Article:
Gan, S, Liang, S, Li, K et al. (2 more authors) (2018) Trajectory Length Prediction for Intelligent Traffic Signaling: A Data-Driven Approach. IEEE Transactions on Intelligent Transportation Systems, 19 (2). pp. 426-435. ISSN 1524-9050

https://doi.org/10.1109/TITS.2017.2700209

Trajectory length prediction for intelligent traffic signalling: a data driven approach

Shaojun Gan, Shan Liang, *Member, IEEE,* Kang Li, *Senior Member, IEEE,* Jing Deng, *Member, IEEE,* and Tingli Cheng

***Abstract*—Yangtze River is one of the world's most important cargo-carrying rivers. However, the traffic capacity is becoming the bottleneck for further developments. This has been highlighted in recent Yangtze River economic zone proposal in which the improvement of the Yangtze River traffic capacity is a key project. Efficient traffic management based on ships' trajectory length prediction is a key way to improve the traffic capacity. Yet, in existing intelligent traffic signalling systems (ITSSs), ships are supposed to travel exactly along the central line of the Yangtze River which is often not a valid assumption and has caused a number of problems. Over the past few years, traffic data have been accumulated exponentially, leading to the big data era. This trend allows more accurate prediction of ships' travel trajectory length based on historical data. In this paper, ships' historical trajectories are first grouped by using the Fuzzy C-Means clustering algorithm. The relationship between some known factors (i.e. ship speed, loading capacity, self-weight, maximum power, ship length, ship width, ship type and water level) and the resultant memberships are then modeled using Artificial Neural Networks (ANN). The trajectory length is then estimated by the sum of the predicted probabilities multiplied by the trajectory cluster centers' length. The experimental results show that the proposed method can reduce the probability of generating wrong traffic control signals by 89% over existing ITSSs. This will significantly improve the efficiency of the Yangtze river traffic management system, and increase the traffic capacity by reducing the travelling time.**

***Index Terms*—trajectory prediction, data driven, fuzzy c-means (FCM), artificial neural networks (ANN), intelligent traffic signalling system (ITSS).**

I. Introduction

Yangtze River has been the world's busiest navigable inland waterway since 2010 as more freights are transported through the Yangtze River than the other inland waterways [1]. Record shows that 2.18 billion tons of cargo were shipped through the main reaches of the Yangtze river in 2015, which accounts for 80% of the river freight in China [2]. The figure is expected to reach 6.2 billion tons in 2030, which is about 17% of the global total shipping volume [3].

Controlled waterways are special areas in the Yangtze River with unfavorable geographical conditions, e.g. narrow channels, sharp curves and raging water. Fig. 1 shows the map of the Shenbeizui Controlled Waterway in Sichuan Province, China. Ships are only allowed to pass in one direction at a time in this U-shape controlled waterways for safety reasons. This means, if two ships intend to pass through the Shenbeizui Controlled Waterway from both directions, only one ship is allowed to pass at one time while the other must wait outside the controlled zone until it is cleared. This becomes much more complicated if more ships arrive at both directions. A number of intelligent traffic signalling systems have been deployed along the Yangtze River to control the traffic. However, the accuracy of the generated traffic signals is quite low, which becomes the key factor limiting the traffic capacity of the Yangtze river [4].

The traffic signalling of the controlled waterways entirely relies on the time a ship needs to pass through the controlled waterways. This can be computed through the trajectory length and the speed of the each ship. In our previous work, a novel algorithm has been proposed to build ship speed model for long-term prediction in the Yangtze River traffic management [5]. Although the trajectories of each downstream ship follow nearly the same trend and end up with similar length, the upstream ships are in a different scenario. It varies a lot due to ragging water, limited engine power, etc. Thus, this paper mainly focuses on the trajectory length prediction for upstream ships.

Ships' trajectory has been studied from different perspectives in the literature, and a few were reported for inland waterway applications. However, most efforts were made to avoid ship collisions instead of improving the traffic efficiency. A three layered BP neural network was built to predict the ship's trajectory where the ship's speed and course were used as the input of the BP network, and the ship's position change was the output [6]. Even though the proposed model is computationally efficient, It can only achieve 1 minute ahead trajectory prediction with satisfactory accuracy. Sutulo et al. proposed a simplified realistic dynamic mathematical model by eliminating a number of secondary effects and using a very small number of input data. The model can be computed at high speed, but it is still limited to short time prediction [7]. Perera et al. proposed an extended Kalman filter (EKF) to predict ship states and trajectory for both vessel navigation systems and vessel traffic monitoring systems [8]. The EKF was shown to have perfect performance in estimating ship speed and acceleration from noised data. However, it has the same limitations listed above, and thus is not suitable for trajectory length prediction in Yangtze River. Gerben et al.

S. Gan and S. Liang are with the Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing, 400044, P. R. China, also with the School of Automation, Chongqing University, Chongqing, 400044, P. R. China e-mail: cqugsj@gmail.com; lightsun@cqu.edu.cn.

K. Li and J. Deng are with School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast, BT9 5AH, UK.

T. Cheng is with School of Electrical Engineering, Chongqing University, Chongqing, 400044, P. R. China.

S. Liang is the corresponding author (e-mail: lightsun@cqu.edu.cn).

proposed an unsupervised long-term ship trajectory modeling and prediction method for a certain marine region [9]. The trajectories used in the experiment is more or less the same within a week, thus the trajectory can be predicted directly by using clustering algorithm, which is not the case in the controlled waterways in the Yangtze River. Serrano built a three-degree ship freedom model to control the ships to follow trajectories previously established [10]. The model was built based on the surge, sway and yaw which are very difficult to obtain for most manual operation ships travelling along the Yangtze river.

According to our previous study, the ships' trajectories heavily depend on the ship speed, loading capacity, self-weight, engine power, ship size, ship type and water level. Due to the fast development of information and communication technologies, the volume of historic traffic data has been growing rapidly, leading to the era of big data. Transportation management and control is becoming more data-driven based. Although many ship trajectory prediction systems and models have been proposed, the majority can only achieve short-term trajectory prediction and the performance is still unsatisfactory. This inspires us to reconsider the ship trajectory length prediction problem based on big historical traffic data.

In this paper, the historical trajectories were first partitioned into 5 segments according to its variance and then clustered into smaller groups using the Fuzzy C-Means (FCM) algorithm for each segment. The clustered trajectories along with their memberships which are obtained by the FCM algorithm, and together with other traffic data were then used to train the ANN models. The predicted trajectory length is obtained by summing the weighted mean length of clustered trajectories. The experimental results confirm that the resultant models can predict ships' trajectory lengths with a fairly satisfactory accuracy and significantly reduce the error of traffic control signals. The contribution of this work are four-fold: 1) Overall trajectory length prediction. 2) Build a model which relates the unknown ship trajectory with its known factors (ship speed, loading capacity, self-weight, maximum power, ship length, ship width, ship type and water level). 3) Data peprocessing. 4) Reduce the probability of generating wrong traffic control signals by 89% over existing ITSSs. The developments will become an important part of the ITSSs, where the optimal traffic control signals are highly dependent on the accuracy of the ships' predicted trajectory length.

The rest of the paper is organized as follows. Section 2 gives a brief introduction of the controlled waterways in Yangtze River and issues around the traffic management in controlled waterways. In section 3, the adopted method will be presented in detail. Then in section 4, the ship trajectories are clustered and modelled, and the effectiveness of the proposed models are tested and verified. Finally, section 5 concludes the paper with remarks on future work.

II. Problem Statement

A. Traffic Management in Controlled Waterways of Yangtze River

Controlled Waterways are dangerous areas which widely exist in the upper reaches of the Yangtze River as well as other inland waterways around the world. In a controlled waterway, due to the hazard condition, such as narrow channel, sharp curve, and raging water, ships are only allowed to pass in one direction at a time to ensure transportation safety. Ships travel toward the opposite direction have to wait outside the controlled waterway until it is cleared. In order to guide ships passing through the controlled waterways efficiently and safely, ITSSs are used to generate control signals based on current traffic conditions, e.g., the number of ships in waiting areas, the distance of ships from controlled waterway, and the ship speeds etc.

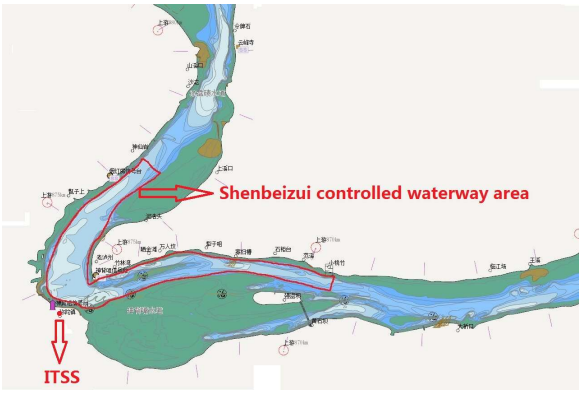

Fig. 1. Electronic Navigational Chart of Shenbeizui Controlled Waterway.

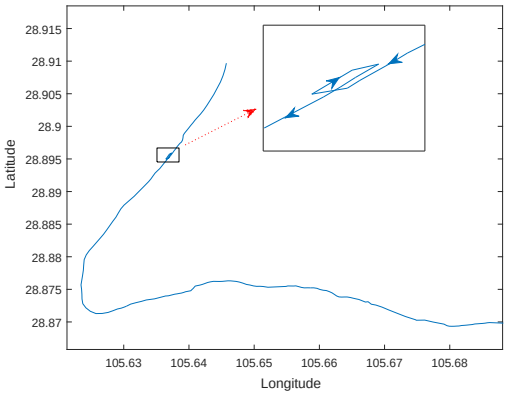

Fig. 2. The trajectory of a downstream ship turn round outside Shenbeizui Controlled Waterway.

Another issue is that downstream ships have to turn round and head upstream in order to make a stop because of the torrential water flow in the controlled waterways, as shown in Fig. 2. Due to harsh geological features of the controlled waterways, it is very dangerous for ships to turn round. Thus, the ITSSs were deployed to improve the controlled waterways' traffic efficiency while avoid the stop and waiting of downstream ships. Due to the above reason, downstream ships should be given higher priority when passing through the controlled waterways.

However, as current ITSSs often assume that all ships travel exactly along the middle line of the Yangtze River, the predicted ship passing time is inaccurate, leading to ineffective even wrong traffic control signals. In reality, the

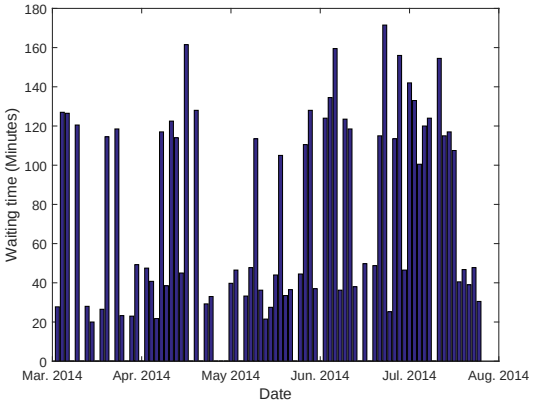

Fig. 3. Waiting time of a ship passing through Shenbeizui Controlled Waterway from Mar. 2014 to Aug. 2014

controlled waterways are often occupied by upstream ships when downstream ships arrive. Therefore, the downstream ships have to wait and thus need to make unnecessary U-turns. Such non-optimal traffic control signals not only lead to hours of waiting, but also impose potential risks to the ships when making the U-turns. Fig. 3 shows the waiting time of a randomly selected ship passing through the Shenbeizui Controlled Waterway from March 2014 to August 2014. Each bar represents the waiting time of this ship when passed through the controlled waterway. During this six months period, this ship has travelled through the Shenbeizui Controlled Waterway for 86 times. Its total waiting time is 90 hours, with an average of 62.6 minutes per journey. Since hundreds of thousands of ships are operating on the Yangtze River every day, accurate trajectory length prediction is urgently needed for intelligent traffic signalling.

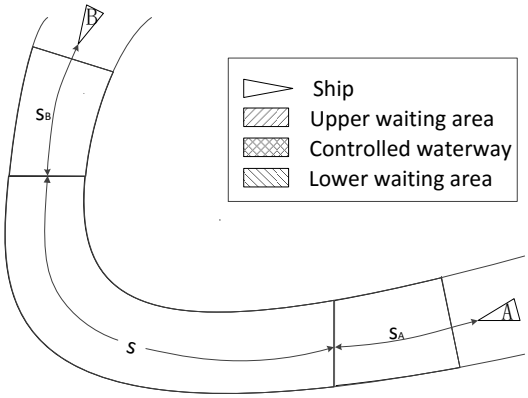

Fig. 4. Simplest situation in Controlled Waterway

Fig. 4 illustrates a basic situation of traffic management in a controlled waterway. Ships A and B are going to travel through the controlled waterway. The existing ITSSs generate the traffic signals by the following simplified model if two ships are about to pass through the controlled waterway from different directions at the same time.

$$\left.\begin{array}{l} t_A = \frac{S_A}{V_A}; \quad t_B = \frac{S_B}{V_B}; \\ t_{SA} = \frac{S}{V_A}; \quad t_{SB} = \frac{S}{V_B}; \\ A = 1, \quad B = 0 \\ WT_B = t_B - t_A - t_{SA} \\ A = 0, \quad B = 1 \\ WT_A = t_B + t_{SB} - t_A \end{array}\right\} \begin{array}{l} if\ t_B \geq t_A + t_{SA} \\ \\ \\ \\ if\ t_B < t_A + t_{SA} \end{array} \quad (1)$$

where S denotes the length of the controlled waterway, S_A and S_B are the distances of ships A and B from the controlled waterway. V_A and V_B represent the velocity of ships A and B respectively. t_A and t_B are the time of ships A and B arriving at the controlled waterway. t_{SA} and t_{SB} indicate their time of passing through the controlled waterway. WT_A and WT_B are their waiting times. $A, B \in \{0, 1\}$, 0 and 1 are red waiting signal and green pass signal respectively. The objective of the decision making process is to minimise the total waiting time while guaranteeing the priority of downstream ships. It is obvious that the performance of the model (1) highly depends on the accuracy of the trajectory length prediction.

Upstream ships also have some difficulties in choosing a short trajectory due to their power limitation and ragging water conditions. They need to turn frequently in the controlled waterways to compensate for adverse water stream force, and end up with tranquil but long trajectories. This situation is often more frequent when ships are full loaded and underpowered.

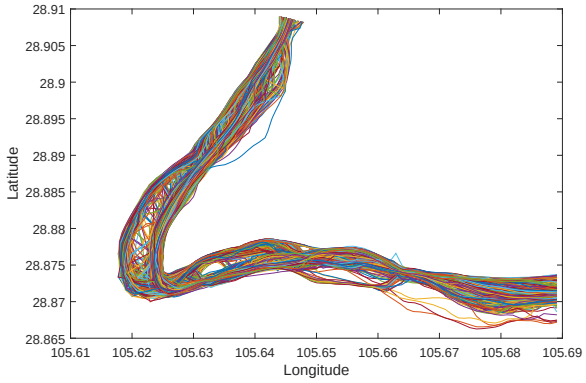

Fig. 5. Trajectories of upstream ships passing through the Shenbeizui Controlled Waterway

Fig. 5 shows the trajectories of 2000 randomly selected upstream ships passing through the Shenbeizui Controlled Waterway. It is clear that the actual ship trajectories don't follow the middle line which is adopted in current ITSSs to calculate the trajectory as well as passing time. Thus, the non-optimal signal generation becomes inevitable due to incorrect prediction of S_A and S_B. Fig. 6 shows the length of the 2000 ships' trajectories, which indicates large variation from the middle line (red). The incorrect trajectory length prediction in current ITSSs not only leads to hours of unnecessary waiting, but also increases the chance of waiting for downstream ships which is against the controlled waterway traffic management regulations.

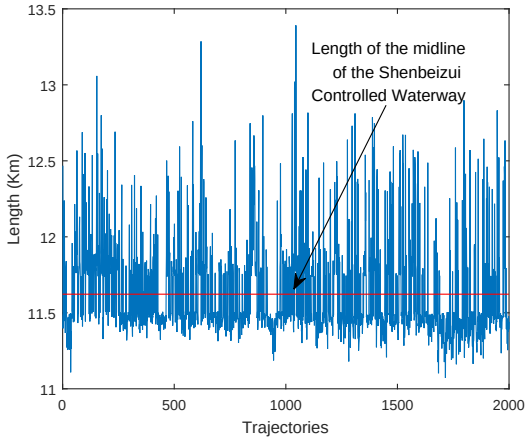

Fig. 6. Trajectory Length

B. Problem Formulation

Based on the above analysis, accurate trajectory length prediction is vital for the traffic signalling on the Yangtze River. However, manually controlled ships have lots of uncertainties in their travelling trajectories. thus it is very difficult to predict the trajectory or its length accurately based on historical data. However, the trajectories can be clustered into small groups based on their similarities. The trajectories within the same cluster have similar length distribution. Our previous studies show that, the upstream ships' trajectories are related to ship speed, loading capacity, self-weight, maximum power, ship size and water level. Therefore, a mathematical model based on these known factors could be built to calculate the probability of a trajectory belonging to each cluster. The length of the predicted trajectory can be presented by summing the probability weighted cluster means.

Suppose $\{x_i\}_{i=1}^{m} \subseteq R^n$ denotes the trajectories which will be clustered into C groups, $g_1, g_2, \dots g_C$. The aim of the first stage is to estimate the membership w_{ij} which denotes the probability of trajectory x_i belonging to cluster g_j. The second stage will calculate the probability of w_{ij} for all possible i and j ($j \in [1, C]$) based on known factors of ship i, i.e. speed v_i, loading capacity lc_i, self-weight sw_i, maximum power mp_i, ship length len_i, ship width wid_i, ship type ty_i and water level wl_i, as shown in (2). The function f is normally unknown and needs to be identified. In this paper, an Artificial Neural Network (ANN) is adopted.

$$w_{ij} = f(v_i, lc_i, sw_i, mp_i, len_i, wid_i, ty_i, wl_i) \tag{2}$$

where $i = 1, 2, \dots, m, j = 1, 2, \dots, C$.

Thus, the predicted length of the ith trajectory $\hat{l}_i$ can be obtained by summing the probability weighted cluster centers length.

$$\hat{l}_i = \sum_{j=1}^{C} w_{ij} \left\| g_j \right\| \tag{3}$$

where $\|g_j\|$ denotes the mean length of cluster j.

III. Methodology

A. FCM clustering

FCM is a soft clustering algorithm which assigns memberships to each data corresponding to all clusters based on the distances of the data and the cluster centers [11]. Let $X = (x_1, x_2, \dots, x_N)$ denotes N data samples to be clustered into C groups. The algorithms aims to minimize the objective function (4) iteratively.

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} \mu_{ij}^{m} \|x_i - g_j\|^2 \tag{4}$$

where $m = 1, 2, \cdots, \infty$ is an index that controls the fuzziness of the resulting partition. In the absence of experimentation or domain knowledge, m is usually set to 2. $\mu_{ij} \in [0, 1]$ is the membership indicating the probability of x_i belonging to cluster j, thus,

$$\sum_{i=1}^{N} \mu_{ij} = 1 \ \ (j = 1, 2, \dots, C) \tag{5}$$

g_j is the center of j_{th} cluster, and $\|*\|$ denotes the similarity between data x_i and the cluster center c_j.

The objective function J is minimized when larger membership values are assigned to data closer to the cluster centers, and vice versa. The memberships and the cluster centers are updated by:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - g_j\|}{\|x_i - g_k\|} \right)^{\frac{2}{m-1}}} \tag{6}$$

$$g_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}} \tag{7}$$

if m is set to 1, the memberships μ_{ij} converge to 0 or 1, which transforms to a hard clustering method.

The pseudo code of FCM algorithm is summarised in Algorithm 1. Specifically, an appropriate number of clusters C are determined first followed by the initialization of C centres randomly selected from the data set. The centers and memberships are then updated iteratively until a stop criterion is met, such as the changes of memberships become small as shown in Eq.(8) or the centres don't change in two successive iteration steps.

$$\max_{ij} \left\{ \left| \mu_{ij}^{(k+1)} - \mu_{ij}^{(k)} \right| \right\} < \varepsilon \tag{8}$$

B. Artificial Neural Network

Artificial neural networks, which were initially inspired by biological neural net, usually contain at least 3 layers: input layer, hidden layer and output layer. The input data are fed to the network through the input layer and then delivered to the hidden layers by multiplying layer weights. In the hidden layers, the weighted sum of the input data is fed to a nonlinear activation function, e.g. sigmoid function. The processed data

Algorithm 1: Pseudo Code of FCM

Input:
Number of clusters C, Data set X, Fuzziness index m;

Output:
Cluster centers c, membership of all data μ;

```
1: Randomly select C cluster centers;
2: Calculate the initial memberships;
3: repeat:
3:    for(i=1;i<=N;i++) do
4:       for(j=1;j<=C;j++) do
5:          update the cluster centers according to (7);
6:       end for
7:    end for
8:    for(i=1;i<=N;i++) do
9:       for(j=1;j<=C;j++) do
10:         update membership values according to (6);
11:      end for
12:   end for
13: Until stop criteria (8)
```

is finally transferred to the output layer to calculate neural network outputs. All layers in neural network consist of neurons which are the fundamental processing elements, and the neurons from neighbouring layers are interconnected with weights. It has been proved that the artificial neural network is capable of representing any continuous function with an arbitrary degree of accuracy by tuning the weights [12].

Fig. (7) schematically illustrates a typical $m-k-n$ network. Let w_{ji}^{l} denotes the weight of the ith neuron in $(l-1)$th layer to the jth neuron in lth layer. The output of jth neuron in lth layer can be computed as

$$a_j^l = f\left(\sum_i w_{ji}^l a_i^{l-1}\right) \tag{9}$$

where f is the activation function of the neurons in jth layer. Consequently, the output of the network can be calculated as:

$$a_j^3 = \tau\left(\sum_{i=1}^{k} w_{ji}^3 \sigma\left(\sum_{l=1}^{m} w_{jl}^2 a_l^1\right)\right) \tag{10}$$

where $j = 1, 2, \ldots, n$ is the number of neurons in the output layer. τ and σ are the activation function of output layer and hidden layer respectively.

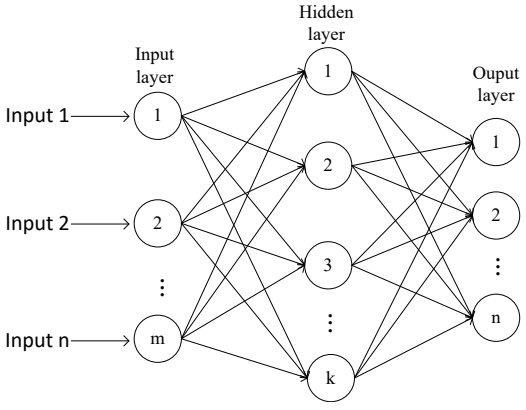

Fig. 7. Basic Structure of BP network

The most commonly used artificial neural network for classification problem is the multilayer perceptron (MLP)

Algorithm 2: Pseudo Code of BP

Input
Training data set
Output
Neural network

```
1:Initialize all weights with random values
2:repeat:
  //Propagated the input forward through the network:
3:  for each layer in the network
4:    for every node in the layer
5:      Calculate the weighted sum of inputs to the nodes
6:      Calculate the node outputs
7:    end
8:  end
  //Propagate the errors backward through the network:
9:  for every node in the output layer
10:   calculate the error signal
11: end
12: for all hidden layers
13:   for every node in the layer
14:     Calculate the node's signal error
15:     Update each node's weight in the network
16:   end
17: end
18:until stop criteria
```

trained by the back-propagation (BP) algorithm [13]. BP is a supervised training method for neural networks which relies on errors between network outputs and desired outputs. This error is minimized by tuning the weights using gradient descent optimization methods. Therefore, the training process can be divided into two phases. In the forward phase, the input data is fed to generate the network output. While in the back propagation phase, the network error was used to tune the weights backwards. The pseudo code for the BP algorithm is given in Algorithm 2.

IV. Experiments

A. Data Description

The real data collected from the Shenbeizui Controlled waterway ITSS are used in the experiment as a numerical example. The ITSS retrieves data from a AIS (Automatic Identification System) station which transmits ships' information by radio. AIS is an automatic tracking system used for identifying ships by exchanging information (i.e. Maritime Mobile Service Identity (MMSI), ship name, type, dimension, draught, navigation status, rate of turn, speed over ground, longitude, latitude, true heading etc.) with nearby AISs through 161.975 and 162.025 MHz VHF (Very High Frequency) radio. Fig. 8 depicts the basic structure of AIS data acquisition system. In the Yangtze Rive, all ships are required to deploy AIS. Thus, the ITSS could estimate traffic conditions by analysing the ships' AIS information.

In this paper, 2000 randomly selected upstream ships' trajectories corresponding with their speed, maximum power, self-weight, loading capacity, ship type, ship width, ship length and the water level were used. All the 2000 trajectories were first clustered by the FCM algorithm to obtain their memberships belonging to the clusters. Then, 70% of the data were used for ANN training, while the remaining 30% data were reserved for testing. The detailed procedure of the experiment is illustrated in Fig.9.

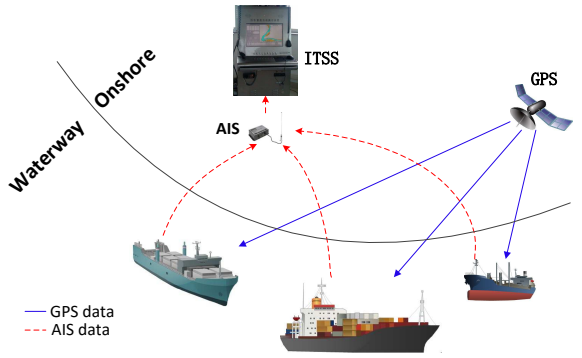

Fig. 8. The structure of AIS data acquisition

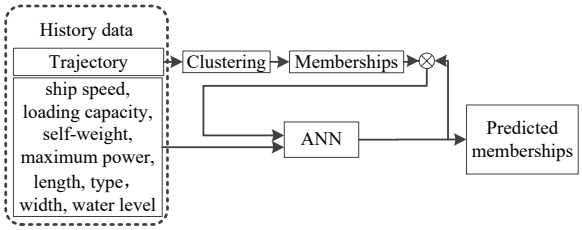

Fig. 9. Block diagram of the experiment procedure

B. Data pre-processing

1) Dimension reduction: In practice, ships broadcast their status at different frequencies relating to their speed [14]. This causes the ship trajectories nonaligned, and the sampling rates vary. FCM cannot be applied to these raw data. Thus data pre-processing becomes necessary. Firstly, all the 2000 trajectories were aligned to the same starting and ending lines, and then re-sampled to the same size, i.e. same number of points in each trajectory. Fig. 10 shows the pre-processing of two raw trajectories, $\bullet$ and $*$ are the position points of two different trajectories. Let $trajectories = \{tra_1, tra_2, \ldots tra_{2000}\}$ denotes the total 2000 trajectories to be used in the experiment, where $tra_i = \left\{p_i^j\right\}_{j=1}^{m} \subseteq R^2$ is the ith trajectory with m sampling points, $p_i^j = \left\{x_i^j, y_i^j\right\}$ represents the jth point in ith trajectory and x_i^j and y_i^j are the longitude and latitude respectively.

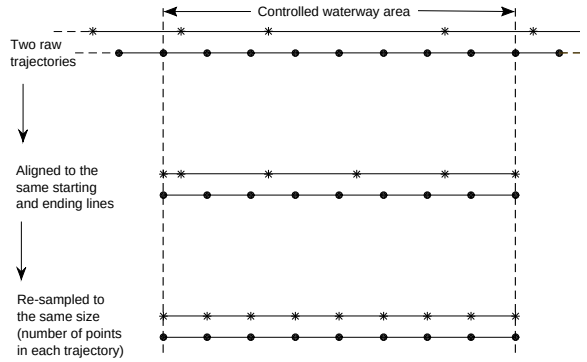

Fig. 10. Example of alignment and re-sampling of two raw trajectories

It has been proven that high data dimension increases difficulty for clustering algorithm due to more expensive computational cost and sensitivity to redundancies [15], [16]. To tackle this issue, all the trajectories' coordinates were transformed from GPS to the distances from the bank of the controlled waterway, shown in Fig. 11. The resultant trajectories can be represented as $tra_i = \left\{d_i^j\right\}^m \in R$ where d_i^j denotes the distance of ith point in jth trajectory to the bank of the Shenbeizui Controlled Waterway. This leads to reduced data dimension without losing any information in the original trajectories. The data dimension has reduced from $2m$ to m.

TABLE I
NUMERIC NOTATION OF SHIP TYPES

Ship type	Code	Ship type	Code
Cargo ship	1	Ferry	5
Oil tanker	2	Motor boat	6
Multi-purpose ship	3	Bulk carrier	7
Container ship	4	Passenger ship	8

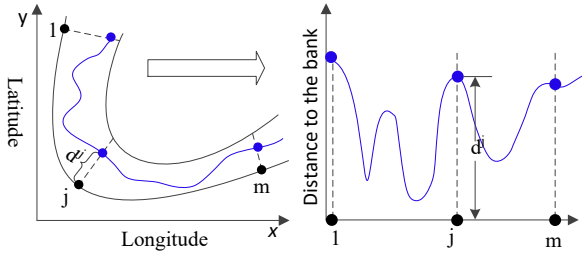

Fig. 11. Example of the coordinates transformation

2) Trajectory partitioning: Another challenge in trajectory clustering is to deal with large uncertainties and complicated waterway geography. [17].

From historical data, it has been noticed that the branch area in controlled waterway significantly restricted the freedom of travelling. Thus the ship trajectories are very close to each other at these points as shown in Fig. 12. This can be further confirmed by their variations. As a result, the trajectories were partitioned into 5 segments as illustrated in Fig. 12. The FCM algorithm was then applied to each of these five segments. In addition, The ANN model was also built for each segment to learn the latent relationship between ships' trajectory and other known factors where ship types were coded according as in table I.

C. Index of performance

To evaluate the performances of the proposed algorithm, three indexes, i.e. normalized prediction error (NPE), mean absolute error (MAE) and mean squared error (MSE) were used in this paper. They are defined as

$$NPE = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{l}_i - l_i\right)^2}{\sum_{i=1}^{n} {l_i}^2}} \times 100\% \tag{11}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{l}_i - l_i\right| \tag{12}$$

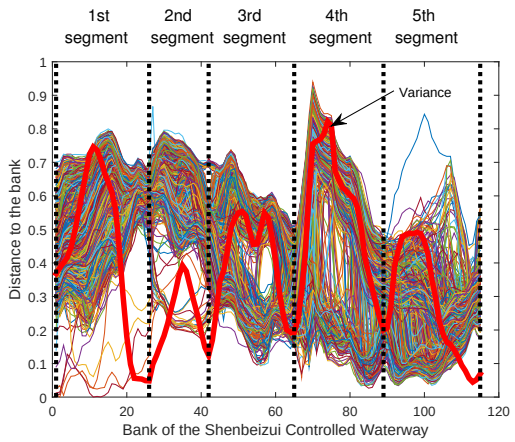

Fig. 12. The divided trajectory segments in the transformed coordinates

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{l}_i - l_i\right)^2 \tag{13}$$

where l_i is the actual length of ith trajectory, and $\hat{l}_i$ is the predicted length.

D. Trajectory clustering

FCM clustering needs the similarity information between trajectories. In this case, the similarity between trajectories i and k was defined as the sum of gaussian distance between the corresponding points

$$d(tra_i, tra_k) = \sum_{j=1}^{m} dis(p_i^j, p_k^j) \tag{14}$$

where $dis(p_i^j, p_k^j)$ gives the gaussian distance between the jth point of trajectories i and k.

Another parameter to be determined for FCM is the number of clusters. This requires some prior knowledge about the data set otherwise a guess has to be made. In this study, prior knowledge and Akaikes information criterion are adopted to estimate the optimal number of clusters [18] which are 3, 2, 2, 3 and 4 for the above five segments. Trajectories were assigned to the clusters based on their memberships. The clustering results are shown in Fig. 13. The length of each cluster center is given in Table II. Note that there is no standard right answer for the clustering results, the effectiveness of the clustering results depends on the clustering purpose.

TABLE II
THE LENGTH OF CLUSTER CENTRES IN EACH SEGMENT

Cluster Labels	1	2	3	4
1st segment (km)	2.3912	2.2843	2.3234	-
2nd segment (km)	1.8219	1.8110	-	-
3rd segment (km)	2.2547	2.3770	-	-
4th segment (km)	2.4573	2.9713	2.2171	-
5th segment (km)	2.7040	2.6230	2.6681	2.7499

E. Trajectory modelling

The second phase of the experiment was to build ANN models to predict the ships' probabilities of selecting different trajectory clusters based on their known information, i.e. ship speed, loading capacity, self-weight, maximum power, ship size, ship type and water level. These variables were selected as the inputs of the ANN and the outputs were memberships to each cluster. This offers several advantages over traditional methods which use clustering label as the targets. In this application, the clusters do not have clear boundaries, thus a small error in membership would lead to a completely wrong cluster label. Using membership values as the targets can avoid these problems and improve the modelling accuracy.

Softmax function [19] was selected as the activation function of the output layer in the ANN to make sure the model gives a valid probability distribution, i.e, all outputs are greater than 0 and their sum equals to 1. The output of ith neuron in last layer is calculated using (15). The cross entropy cost function [20] was used to obtain an error vector in the output layer, which was then backpropagated to the hidden layer and input layer to tune the weights of the ANN.

$$s(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{L} e^{z_j}} \tag{15}$$

where L is the number of neurons in output layer, and e^{z_j} is the input of jth neuron in this layer.

Five ANN models were then built to predict the ships' probabilities of choosing different trajectory clusters in the above five segments. The predicted trajectory length is given by the sum of cluster center lengths weighted by the predicted memberships. The experimental results show that the built model can achieve much better performance in predicting the ships' trajectory length when compared to the existing method. The MSE, MAE and NPE from both proposed method and existing systems were shown in table III. The built models have achieved the best performance for the 2nd trajectory segment with 0.0014 MSE while it is 0.0715 for segment 4. Obviously, a significant improvement has been made over existing mid-line prediction method.

TABLE III
PERFORMANCE OF THE BUILT ANN MODELS AND THE EXISTING METHOD

		MSE	MAE	NPE
ANN	1st segment	0.0025	0.0332	2.13%
	2nd segment	0.0014	0.0285	2.05%
	3rd segment	0.0079	0.0775	3.74%
	4th segment	0.0502	0.1531	9.52%
	5th segment	0.0033	0.0420	2.13%
	Total	0.0715	0.2208	2.33%
Mid-line	Total	0.1237	0.3150	3.28%

F. Discussion

Trajectory length prediction is vital for accurate and efficient traffic signalling in the Yangtze River. Even a small improvement in trajectory length prediction could make a big difference in traffic signaling optimization. For example, ship A with the information listed in Table IV travelled through

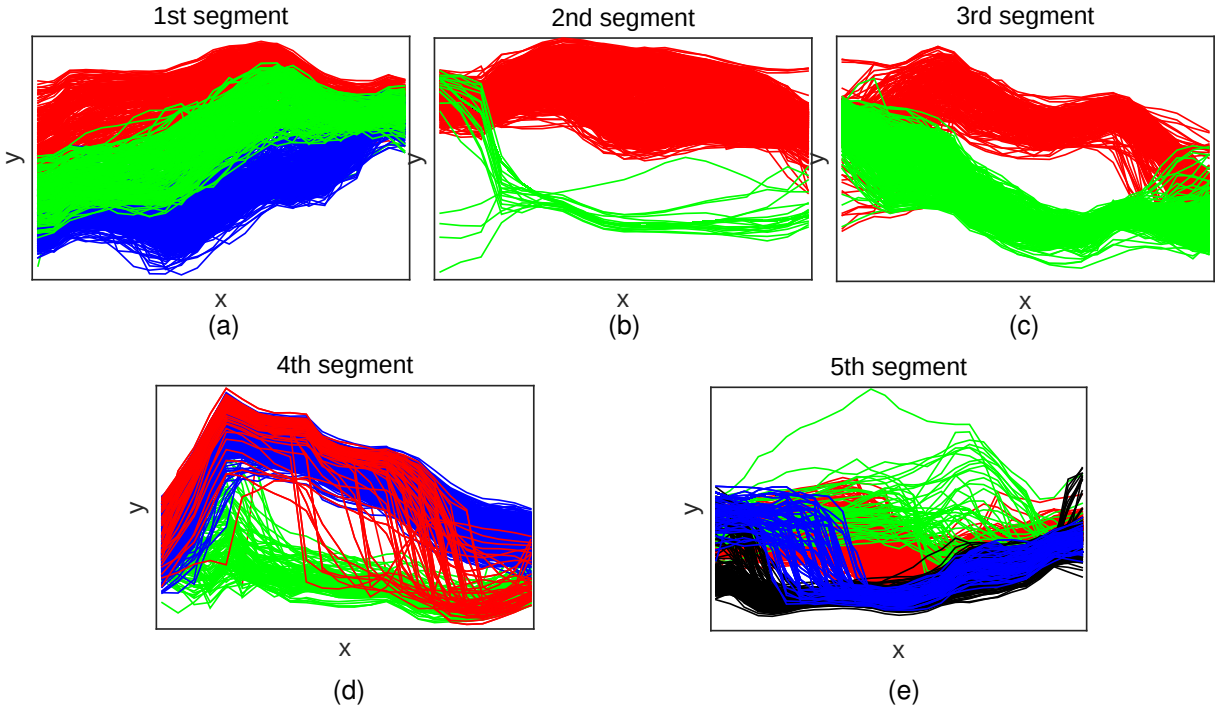

Fig. 13. Clustering results of the trajectory in each segment (x and y are the new coordination which represent the bank of the Shenbeizui Controlled Waterway and the distances to the bank respectively). (a) 1st segment. (b) 2nd segment. (c) 3rd segment. (d) 4th segment. (e) 5th segment.

the Shenbeizui Controlled Waterway on November 2, 2015. The trajectory length predictions based on the built ANN models and the existing mid-line method are 11.4749 and 11.6255 km respectively, while the actual length is 11.4900 km. The time ship A spent in the lower waiting area and the controlled waterway based on different trajectory length prediction methods are listed in table V. Suppose another ship B was approaching the controlled waterway at the same time, a traffic signal needs to be generated according to the simplified signalling model (1). Table VI shows the traffic signals from different approaches under 4 scenarios. Although none of these method can lead to correct signals for all different t_B, the built ANN models are least likely to generate a wrong signal except when t_B is between 92.39 and 92.51 minutes, while in the existing mid-line prediction method, a wrong signal would be generated when t_B is between 92.51 and 93.59 minutes . In this case, the built models reduce the wrong signal time period by 0.96 minutes which is 89% in probabilities by compare to existing method. In practice, a wrong signal could lead to hours of unnecessary waiting time or risky U-turn for downstream ships. Thus, improved accuracy in trajectory length prediction is always vital in ITSSs, especially for the busy Yangtze River.

TABLE IV
SHIP A'S INFORMATION

Factor	Value	Factor	Value
Loading capacity	681 tons	Ship type	Cargo ship
Self-weight	306 tons	Length	61 meters
Maximum power	280 kilowatt	Width	12 meters
Speed	2.9154 knot	Water level	3.97 meters

TABLE V
TIME OF SHIP A SPEND TO PASS THROUGH SHENBEIZUI CONTROLLED WATERWAY AREA BASED ON DIFFERENT TRAJECTORY PREDICTION APPROACHES

	Controlled waterway (t_A)	Lower waiting area (t_{SA})	$t_{SA}+t_A$
Mid-line prediction	58 min	35.59 min	93.59 min
ANN	57.26 min	35.13 min	92.39 min
True value	57.33 min	35.18 min	92.51 min

TABLE VI
TRAFFIC SIGNALLING FOR SHIPS A AND B BASED ON DIFFERENT TRAJECTORY LENGTH PREDICTION APPROACHES

$t_B(min)$	Mid-line prediction	ANN prediction	Desired signal
$\leq$92.39	A=0;B=1	A=0;B=1	A=0;B=1
92.39-92.51	A=0;B=1	**A=1;B=0**	A=0;B=1
92.51-93.59	**A=0;B=1**	A=1;B=0	A=1;B=0
$\geq$93.59	A=1;B=0	A=1;B=0	A=1;B=0

*Wrong signals are marked in bold.

V. CONCLUSION

In this paper, strategies to predict ships' trajectory length based on the known ship information (i.e. ship speed, loading capacity, self-weight, maximum power, ship type, ship length, ship width and water level) has been developed to improve the traffic signalling in the controlled waterways of Yangtze River. This is achieved by two phases. Firstly, the whole trajectories were partitioned into 5 segments according to the variances and then the clustering was implemented each segment of all trajectories. Secondly, ANN models were built to predict ships' probability of selecting different trajectory

clusters in each segment. The trajectory length prediction was obtained by summing the weighted cluster center length. The experiment confirmed that the proposed approach could reduce the probability of false signalling by more than 89% compared to the existing system.

Although the research has reached its aim, there are still some limitations. First, this research predicts the trajectory length by summing up the weighted cluster centre length. Thus, the exact trajectory is not predictable by the proposed method, this may limit its further applications. Second, the proposed method performs much better in controlled waterways than in broad and smooth waterways, that may be because ships travel more arbitrary in broad and smooth waterways.

In the future, other approaches will be investigated to further improve the performance of ship trajectory length prediction. In addition, other factors which influence the traffic signalling in the Yangtze River may also be considered.

ACKNOWLEDGMENT

S. Gan would like to thank the sponsorship of Chinese Scholarship Council for funding his research at Queen's University Belfast.

REFERENCES

[1] D. Zhang, X. Yan, Z. Yang, and J. Wang, "An accident data–based approach for congestion risk assessment of inland waterways: A yangtze river case," *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of risk and reliability*, vol. 228, no. 2, pp. 176–188, 2014.

[2] X. Yang, "14 chinas central region," *China's Evolving Industrial Policies and Economic Restructuring*, p. 241, 2014.

[3] "China shipping development outlook 2030," shanghai international shipping institute, Shanghai, China, Tech. Rep., 5 2015.

[4] S. Zhao, C. Tang, S. Liang, and D. Wang, "Track prediction of vessel in controlled waterway based on improved kalman filter," *Jisuanji Yingyong/ Journal of Computer Applications*, vol. 32, no. 11, pp. 3247–3250, 2012.

[5] S. Gan, S. Liang, K. Li, J. Deng, and T. Cheng, "Long-term ship speed prediction for intelligent traffic signalling," *Intelligent Transportation Systems, IEEE Transactions on*, 2016. doi:10.1109/TITS.2016.2560131.

[6] T. Xu, X. Liu, and X. Yang, "A novel approach for ship trajectory online prediction using bp neural network algorithm." *Advances in Information Sciences & Service Sciences*, vol. 4, no. 11, 2012.

[7] S. Sutulo, L. Moreira, and C. G. Soares, "Mathematical models for ship path prediction in manoeuvring simulation systems," *Ocean engineering*, vol. 29, no. 1, pp. 1–19, 2002.

[8] L. P. Perera, P. Oliveira, and C. Guedes Soares, "Maritime traffic monitoring based on vessel detection, tracking, state estimation, and trajectory prediction," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 13, no. 3, pp. 1188–1200, 2012.

[9] G. de Vries and M. van Someren, "Unsupervised ship trajectory modeling and prediction using compression and clustering," in *Proceedings BeneLearn*. Citeseer, 2009, pp. 7–12.

[10] M. E. Serrano, G. J. Scaglia, S. A. Godoy, V. Mut, and O. A. Ortiz, "Trajectory tracking of underactuated surface vessels: A linear algebra approach," *Control Systems Technology, IEEE Transactions on*, vol. 22, no. 3, pp. 1103–1111, 2014.

[11] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[12] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks*, vol. 2, no. 5, pp. 359–366, 1989.

[13] D. P. Mohapatra and S. Patnaik, *Intelligent Computing, Networking, and Informatics: Proceedings of the International Conference on Advanced Computing, Networking, and Informatics, India, June 2013*. Springer Science & Business Media, 2013, vol. 243.

[14] C. B. J. Tetreault, "Use of the automatic identification system (ais) for maritime domain awareness (mda)," in *OCEANS, 2005. Proceedings of MTS/IEEE*. IEEE, 2005, pp. 1590–1594.

[15] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result analysis of the nips 2003 feature selection challenge," in *Advances in neural information processing systems*, 2004, pp. 545–552.

[16] C. Boutsidis, A. Zouzias, M. W. Mahoney, and P. Drineas, "Randomized dimensionality reduction for-means clustering," *Information Theory, IEEE Transactions on*, vol. 61, no. 2, pp. 1045–1062, 2015.

[17] J.-G. Lee, J. Han, and K.-Y. Whang, "Trajectory clustering: a partition-and-group framework," in *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM, 2007, pp. 593–604.

[18] H.-M. Lu and C.-H. Lee, "A twitter hashtag recommendation model that accommodates for temporal clustering effects," *Intelligent Systems, IEEE*, vol. 30, no. 3, pp. 18–25, 2015.

[19] T. Mikolov, S. Kombrink, L. Burget, J. H. Černockỳ, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5528–5531.

[20] D. M. Kline and V. L. Berardi, "Revisiting squared-error and cross-entropy functions for training neural network classifiers," *Neural Computing & Applications*, vol. 14, no. 4, pp. 310–318, 2005.