported another case with a transportation classified as a Gill Net. This time a fishery administration ship was informed in time and caught the transportation with gill nets onsite.

To evaluate the choice of the classifier, we compare XGBoost with other typical classifiers, including the Logistic Regression (LR) (Shewhart and Wilks, 2005), Support Vector Machine (SVM) (Cortes and Vapnik, 1995), and k-Nearest Neighbor (kNN) (Cover and Hart, 1967). The results in Table 3 show that all the classifiers achieve high classification accuracy, while XGBoost has the best value of 95.42%. This also confirms that our feature vector is effective, which makes the classification

results little dependent on classifiers.

Considering time cost, XGBoost takes 148.16 s to train nine classifiers with all the train set of registered vessels, and takes 8.78 s to identify the types of 2508 testing vessels. LR takes up largest training time because there is a phase of generating polynomial and interactive features. kNN uses the largest recognizing time because every sample is compared with all clusters. Though XGBoost does not achieve the lowest time on both training and recognizing, the time cost of training is acceptable and the type identification time is qualified as quasi-real-time, especially for the requirement on determining whether Transportations conduct illegal fish catching.

Table 3 Comparison on classifiers of all models

Classifier	Accuracy	Train time (s)	Test time (s)
LR	0.893	941.12	4.81
SVM (RBF)	0.9145	68.61	62.57
kNN	0.9244	33.72	136.21
XGBoost	0.9542	148.16	8.78

3.2 Fishing Density

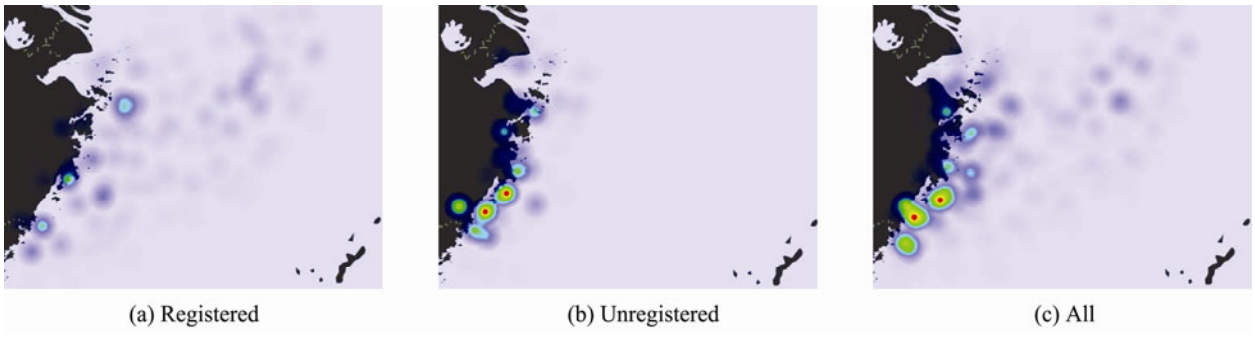
After evaluating the accuracy of FVID, we apply it to identify the unregistered 1350 fishing vessels in the East China Sea in March 2017. Based on the type identification, we further distinguish the sailing and fishing behaviors from the VMS trajectory for each unregistered fishing vessel. Then we calculate the accumulation of fishing time for each fishing vessel type with registered and unregistered vessels.

Table 4 shows the results of type identification on unregistered vessels and fishing time. It shows that the maximal number of unregistered vessels are Gill Nets, counting for near one-third of registered ones. The reason is that there is no regulation to limit the fishing region of Gill Nets in China. Nevertheless, this huge number still astonishes the province fishery bureau. When digging into fishing time comparison on the registered and unregistered Gill Nets, another important fact is revealed that the fishing hours of unregistered vessels take up to above 56% of the registered vessels *i.e.* each unregistered Gill Net conduct 170% fishing activities comparing to the registered Gill Net on average.

Fig.7 further shows the comparison of fishing density distribution for Gill Nets between registered and unregistered vessels. Fig.7(a) shows that the fishing activities of registered Gill Nets spread all over the East China Sea. There are no very highly dense fishing regions for the registered vessels. To the contrast, the unregistered fishing vessels concentrate on the south-west nearshore regions, as shown in Fig.7(b). When adding the fishing density of registered and unregistered Gill Nets together, Fig.7(c) shows that the dense regions are almost determined by the unregistered vessels, except that only one south-west region are the result of density addition. The fishery bureau may have a very different conclusion on fishery resource management for these three dense fishing regions, if the unregistered Gill Nets are not identified and counted with their fishing activities. The special case on Gill Nets calls for further coordination between the nearby province's fishery bureaus.

Table 4 Vessel Number and Fishing time statistics for registered and unregistered vessels of each type

Type	Description	Registered		Unregistered	
		Vessel number	Fishing hour	Vessel number	Fishing hour
T_1	Shrimp Trawl	2761	944492.87	298	46983.07
T_2	Otter Trawl	1581	450118.61	46	6857.35
T_3	Pair Trawl	1364	387447.41	105	13612.48
T_4	Gill Net	2093	211430.94	685	119718.72
T_5	Canvas Stow Net	727	60035.3	22	1898.6
T_6	Crab Cage	292	38637.67	28	2757.03
T_7	Square Net	271	34474.15	3	158.03
T_8	Light Seine	135	8852.62	3	32.5
T_9	Transportation	807		160	



(a) Registered (b) Unregistered (c) All

Fig.7 Fishing density comparison on Gill Net before and after adding unregistered vessels.

The second largest type in proportion to registered ones is Transportation, taking up to nearly 20%. All the transportations come from the southern province and most of them are in cooperation with the unregistered Gill Nets. Because the Transportations do not take fishing activities, they have little effect on fishery resource distributions.

Nevertheless, they may have an impact on the market price of fishery among provinces.

There are very few unregistered vessels of Canvas Stow Net, Crab Cage, Square Net and Light Seine, due to that those vessels do not have the capacity for remote operation. After tracking these unregistered vessels, we find that all such vessels sail off from the fishing ports of Zhejiang Province and take fishing activities on the border of the fishing regions between two provinces. After some background checks, we get that most vessels were once registered in Zhejiang province and transferred their registrations to the southern province of Fujian. That is why these vessels catch fishes in the related area.

For Shrimp Trawl, Otter Trawl and Pair Trawl, the unregistered vessels are of 10%, 3% and 7.3% of the numbers of registered vessels respectively. Meanwhile, their fishing activity time takes up of 5.0%, 1.5% and 3.5% of the ones of registered vessels respectively. These numbers show that their fishing time per vessel is nearly only half of the registered ones. When tracking their trajectories, the fishing regions of unregistered vessel spread out with the registered vessels. Therefore, the unregistered vessels of these three types have no important impact on the fishery resource management.

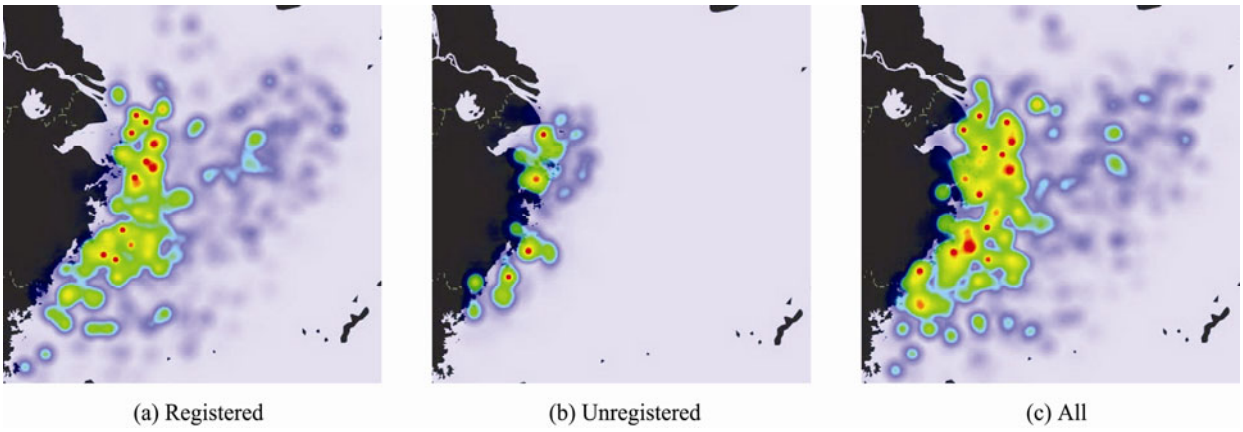


Fig.8 Comparison of fishing density in the East China Sea of registered vessels, unregistered vessels, and all vessels.

When adding the fishing density of all types together, we have the comparison before and after considering unregistered vessels, as shown in Fig.8. Fig.8(a) shows the dense fishing region of registered vessels is along the complete coastal line of the East China Sea, but outside the coastline for a certain distance. For unregistered vessels, their dense fishing regions are only of a small number of regions, which are close to the coastline. The closeness to the coastline reflects that the fishing vessels from other provinces may not have much experience on fishing in the long distance areas offshore in the East China Sea.

Fig.8(c) exhibits the results of adding Figs.8(a) and 8(b). Comparing Figs.8(a) with 8(c), there are two major differences. One lies in that the borders of the fishing regions are closer to the coastline as discussed above. The other difference is that the fishing density of south-west regions is denser in Fig.8(c) than in Fig.8(a). This is because many unregistered vessels conduct fishing in the south-west region, as illustrated in the example of Gill Nets. These differences imply the importance for the fishery bureau to identify the unregistered vessel type and consider their fishing activities.

4 Conclusions

In this paper, we proposed fishing vessel type identification scheme (FVID) based on VMS trajectories, which exploited feature engineering and machine learning scheme of XGBoost as its two key blocks and identified nine fishing vessel types. We apply FVID to the unregistered fishing vessels for type classification. Then, we recognized the unregistered fishing vessels' fishing activities according to their types, and calculated and compared the difference between fishing density distributions. The fishing density distribution difference between registered and unregistered vessels confirmed the importance on applying FVID. The fishery bureau now adopts FVID not only to recognize the types of unregistered fishing vessels and calculate the fishing density distribution, but also to provide clues on illegal fishing monitoring, especially on Transportation. These applications of FVID confirms its potentials for fishery resource research and management with unregistered fishing vessels by the VMS datasets.

Acknowledgements

We thank the Zhejiang Ocean and Fishery Bureau for providing VMS data. This research was partially supported by National Key R&D Program (No. 2016YFC 1401900), the National Natural Science Foundation of China (Nos. 61379127, 61379128, 61572448), the Fundamental Research Funds for the Central Universities (No. 201713016), and Qingdao National Laboratory for Marine Science and Technology Open Research Project (No. QNLM2016ORP 0405).

References

Bastardie, F., Nielsen, J. R., Ulrich, C., Egekvist, J., and Degel, H., 2010. Detailed mapping of fishing effort and landings by coupling fishing logbooks with satellite-recorded vessel geo-

location. *Fisheries Research*, **106** (1): 41-53.

Bertrand, S., Díaz, E., and Lengaigne, M., 2008. Patterns in the spatial distribution of *Peruvian anchovy* (*Engraulis ringens*) revealed by spatially explicit fishing data. *Progress in Oceanography*, **79** (2-4): 379-389.

Boutell, M. R., Luo, J., Shen, X., and Brown, C. M., 2004. Learning multi-label scene classification. *Pattern Recognit*, **37** (9): 1757-1771.

Campanis, G., 2008. Advancements in vms data analyses. *NAFO Annual Report for 2008*.

Castro, J., Punzón, A., Pierce, G. J., Marín, M., and Abad, E., 2010. Identification of métiers of the Northern Spanish coastal bottom pair trawl fleet by using the partitioning method CLARA. *Fisheries Research*, **102** (1-2): 184-190.

Chang, S. K., 2011. Application of a vessel monitoring system to advance sustainable fisheries management-Benefits received in Taiwan. *Marine Policy*, **35** (2): 116-121.

Chang, S. K., Liu, K. Y., and Song, Y. H., 2010. Distant water fisheries development and vessel monitoring system implementation in Taiwan – History and driving forces. *Marine Policy*, **34** (3): 541-548.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P., 2011. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16** (1): 321-357.

Chen, T., and Guestrin, C., 2016. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 785-794.

Clarkson, V., Kootsookos, P. J., and Quinn, B. G., 1994. Analysis of the variance threshold of Kay's weighted linear predictor frequency estimator. *IEEE Transactions on Signal Processing*, **42** (9): 2370-2379.

Coro, G., Fortunati, L., and Pagano, P., 2013. Deriving fishing monthly effort and caught species from vessel trajectories. *Oceans, IEEE*, 1-5.

Cortes, C., and Vapnik, V., 1995. Support-vector networks. *Machine Learning*, **20** (3): 273-297.

Cover, T. M., and Hart, P. E., 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13** (1): 21-27.

Deng, R., Dichmont, C., Milton, D., Haywood, M., Vance, D., Hall, N., and Die, D., 2005. Can vessel monitoring system data also be used to study trawling intensity and population depletion? The example of Australia's northern prawn fishery. *Canadian Journal of Fisheries and Aquatic Sciences*, **62** (3): 611-622.

Feng, S. H., and Lang, C. Y., 2017. Graph regularized low- rank feature mapping for multi-label learning with application to image annotation. *Multidimensional Systems & Signal Processing*, **11**: 1-22.

García-de-la-Fuente, L., Fernández-Vázquez, E., and Ramos-Carvajal, C., 2016. A methodology for analyzing the impact of the artisanal fishing fleets on regional economies: An application for the case of Asturias (Spain). *Marine Policy*, **74**: 165- 176.

Gloaguen, P., Mahévas, S., Rivot, E., Woillez, M., Guitton, J., Vermard, Y., and Etienne, M. P., 2015. An autoregressive model to describe fishing vessel movement and activity: An autoregressive model to describe fishing vessel movement and activity. *Environmetrics*, **26** (1): 17-28.

Joo, R., Bertrand, S., Chaigneau, A., and Ñiquen, M., 2011. Optimization of an artificial neural network for identifying fishing set positions from VMS data: An example from the Peruvian anchovy purse seine fishery. *Ecological Modelling*, **222** (4): 1048-1059.

Joo, R., Salcedo, O., Gutierrez, M., Fablet, R., and Bertrand, S., 2015. Defining fishing spatial strategies from VMS data: Insights from the world's largest monospecific fishery. *Fisheries Research*, **164**: 223-230.

Kim, J. S., and Jeong, J. S., 2015. Pattern recognition of ship navigational data using support vector machine. *The International Journal of Fuzzy Logic and Intelligent Systems*, **15** (4): 268-276.

Lambert, G. I., Jennings, S., Hiddink, J. G., Hintzen, N. T., Hinz, H., Kaiser, M. J., and Murray, L. G., 2012. Implications of using alternative methods of vessel monitoring system (VMS) data analysis to describe fishing activities and impacts. *ICES Journal of Marine Science*, **69** (4): 682-693.

Lee, J., 2010. Developing reliable, repeatable and accessible methods to provide high-resolution estimates of fishing-effort distributions from vessel monitoring system (VMS) data. *ICES Journal of Marine Science*, **67** (6): 1260-1271.

Murray, L. G., Hinz, H., Hold, N., and Kaiser, M. J., 2013. The effectiveness of using *CPUE* data derived from vessel monitoring systems and fisheries logbooks to estimate scallop biomass. *ICES Journal of Marine Science*, **70** (7): 1330-1340.

Perera, L. P., Oliveira, P., and Guedes Soares, C., 2012. Maritime traffic monitoring based on vessel detection, tracking, state estimation, and trajectory prediction. *IEEE Transactions on Intelligent Transportation Systems*, **13** (3): 1188- 1200.

Russo, T., Carpentieri, P., Fiorentino, F., Scardi, M., Cioffi, A., and Cataudella, S., 2016. Modeling landings profiles of fishing vessels: An application of self-organizing maps to VMS and logbook data. *Fisheries Research*, **181**: 34- 47.

Russo, T., Parisi, A., Prorgi, M., Boccoli, F., Cignini, I., Tordoni, M., and Cataudella, S., 2011. When behaviour reveals activity: Assigning fishing effort to métiers based on VMS data using artificial neural networks. *Fisheries Research*, **111** (1-2): 53-64.

Shewhart, W. A., and Wilks, S. S., 2005. *Applied Logistic Regression*, 2nd Edition, 1-22.

Vermard, Y., Rivot, E., Mahévas, S., Marchal, P., and Gascuel, D., 2010. Identifying fishing trip behaviour and estimating fishing effort from VMS data using Bayesian Hidden Markov Models. *Ecological Modelling*, **221** (15): 1757-1769.

Walker, E., and Bez, N., 2010. A pioneer validation of a state-space model of vessel trajectories (VMS) with observers' data. *Ecological Modelling*, **221** (17): 2008-2017.

Wang, H., Osen, O. L., Li, G. Y., Li, W., Dai, H. N., and Zeng, W., 2015. Big data and industrial internet of things for the maritime industry in Northwestern Norway. *TENCON 2015 - 2015 IEEE Region 10 Conference IEEE*, 1-5.

Wang, Q., and Dai, H. N., 2013. On modeling of eavesdropping behavior in underwater acoustic sensor networks. *International Symposium on A World of Wireless*, *Mobile and Multimedia Networks IEEE*, 1-3.

Wang, Y., Liu, Y., and Guo, Z., 2012. Three-dimensional ocean sensor networks: A survey. *Journal of Ocean University of China*, **11** (4): 436-450.

Williams, E., 2011. Aviation Formulary V1. 46. Aviation.

Witt, M. J., and Godley, B. J., 2007. A step towards seascape scale conservation: Using Vessel Monitoring Systems (VMS) to map fishing activity. *PLoS One*, **2** (10): e1111.

Zong, Y., Huang, H., Hong, F., Zhen, Y., and Guo, Z., 2016. Recognizing fishing activities *via* VMS trace analysis based on mathematical morphology. *Techno-Ocean IEEE*, 465-470.

(Edited by Ji Dechun)