CrossMark

THE JOURNAL OF NAVIGATION (2021), **74**, 156–174. © The Royal Institute of Navigation 2020
doi:10.1017/S0373463320000442

Vessel Trajectory Prediction Using Historical Automatic Identification System Data

Danial Alizadeh[1], Ali Asghar Alesheikh[1] and Mohammad Sharif[2]

[1](*Department of Geospatial Information Systems, Faculty of Geodesy and Geomatics Engineering, K. N. Toosi University of Technology, Tehran, Iran*)
[2](*Department of Geography, Faculty of Humanities, University of Hormozgan, Bandar Abbas, Iran*)
(E-mail: m.sharif@hormozgan.ac.ir)

For maritime safety and security, vessels should be able to predict the trajectories of nearby vessels to avoid collision. This research proposes three novel models based on similarity search of trajectories that predict vessels' trajectories in the short and long term. The first and second prediction models are, respectively, point-based and trajectory-based models that consider constant distances between target and sample trajectories. The third prediction model is a trajectory-based model that exploits a long short-term memory approach to measure the dynamic distance between target and sample trajectories. To evaluate the performance of the proposed models, they are applied to a real automatic identification system (AIS) vessel dataset in the Strait of Georgia, USA. The models' accuracies in terms of Haversine distance between the predicted and actual positions show relative prediction error reductions of 40·85% for the second model compared with the first model and 23% for the third model compared with the second model.

KEYWORDS

1. Trajectory. 2. Movement Models. 3. Ship Behaviour. 4. Automatic Identification System.

Submitted: 30 March 2020. Accepted: 28 July 2020. First published online: 26 August 2020.

1. INTRODUCTION. Maritime transportation is a major aspect of the global economy; almost 90% of global trade volume and 70% of global trade value are conducted through the marine environment. It is also energy efficient – only 7% of all energy consumed by transportation activities is related to this mode of transportation (Rodrigue et al., 2016). Therefore, increasing maritime safety and security is a paramount task. A study on marine collision reports revealed that 71% of human error is caused due to lack of situational awareness. Of such situational awareness errors, about 9% were related to the lack of awareness about future actions in the marine environment (Grech et al., 2002).

To enhance marine safety and security, since 2004 all vessels above a specific size have been obliged by the International Maritime Organization to be equipped with automatic identification systems (AISs) (SOLAS, 2003). AISs repeatedly broadcast both static

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

and dynamic information. Static information contains marine mobile service identifier (MMSI), vessel name, type, and dimensions, whereas dynamic information includes latitude, longitude, speed over ground (SOG), and course over ground (COG) (Series, 2010). AIS messages are reported at different update rates, which vary according to vessel speed and status, and range from 2 s (while the vessel operates at high speed) to 3 min (while the vessel is anchored) (Harati-Mokhtari et al., 2007). To enhance maritime situational awareness, various studies have been conducted using AIS data for detection of anomalous behaviour (Sidibé and Shu, 2017; Zhao and Shi, 2019), extraction of shipping route patterns (Fujino et al., 2018; Sheng and Yin, 2018), prediction of vessel trajectory (Pallotta et al., 2013; Wijaya and Nakamura, 2013; Borkowski, 2017; Duca et al., 2017; Gao et al., 2018; Graser et al., 2018; Mao et al., 2018), and traffic prediction (Xiao et al., 2017; Kim and Lee, 2018).

A necessary aspect of increasing maritime situational awareness is predicting target vessel trajectory effectively. In this regard, vessels should be able to predict the trajectories of other vessels which are located in their vicinity and adapt their movements to avoid any collision. Various models have been developed for predicting vessel trajectory so far. In general, the proposed models are either data-driven or model-driven. The model-driven models require that process parameters be estimated, which is not straightforward (Dalsnes et al., 2018). Besides, these models become complicated and need more time when non-linearity of the movement increases (Kawan et al., 2017). Data-driven models not only address the drawbacks of model-driven models, but also predict vessel trajectory for longer durations, which is required in collision avoidance. However, data-driven models should be able to cope with a large amount of data (Graser et al., 2019).

In this paper, vessel trajectory prediction is studied using the data-driven procedure. Predictions are carried out based on the assumption that vessels similar in pattern behaviour in the past will have similar patterns in the future. The sharp distinctions between this research and previous research are: (1) AIS messages are not considered individually, and the spatial correlation between AIS messages is taken into account. Therefore, the trajectory-based similarity is measured instead of point-based measure. (2) Distances between the target trajectory and the most similar trajectory are no longer considered constant during the prediction.

The structure of this paper is as follows. Section 2 provides an overview of the previous studies in vessel trajectory prediction. Section 3 introduces three prediction models. In Section 4, the proposed models are experimentally evaluated by preforming them on real AIS data sets, and the results are presented. Finally, Section 5 concludes the key remarks of the study and suggests promising research topics for future work.

2. RELATED WORK.

Various models have been developed for trajectory prediction, which are generally classified as model-driven or data-driven approaches. The Gaussian mixture trajectory prediction (GMTP) model presented by Qiao et al. (2015) revealed the complex motion patterns based on Gaussian mixture models. They calculated the probability distribution of different types of motion patterns. The possible trajectories of input trajectory were then determined by Gaussian process regression. The results revealed that GMTP could improve prediction accuracy compared with the Gaussian process regression and the Kalman filter. Perera and Soares (2010) estimated ship position, velocity, and acceleration using an extended Kalman filter. Model-driven approaches require a mathematical

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

model to be built for each trajectory, therefore they are not very suitable for dynamic environments (Tang et al., 2019).

Data-driven approaches have made remarkable achievements in trajectory prediction. First, Gan et al. (2016) used the K-means algorithm to group AIS data. Secondly, the clusters achieved by the first step were used to train the artificial neural network (ANN) to predict the vessels' trajectories. The results showed a 70% prediction accuracy. Tang et al. (2019) constructed a neural network with two long short-term memory (LSTM) layers, which can observe the first 10 min of the vessel's state to predict the location of the vessel after 20 min. The experimental comparison revealed that the model performs better than the Kalman filter and backpropagation neural network. Nishizaki et al. (2018) trained a support vector machine (SVM) to predict the next course of vessels that exited from the Nakanose traffic route, known as an exit point. They used the data related to the exit point, 5 min before the exit point, and 10 min before the exit point to train the SVM model. The result shows prediction accuracy of 87·9%, 80·3%, and 77% for the data related to exit point, 5 min, and 10 min before the exit point, respectively. However, the main drawback of their work was the disproportion of training data, which means the trained data mostly belonged to a particular class. Graser et al. (2018) applied three models – linear prediction, learned statistical, and similarity search – to vessel trajectories. The results proved that the similarity search approach performed accurately compared with the other two models for prediction. To predict the next location of the target point, Wijaya and Nakamura (2013) extracted the most similar points to the target point based on the assumption that spatial distances between the target point and the most similar point to it remain constant during the prediction. Long-term prediction is important in the maritime environment to gain sufficient time for the actions needed to control vessel movements (Kim and Lee, 2018). However, previous works either carried out short-term prediction (Wijaya and Nakamura, 2013; Graser et al., 2018) or obtained unsatisfactory results in long-term prediction (Zorbas et al., 2015). To overcome these problems, three distinct models are proposed based on similarity searches that can predict vessels' trajectories for short and long durations.

3. METHODOLOGY.

This section is dedicated to the pre-processing of raw AIS data and proposes three distinct models for predicting vessels' positions. All three models utilise similarity searches to predict the geographic location of a vessel at the moment ahead. The schematic diagram of the methodology presented in this work is illustrated in Figure 1.

3.1. *Data pre-processing.*

Because of high redundancy in AIS data and complexity in extracting vessel movement patterns from raw AIS data, the data must be pre-processed. First, because only moving vessels are important, vessels with SOG values less than 1 knot, including anchored vessels, are eliminated from the data set. Second, to compose trajectories related to each vessel, the entries are grouped based on their MMSIs. In the next step, each unique trajectory is divided into individual trips, which are called voyages, if the gap observation between two consecutive AIS messages is greater than a time threshold (ε_1). Finally, it is required that all records of each voyage be resampled into T min. The parameters of the minimum gap observation between two consecutive AIS records (ε_1), bins for time resampling (T), minimum duration of each voyage (ε_2), and minimum SOG (ε_3) were set at 5 min, 10 min, 120 min, and 1 knot, respectively. As mentioned before, AIS messages are transmitted every 2 s to 3 min. Therefore, continuous trajectories have been split with gap observation (ε_1) of 5 min.

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

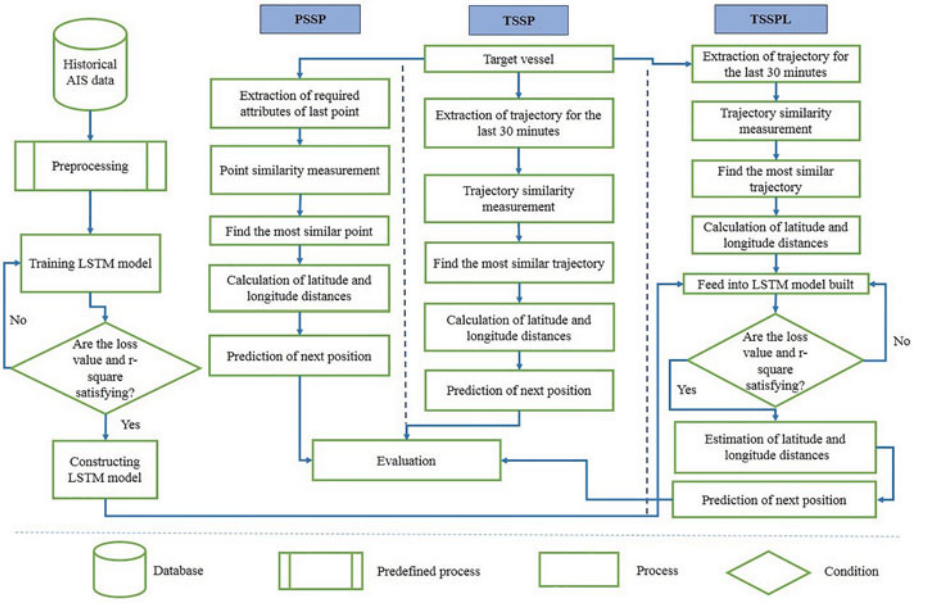


Figure 1. Flow diagram of PSSP, TSSP, and TSSPL models.

3.2. *Point-based similarity search prediction (PSSP).* To predict the next locations of the target vessels, the most similar vessel to the target vessel is considered. In the PSSP model, each AIS record is considered individually as a point and the distances between the last point of the target vessel (i.e., target point) and all the historical points are measured in terms of spatial, speed, and course variables. After extracting the most similar point to the target point from the historical points (i.e., the most similar extracted point), the next locations of the target vessel will be predicted using the most similar extracted point as follows. The main concept of this model has been inspired by Wijaya and Nakamura (2013); however, it has been modified here to increase the prediction accuracy.

3.2.1. *Similarity measure phase.* To measure how similar two points are in terms of spatial, speed, and course criteria, three similarity distance functions are defined.

3.2.1.1. *Spatial similarity measure.* The Haversine distance measures the spatial similarity of two points. It determines a great-circle distance between two points on a sphere using their latitudes and longitudes along the surface. As the Earth is nearly spherical, Haversine distance provides a good approximation for the shortest distance involving points over the Earth's surface, as is expressed by Equation (1) (Valsamis et al., 2017). Smaller value of the calculated Haversine distance between two points indicates more spatial similarity.

$$d_d = 2r * \sin^{-1}\left(\sin^2\left(\frac{\Delta\text{Latitude}}{2}\right) + \cos(\text{Latitude}_1)\cdot\cos(\text{Latitude}_2) \cdot \sin^2\left(\frac{\Delta\text{Longitude}}{2}\right)\right) \quad (1)$$

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

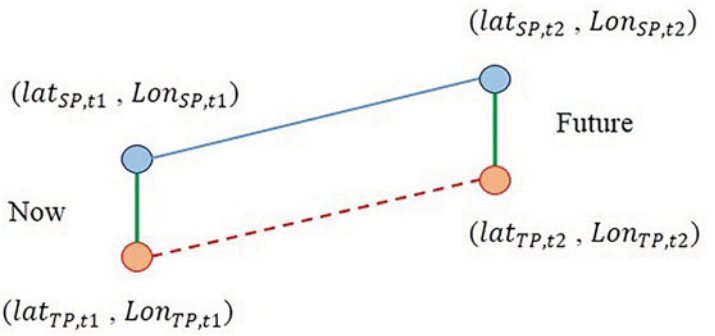


Figure 2. The location prediction approach proposed in PSSP model.

where r is the radius of the Earth (6371 km). d_d is the calculated Haversine distance between two points. Δ Latitude and Δ Longitude denote differences between the latitudes and longitudes of two points, respectively. Latitude_1 shows the latitude of the first point, and Latitude_2 is the latitude of the second point.

3.2.1.2. *Speed similarity measure.* To measure how similar two points are in terms of speed (SOG), the absolute difference of the speed of two points (d_s) is calculated using Equation (2). The smaller the value of the absolute difference, the greater the similarity between two points is.

$$d_s = |\text{SOG}_1 - \text{SOG}_2| \tag{2}$$

where SOG_1 and SOG_2 denote the speeds of the first and second points, respectively.

3.2.1.3. *Course similarity measure.* To measure how similar two points are in terms of their course (COG), the absolute difference of the course of two points (d_c) is calculated using Equation (3). The smaller the absolute difference, the greater the similarity between the two points is.

$$d_c = |\text{COG}_1 - \text{COG}_2| \tag{3}$$

where COG_1 and COG_2 indicate the COG of the first and second points, respectively.

The values achieved from Equations (1)–(3) are normalised to the scale of 1 using the scikit-learn Python package due to the differences in their units of measurement. Finally, the most similar point to the target point in terms of spatial distance, speed, and course is determined using the overall weighted value of three similarity measure functions (Equation (4)).

$$D_t = w_d \times d_d + w_s \times d_s + w_c \times d_c, \tag{4}$$

where d_d, d_s, and d_c stand for the normalised spatial, speed, and course similarity measures, respectively, and w_d, w_s, and w_c are their corresponding weights. D_t is the total similarity distance. The weight of each similarity measure criterion is determined based on the degree of its importance.

3.2.2. *Location prediction phase.* The most similar point extracted from historical data is used to predict the next location of the target point under the assumption that the spatial distances remain constant throughout the prediction. The schematic procedure of location prediction using PSSP is depicted in Figure 2.

In Figure 2, $(\text{Lat}_{TP,t1}, \text{Lon}_{TP,t1})$ represents the coordinate of the target point at the current time and is known. $(\text{Lat}_{TP,t2}, \text{Lon}_{TP,t2})$ represents the coordinate of the target point at the

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

future time, which is unknown and should be predicted. Likewise, $(\text{Lat}_{SP.t1}, \text{Lon}_{SP.t1})$ and $(\text{Lat}_{SP.t2}, \text{Lon}_{SP.t2})$, which are both known, stand for coordinates of a similar point at the current and future times, respectively. The next location of the target point will be predicted using the most similar extracted point by Equations (5) and (6).

$$\begin{aligned} \Delta\text{Longitude} &= \text{Lon}_{TP.t1} - \text{Lon}_{SP.t1} \\ \Delta\text{Latitude} &= \text{Lat}_{TP.t1} - \text{Lat}_{SP.t1} \end{aligned} \tag{5}$$

$$\begin{aligned} \text{Lon}_{TP.t2} &= \Delta\text{Longitude} + \text{Lon}_{SP.t2} \\ \text{Lat}_{TP.t2} &= \Delta\text{Latitude} + \text{Lat}_{SP.t2} \end{aligned} \tag{6}$$

If the target and the most similar extracted points move in the same direction, the PSSP model performs appropriately. However, in our experiments, if these objects move in different directions, the prediction accuracy obtained from the PSSP model will dramatically decrease. Therefore, to improve the accuracy of the PSSP model, a constraint is imposed that declares the absolute difference between the azimuth angle of the target point trajectory and the azimuth angle of the similar point trajectory, which must be less than 22·5°. To recognise which direction the vessel is moving, the start and end points are connected, then the azimuth angle of this connected line is measured by Equations (7) and (8) as illustrated in Figure 3. In Figure 3, the blue line demonstrates a particular trajectory that we incline to measure its azimuth angle, and the red line shows the line that connects the start and end points of the trajectory.

$$g_{\text{Start-End}} = \tan^{-1}\left|\frac{X_e - X_s}{Y_e - Y_s}\right| \tag{7}$$

$$\text{Azimuth}_{\text{Start-End}} = \begin{cases} g_{\text{Start-End}} & \text{if } \Delta X > 0 \text{ and } \Delta Y > 0 \\ 180 - g_{\text{Start-End}} & \text{if } \Delta X > 0 \text{ and } \Delta Y < 0 \\ 360 - g_{\text{Start-End}} & \text{if } \Delta X < 0 \text{ and } \Delta Y > 0 \\ 180 + g_{\text{Start-End}} & \text{if } \Delta X < 0 \text{ and } \Delta Y < 0 \end{cases} \tag{8}$$

where $g_{\text{Start-End}}$ denotes the bearing angle that is calculated from the coordinates of the start and end points of the trajectory. Consequently, the azimuth angle is measured using the bearing angle (Equation (8)). The 16-point compass rose is used to compare the directions of trajectories based on their azimuth angle. Because the measured angle of each section is 22·5°, the absolute difference of the trajectories is considered less than 22.5°.

3.3. *Trajectory-based similarity search prediction (TSSP).* In the TSSP model, the AIS records corresponding to each voyage are considered as a trajectory, and similarity is measured between trajectories instead of individual points. To predict the future locations of the target vessel, the most similar trajectory to the trajectory of the target vessel is extracted from the historical data set. The future locations of the target vessel will be predicted by the extracted trajectory as follows.

3.3.1. *Similarity measure phase.* Various similarity measure methods have been introduced or developed based on either time series such as Euclidean distance, edit distance, longest common subsequence, and dynamic time warping (DTW) or geometric shape matching techniques such as Frechet distance and Hausdorff distance (Sharif et al., 2019). DTW can be applied to trajectories that have unequal lengths, and is applicable for matching trajectories if information is lacking, for managing local time-shifting, or for addressing

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

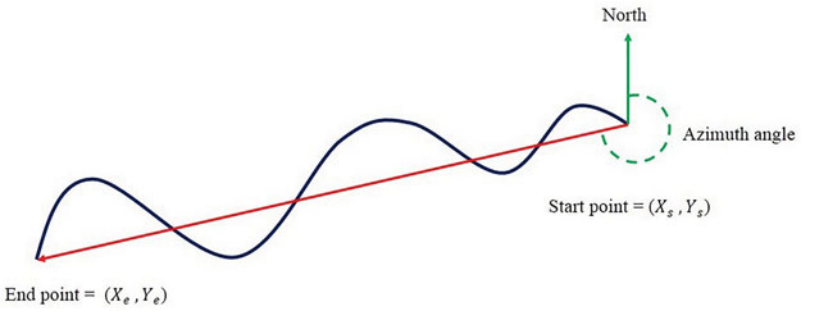


Figure 3. The azimuth angle of a trajectory.

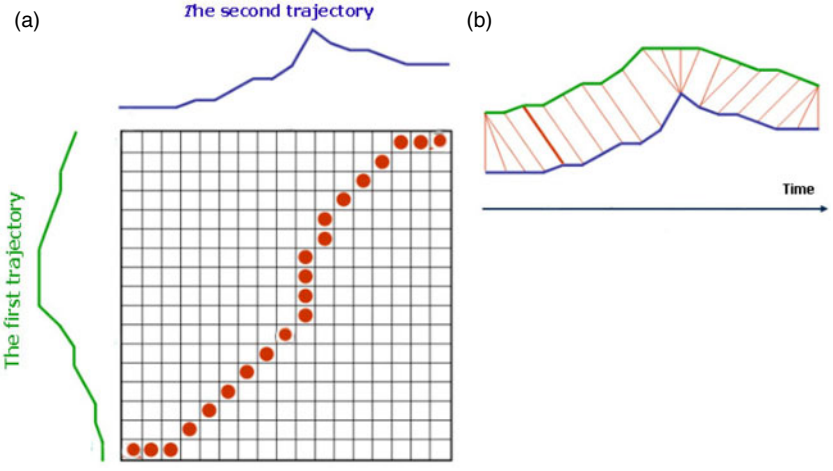


Figure 4. DTW mechanism: (a) elastic achieved by DTW; (b) trajectory alignments.

parametric data. DTW wraps the first trajectory temporally to minimise the distance to the second trajectory. Elements of the first trajectory then become aligned to the elements of the second trajectory in an optimal way (Sharif and Alesheikh, 2017), as shown in Figure 4. DTW performs based on elastic shifting of the time axis; therefore, time is not a major component to be investigated.

DTW is applied to measure how similar two trajectories of distinct voyages are in terms of spatial distance, SOG, and COG. These three DTW distances are in a range from 1 to 2 because of the variety of their units of measurement. The overall weight of ranged values achieved from three DTW results is then calculated by Equation (4) to recognise the trajectory most similar to the target trajectory based on the three criteria.

3.3.2. *Location prediction phase.* To predict the next location of the target vessel, the spatial distances between the last point of the target trajectory and the point of the detected trajectory, which is the most similar to the last point of the target vessel's trajectory, are calculated using Equation (5). The next location of the target vessel will be predicted using Equation (6) based on the assumption that calculated spatial distances remain constant within the duration of the prediction.

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

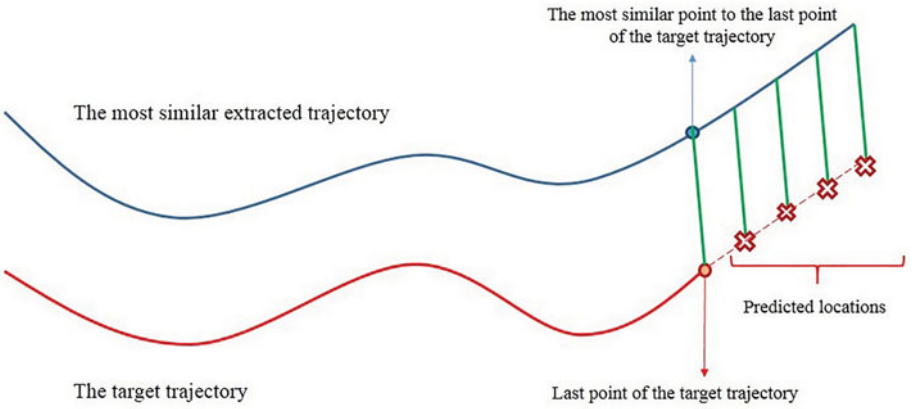


Figure 5. Location prediction proposed in the TSSP model.

The TSSP model is illustrated in Figure 5. The red line shows the target trajectory, and the blue line indicates the most similar trajectory extracted from the historical trajectories. The green lines represent the spatial distances that are considered constant throughout the prediction. The red crosses show the predicted locations.

3.4. *Trajectory-based similarity search prediction using LSTM (TSSPL).* Both PSSP and TSSP models assume that the distance between the target trajectory and the trajectory most similar to it remains constant over the prediction duration. However, vessel movement is affected by external contexts, such as wind, waves and sea currents (Kawan et al., 2017; Sharif and Alesheikh, 2018; Kaffash-Charandabi et al., 2019). In addition, the maritime navigation environment is a complex system that varies over time (Tang et al., 2019). Consequently, it is highly likely that the spatial distances that have been considered constant in both previous models will change, especially when two trajectories are either converging or diverging. To tackle this issue, an LSTM model is built upon the historical AIS data that can predict spatial distances over the duration of the prediction. Tang et al. (2019) developed a LSTM model to predict the future location of vessels. In the proposed model, the first 10 min of a vessel's state, including its location, speed, and course, were observed and inserted into the model as the input sequence, and the location of the vessel at the 20th min was predicted as the output of the LSTM model. However, the similarity search was not used in the LSTM model.

In the TSSPL model, in contrast to the TSSP model, the distance between the target trajectory and the trajectory most similar to it is no longer constant. Therefore, the DTW method is used to find the most similar trajectory. After identifying the trajectory most similar to the target trajectory, spatial distance is calculated between the point on the most similar trajectory that has similarity with the last point of the target trajectory (Equations (5)). In addition, distances are calculated by Equation (5) between the three points before the last point on the target trajectory and their corresponding points on the most similar trajectory. These distances are inserted into the LSTM model as input. The spatial distances predicted by the LSTM model are used to predict the next location at each time interval using Equation (6). An overview of the TSSPL model is illustrated in Figures 6 and 7, for diverging and converging trajectories, respectively. The solid green lines show the calculated distances, the broken green lines show distances predicted using the LSTM model, and the red crosses represent the predicted locations.

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

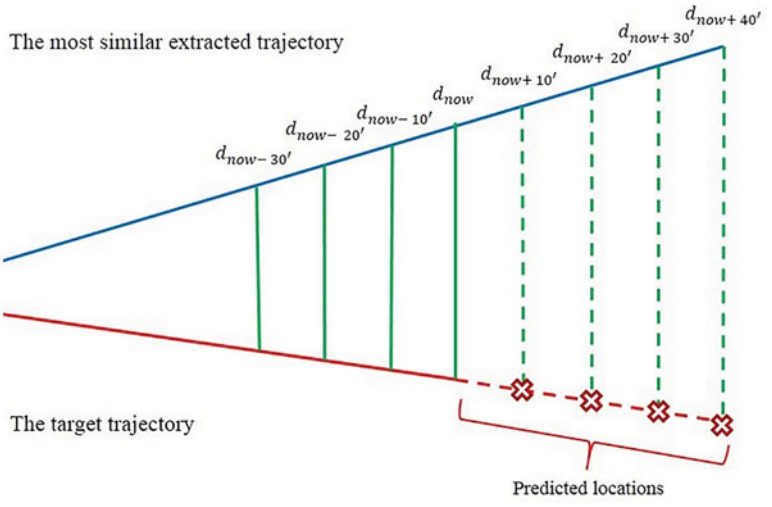


Figure 6. The TSSPL model, when two trajectories are diverging.

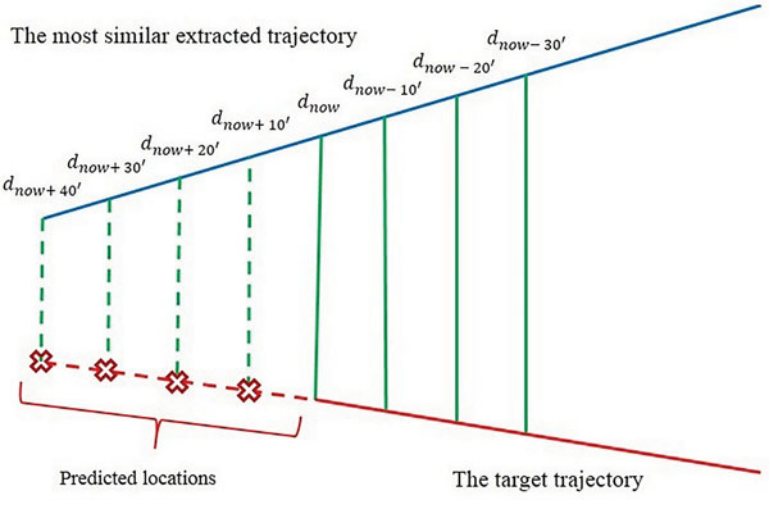


Figure 7. The TSSPL model, when two trajectories are converging.

3.4.1. *Training and implementation of LSTM.* LSTMs are special types of recurrent neural networks (RNNs). The basic formula of RNN is shown in Equation (9), where h_t stands for the new state, x_t indicates the new input, and f shows an activation function with parameters W. RNNs take the previous state represented by h_{t-1} into account. This characteristic makes them efficient, especially when it comes to modelling sequence problems (Karatzoglou et al., 2018). This methodology applies to the present work because distance value at the future timestamp depends on their values at the previous timestamp. However, RNNs suffer from a drawback called vanishing gradients in the backpropagation process (Hochreiter et al., 2001). In this context, LSRMs were introduced by Hochreiter and

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

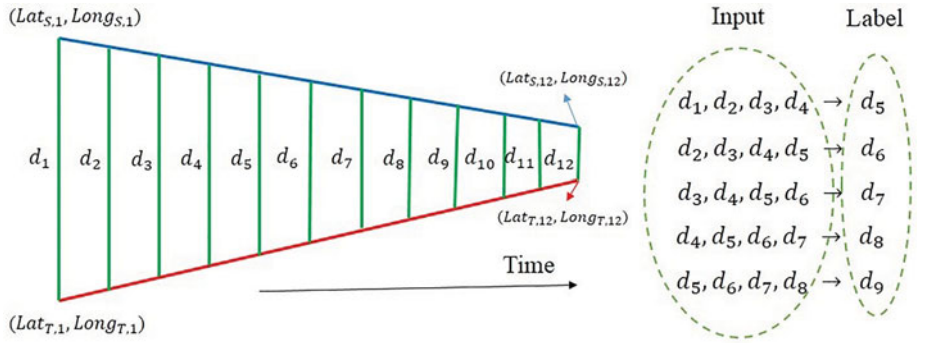


Figure 8. Preparing data to feed into the LSTM model.

Schmidhuber (1997), and further improved by Gers et al. (1999). This was achieved by providing a memory storing and managing mechanism called gated cell. With the aid of this mechanism, LSTMs can store inputs and previous states for a long duration (Karatzoglou et al., 2018). A standard notation of LSTMs is defined by Equations (10) and (11).

$$h_t = f_W(h_{t-1}.x_t) \tag{9}$$

$$h_t = f_{\text{act}}(c_t) \cdot o_t \tag{10}$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot f_{\text{act}}(W_x \cdot x_t + W_h \cdot h_{t-1} + b) \tag{11}$$

Equation (10) indicates how the current state can be affected by the output gate represented by $o_t.f_t$, i_t, f_{act}, W_x, and W_h stand for the forget gate, gate value, activation function, input hidden, and hidden-hidden layer weight matrices, respectively.

To train the network, historical pre-processed data is used. The most similar trajectory to each trajectory from the database is extracted within the database, then spatial distances are calculated between each pair of trajectories that are similar to each other. These calculated distances are labelled to train the network and adjust the weight matrix by comparing the error between the predicted value and expected value (Figure 8).

The structure of the LSTM model is illustrated in Figure 9. h_0, h_1, h_2, h_3, h_4 and x_0, x_1, x_2, x_3, x_4 represent the hidden state and input, respectively. Finally, the last hidden state passes the sigmoid function to achieve the output. The output derived from a layer of the LSTM model may take a larger value. When this output is inserted into further layers, it can be transformed into even greater values that make the model computationally uncontrollable. This problem is solved with the aid of the sigmoid function by squashing a real value into a fixed interval (between 0 and 1). The input of our LSTM model is defined using a matrix with dimension of 4×2 using Equation (12).

$$\text{Input} = \begin{bmatrix} d_1 \\ d_2 \\ d_3 \\ d_4 \end{bmatrix} = \begin{bmatrix} (\text{Long}_{S.1} - \text{Long}_{T.1}) \cdot (\text{Lat}_{S.1} - \text{Lat}_{T.1}) \\ (\text{Long}_{S.2} - \text{Long}_{T.2}) \cdot (\text{Lat}_{S.2} - \text{Lat}_{T.2}) \\ (\text{Long}_{S.3} - \text{Long}_{T.3}) \cdot (\text{Lat}_{S.3} - \text{Lat}_{T.3}) \\ (\text{Long}_{S.4} - \text{Long}_{T.4}) \cdot (\text{Lat}_{S.4} - \text{Lat}_{T.4}) \end{bmatrix} \tag{12}$$

3.4.2. *Evaluation of LSTM.* To train the LSTM model, historical trajectories are divided into two sets: a primary training set and a test set, in a ratio of eight to two. The

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

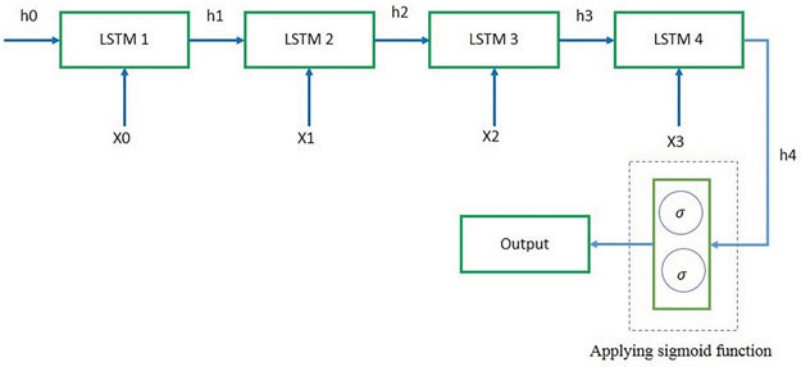


Figure 9. Overview of built LSTM model.

initial training set is then divided into two sets: a training set and a validation set, in a ratio of seven to three. The model is trained using the data in the training set. The validation set is used during the training process to validate the model for enhancing its generalisation ability. The test data are used to achieve the prediction result and assess the performance of the model.

To evaluate the performances of the LSTM model, mean squared error (MSE) and r-squared are used as the evaluation metrics (Equations (13) and (14)).

$$\text{MSE} = \frac{1}{n}\sum_{i=0}^{n}(y_i - \widehat{y_i})^2 \tag{13}$$

$$R^2 = 1 - \frac{\sum_{i=0}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=0}^{n}(y_i - \overline{y_i})^2} \tag{14}$$

where y_i, $\widehat{y_i}$, and $\overline{y_i}$ represent the actual value, predicted value, and the average of all real values, respectively. MSE indicates how close the actual values are to the model's predicted values. MSE is an absolute measure of fit, whereas R^2 is a relative measure of fit.

4. EXPERIMENTAL STUDY. The Strait of Juan de Fuca or Georgia on the west coast of the USA, which is located at latitude 48N to 48·6N and longitude −125·9 W to −125·2 W, is selected as the area of interest (Figure 10). The entrance of the Strait of Juan de Fuca is a congested maritime environment with hundreds of ships passing.

Two different datasets provided by Marinecadastre (2019) are used in this study: one from March 2017 for creating the database and the other from February 2017 for the evaluation of the proposed models. In Table 1, the information extracted from raw AIS data and that extracted from pre-processed data are compared. The considerable difference between these two datasets is that only mobile vessels are preserved.

5. RESULTS AND DISCUSSION. To evaluate the applicability of the PSSP, TSSP, and TSSPL models, the future positions of the 89 distinct vessels at time intervals of 10, 20, 30, and 40 min have been predicted. To assess the models' accuracies, the Haversine distance between the actual position and the predicted position is calculated by Equation (1).

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

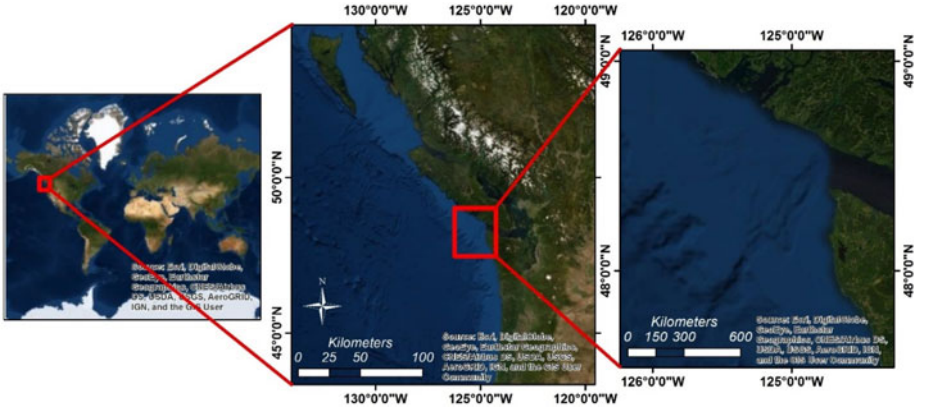


Figure 10. The Strait of Juan de Fuca, west coast of USA.

Table 1. Comparison between raw data and pre-processed data related to original and evaluation datasets.

	Data type	Total duration of voyages (m)	Number of records	Number of vessels	Number of extracted voyages
Original dataset	Raw AIS data	54,362,564	29,917,046	5107	1,115,063
	Pre-processed data	5,497,230	562,381	4087	12,658
Evaluation dataset	Raw AIS data	47,455,927	26,241,972	4491	980,452
	Pre-processed data	4,733,000	485,904	3452	12,604

Table 2. Location prediction errors (in kilometres).

Proposed model	Prediction duration (10 min)	Prediction duration (20 min)	Prediction duration (30 min)	Prediction duration (40 min)
PSSP	0·842	1·700	2·582	3·501
TSSP	0·556	1·040	1·457	1·854
TSSPL	0·391	0·779	1·142	1·569

The calculated distance is defined as the error of prediction that indicates the distance on the Earth's surface between real and predicted positions. According to the importance of the spatial dimension (w_d), the weight is set at 0·6. The weights of the speed (w_s) and course (w_c) criteria are set at 0·2.

Table 2 provides the average prediction error of the 89 target vessels at different prediction durations. As can be seen, in the TSSPL model prediction error decreases by 53·6%, 54·2%, 55·8%, and 55·2% at the time intervals of 10, 20, 30, and 40 min, respectively. Compared with the PSSP model, the TSSPL model shows its efficiency, especially in long-term prediction. As can be inferred from Table 2, the prediction errors for all three models increase during the prediction time. In the PSSP and TSSP models, this occurs because, as time passes, the spatial distances remain constant. For the TSSPL model, the reason can be attributed to the fact that spatial distances are predicted with the help of previous spatial distances predicted by the LSTM model. Therefore, the error of predicted spatial distance in each timestamp would affect the error of predicted spatial distance in the next timestamps.

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

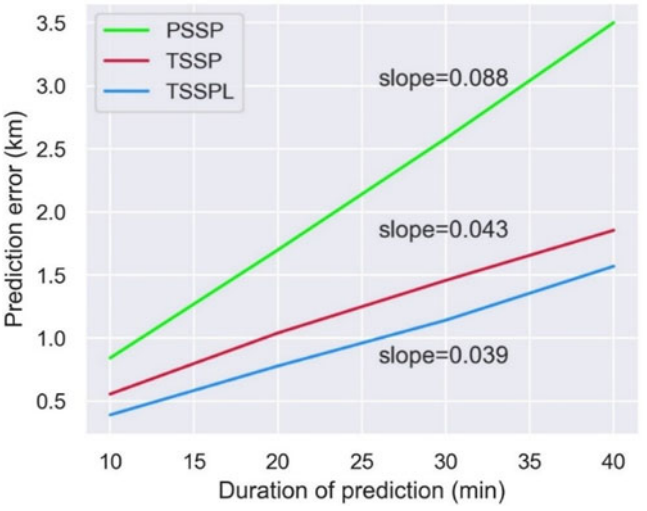


Figure 11. Comparison of performances of trajectory prediction models over time.

A general comparison between the accuracy of the models is demonstrated in Figure 11. The vertical and horizontal axes show the values of the prediction error and the time intervals of the prediction, respectively. The slope of the TSSPL model, which indicates error changes over time, is less than that of the other two models. This assures more reliable prediction, especially in the longer term; the slopes reveal an increase in the prediction error as the duration of prediction continues.

Compared with similar researches, the proposed models perform better in terms of prediction accuracy. The PSSP model achieved accuracy of 0·842, 1·7, 2·582, and 3·501 kilometres for the prediction times of 10, 20, 30, and 40 min, respectively. In addition, prediction error has decreased significantly by 40·85% for the TSSP model compared with the PSSP model and by 23% for the TSSPL model compared with the TSSP model. The results obtained justify the applicability of the proposed models for the long-term prediction that is highly required for vessel traffic monitoring and collision avoidance. However, relying solely on the quantitative results reported is not reliable because each research study conducts its evaluations on different regions and does not consider the spatial variability within and between regions (Graser et al., 2019).

Error distribution analysis has been conducted to assess the models' performance quantitatively. Figure 12 shows the frequency histograms of the error predictions for the next 10, 20, 30, and 40 min. The vertical and horizontal axes indicate the number of predictions for each error and the value of prediction error, respectively. The red line refers to the mean value (μ) of the histogram, and the blue area demonstrates the interval $\mu - \sqrt{2}\sigma$ and $\mu + \sqrt{2}\sigma$, where σ stands for the standard deviation. Concerning the Chebyshev's inequality (Alsmeyer, 2011), at least 50% of values are located within the blue area. The mean values of the TSSPL model are lower than the PSSP and TSSP models for all prediction times.

The constructed LSTM model has been evaluated using metrics described in Section 3.4.2. Table 3 provides quantitative results of the model's performance on the training, validation, and test data sets. As can be seen in Table 3, MSEs related to the training

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

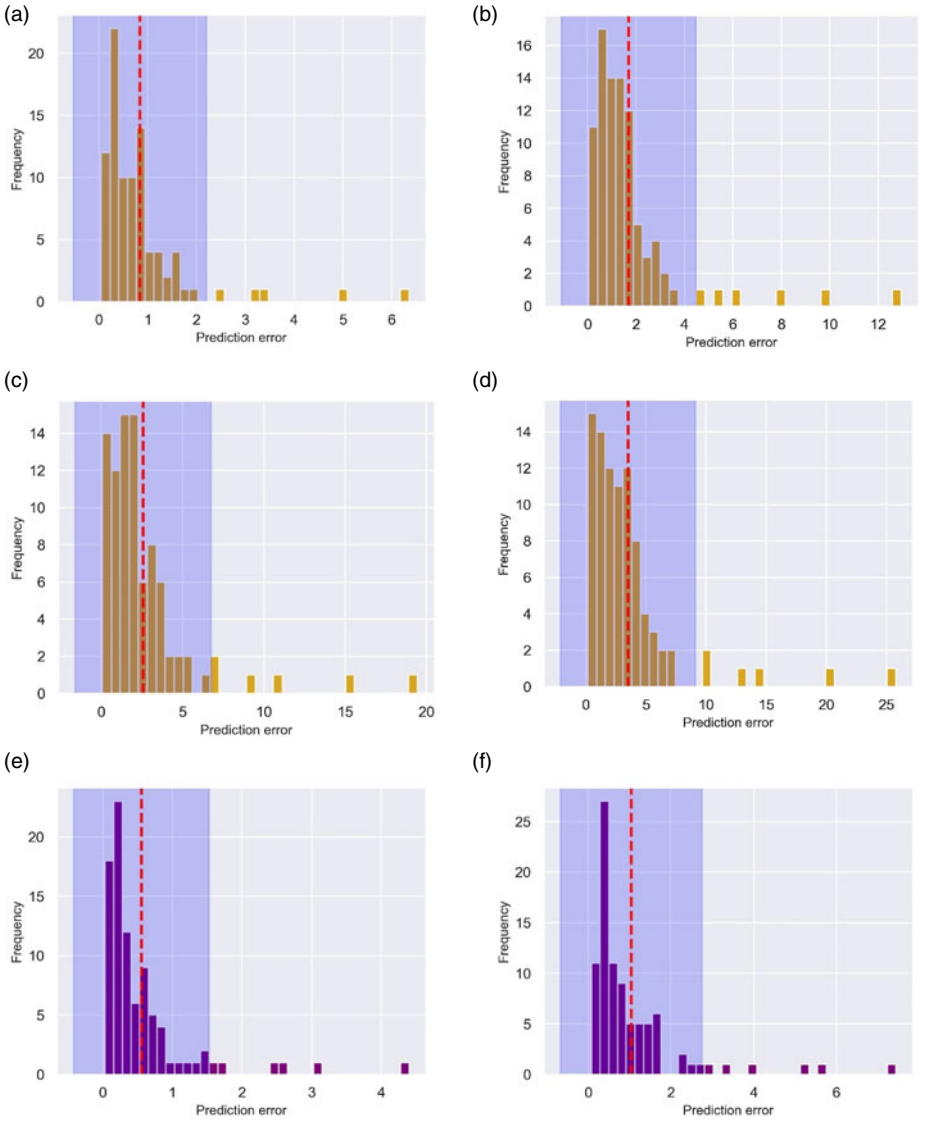


Figure 12. Frequency histograms of error prediction: (a)–(d) for the PSSP model, (e)–(h) for the TSSP model, and (i)–(l) for the TSSPL model for next 10, 20, 30, and 40 min, respectively.

and validation data sets are very close to each other, which means the LSTM model has been trained. In addition, MSE corresponding to the test data set is very close to zero, indicating that the model is perfectly able to predict new data. The values of the two indices during the 1,000 epochs are shown in Figure 13.

In the PSSP model, three different constraints moving in the same direction were set up on the target trajectory and the most similar extracted trajectory. The performance of the PSSP model was evaluated by applying each constraint based on prediction error and the best constraint was chosen as follows.

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

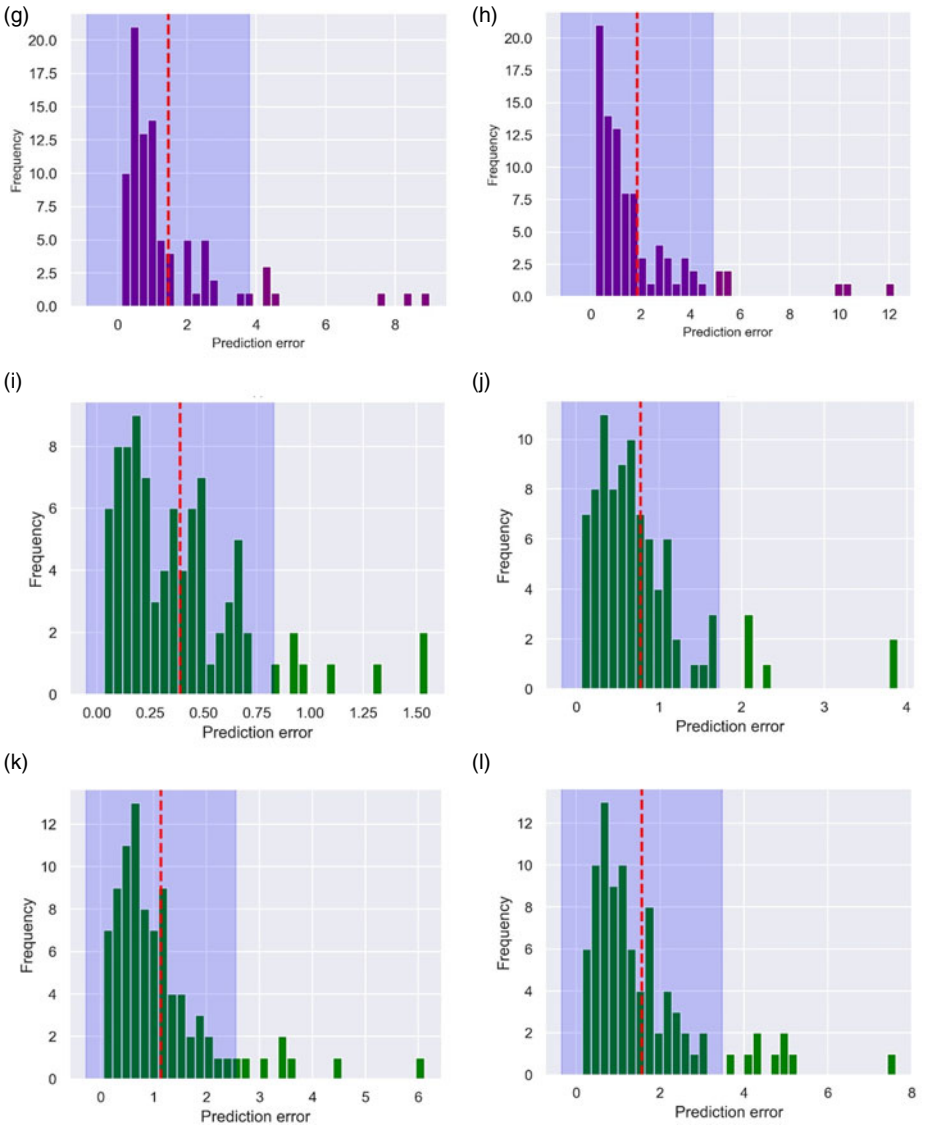


Figure 12. Continued.

Table 3. Evaluation of the model performance.

Data set	MSE	R^2
Training	5·9081e-05	0·9853
Validation	6·2978e-05	0·9846
Test	5·8140e-05	0·9828

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

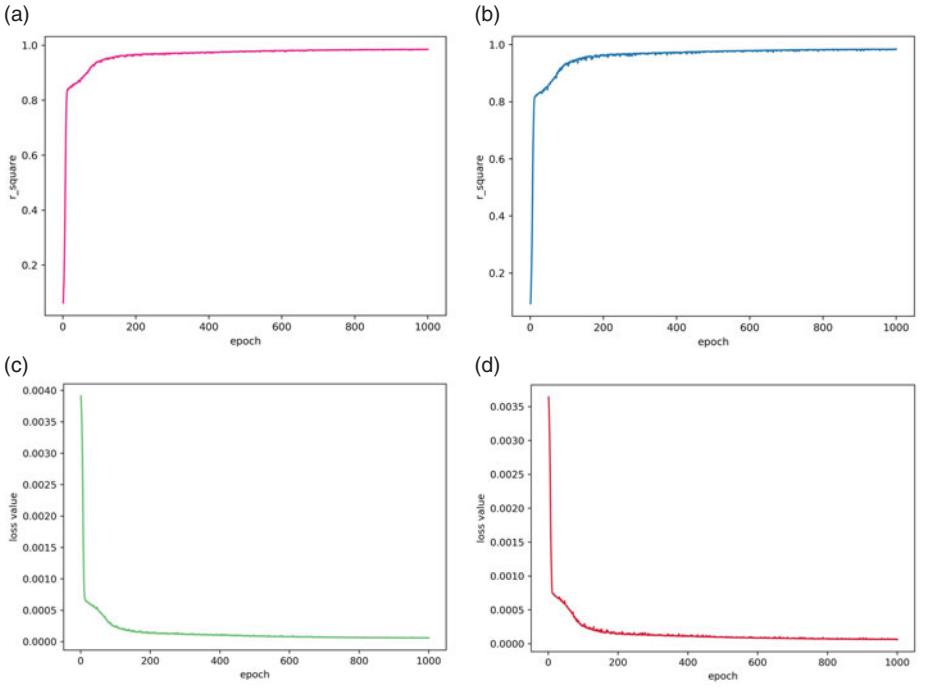


Figure 13. Quantitative result of the model's performance: (a) r-squared values on training dataset; (b) r-squared values on validation data set; (c) MSE values on training data set; (d) MSE on validation data set.

- *Azimuth constraint*: This constraint declares that the absolute difference between the azimuth angle of the target point trajectory and the azimuth angle of the similar point trajectory is less than 22·5°.
- *COG constraint*: As COG determines the actual direction of the vessel, this constraint states that the mean values of COG for the target vessel and the most similar extracted vessel must be the same. In other words, both values must be either negative or positive numbers. If the COG value of two trajectories is positive, it means that they are moving towards the east. Likewise, when the COG values of two trajectories are negative, it means that they are moving towards the west.
- *Longitude constraint*: Another way to distinguish the vessel direction is to calculate the difference between the longitudes of the first and last points of the trajectory, which is called a longitude difference. This constraint expresses that the sign of longitude difference must be the same for the target trajectory and the most similar extracted trajectory.

Evaluation of the performance of each imposed constraint is provided in Table 4. As can be inferred, the azimuth constraint shows the best performance in terms of prediction error compared with the other constraints.

To assess the robustness of each defined constraint, 353 distinct trajectories are inserted into the PSSP model as input trajectories. The number of trajectories that the model can predict is then counted. Robustness value is calculated by Equation (15).

$$\text{Robustness value} = \frac{\text{Predicted trajectories}}{\text{Input trajectories}} \times 100 \tag{15}$$

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

Table 4. Evaluation of the PSSP model performance with different constraint (in kilometres).

Constraint	Prediction duration (10 min)	Prediction duration (20 min)	Prediction duration (30 min)	Prediction duration (40 min)
Azimuth	0·783	1·599	2·421	3·302
COG	1·409	2·842	4·270	5·733
Longitude	1·324	2·677	4·014	5·381

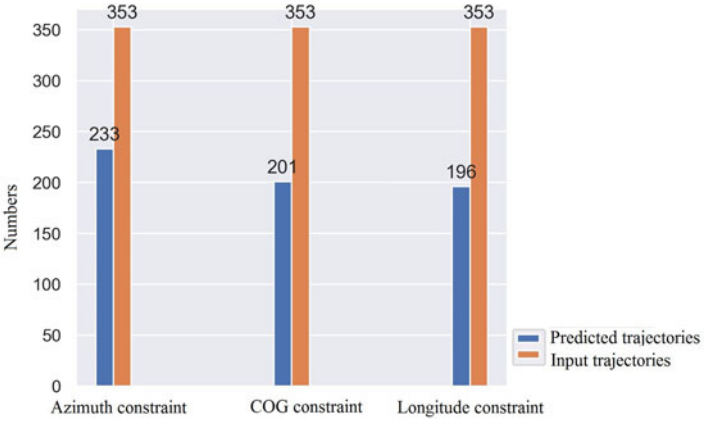


Figure 14. Robustness test of different constraints.

Robustness values for azimuth, COG, and longitude constraints are 66%, 56%, and 55%, respectively. Figure 14 compares the applicability of each constraint. The orange and blue bars, respectively, refer to the input trajectories and the number of trajectories that the PSSP model can predict.

6. CONCLUSION. This research contributes to the prediction of vessels' movements for short and long time durations. Three distinct similarity-based models, namely PSSP, TSSP, and TSSPL, are proposed. In the PSSP model, trajectories are considered as individual points and the most similar point to the target point is found. The prediction is performed by using the most similar point found from the database. However, two sampling points from two distinct trajectories are remarkably the same, whereas the moving objects' behaviours are in different directions. To find out how similar the moving objects' behaviours will be in the future, the similarity between the two trajectories is measured by using the TSSP model. This approach results in a remarkable increase in the accuracy of predictions compared with the PSSP model. The PSSP and TSSP models are based on the assumption that the distance between the target trajectory and the trajectory most similar to it will remain constant. However, because of the dynamic maritime context, this distance is changing within the duration of the prediction. To address this issue, the third proposed model, TSSPL, predicts the mentioned distance using the LSTM model constructed on historical trajectories. Applying the TSSPL model results in a steady decrease in prediction error compared with the TSSP model. The results reveal that the third model is more accurate than the other two models, and it can be utilised to predict a vessel's trajectory

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

effectively, particularly in long-term prediction. This research provides valuable information for coastguards, port management, and insurance companies, supporting them to make their decisions based on knowledge. Moreover, the three proposed models not only show applicability in terms of vessel trajectory prediction but they can also be used in other applications, for example, prediction of hurricanes, and human, and animal trajectories. For further studies, the authors suggest considering the inclusion of contextual information in the similarity measures and location predictions of trajectories. Environmental contexts such as wind, wave, and weather conditions have immense effects on the movement of vessels so that they may deviate from their normal path.

REFERENCES

Alsmeyer, G. (2011). Chebyshev's inequality. In: Lovric, M. (ed.). *International Encyclopedia of Statistical Science*. Berlin, Heidelberg: Springer.

Borkowski, P. (2017). The ship movement trajectory prediction algorithm using navigational data fusion. *Sensors*, 17, 1432.

Dalsnes, B. R., Hexeberg, S., Flåten, A. L., Eriksen, B.-O. H. and Brekke, E. F. (2018). The Neighbor Course Distribution Method with Gaussian Mixture Models for AIS-Based Vessel Trajectory Prediction. *2018 21st International Conference on Information Fusion* (*FUSION*). IEEE, pp. 580–587.

Duca, A. L., Bacciu, C. and Marchetti, A. (2017). A K-Nearest Neighbor Classifier for Ship Route Prediction. *OCEANS 2017-Aberdeen*. Aberdeen, UK: IEEE, pp. 1–6.

Fujino, I., Claramunt, C. and Boudraa, A.-O. (2018). Extracting Courses of Vessels from AIS Data and Real-Time Warning Against Off-Course. *Proceedings of the 2nd International Conference on Big Data Research (ICBDR 2018)*. Association for Computing Machinery, New York, NY, USA, pp. 62–69.

Gan, S., Liang, S., Li, K., Deng, J. and Cheng, T. (2016). Ship Trajectory Prediction for Intelligent Traffic Management Using Clustering and ANN. *2016 UKACC 11th International Conference on Control* (*CONTROL*). IEEE, pp. 1–6.

Gao, M., Shi, G. and Li, S. (2018). Online prediction of ship behavior with automatic identification system sensor data using bidirectional long short-term memory recurrent neural network. *Sensors*, 18, 4211.

Gers, F. A., Schmidhuber, J. and Cummins, F. (1999). Learning to forget: Continual prediction with LSTM. *9th International Conference on Artificial Neural Networks: ICANN '99*, Edinburgh, UK, pp. 850–855.

Graser, A., Schmidt, J. and Widhalm, P. (2018). Predicting trajectories with probabilistic time geography and massive unconstrained movement data.

Graser, A., Schmidt, J., Dragaschnig, M. and Widhalm, P. (2019). Data-driven Trajectory Prediction and Spatial Variability of Prediction Performance in Mari-time Location Based Services. *15th International Conference on Location-Based Services*.

Grech, M. R., Horberry, T. and Smith, A. (2002). Human Error in Maritime Operations: Analyses of Accident Reports Using the Leximancer Tool. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. Los Angeles, CA: Sage Publications.

Harati-Mokhtari, A., Wall, A., Brooks, P. and Wang, J. (2007). Automatic Identification System (AIS): data reliability and human error implications. *The Journal of Navigation*, 60, 373–389.

Hochreiter, S., Bengio, Y., Frasconi, P. and Schmidhuber, J. (2001). Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. In: Kolen, J. F., & Kremer, S. C. (eds.), *A Field Guide to Dynamical Recurrent Neural Networks*. New York, USA: IEEE Press.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9, 1735–1780.

Kaffash-Charandabi, N., Alesheikh, A. A. and Sharif, M. (2019). A ubiquitous asthma monitoring framework based on ambient air pollutants and individuals' contexts. *Environmental Science and Pollution Research*, 26, 7525–7539.

Karatzoglou, A., Jablonski, A. and Beigl, M. (2018). A Seq2Seq Learning Approach for Modeling Semantic Trajectories and Predicting the Next Location. *Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 528–531.

Kawan, B., Wang, H., Li, G. and Chhantyal, K. (2017). Data-driven modeling of ship motion prediction based on support vector regression. *Proceedings of the 58th Conference on Simulation and Modelling (SIMS 58)*, Reykjavik, Iceland, vol. 138, pp. 350–354.

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press

Kim, K.-I. and Lee, K. M. (2018). Deep learning-based caution area traffic prediction with automatic identification system sensor data. *Sensors*, 18, 3172.

Mao, S., Tu, E., Zhang, G., Rachmawati, L., Rajabally, E. and Huang, G.-B. (2018). An Automatic Identification System (AIS) Database for Maritime Trajectory Prediction and Data Mining. In: Cao, J., Cambria, E., Lendasse, A., Miche, Y., Vong, C. (eds.), *Proceedings of ELM-2016. Proceedings in Adaptation, Learning and Optimization*, vol. 9. Springer, Cham, Switzerland.

Marinecadastre. (2019). Available at: www.marinecadastre.gov [Accessed 20 Dec. 2019].

Nishizaki, C., Terayama, M., Okazaki, T. and Shoji, R. (2018). Development of Navigation Support System to Predict New Course of Ship. *World Automation Congress* (*WAC*). IEEE.

Pallotta, G., Vespe, M. and Bryan, K. (2013). Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. *Entropy*, 15, 2218–2245.

Perera, L. P. and Soares, C. G. (2010). Ocean Vessel Trajectory Estimation and Prediction Based on Extended Kalman Filter. *The Second International Conference on Adaptive and Self-Adaptive Systems and Applications*, Lisbon, Portugal, pp. 14–20.

Qiao, S.-J., Jin, K., Han, N., Tang, C.-J. and Gesangduoji, G. (2015). Trajectory prediction algorithm based on Gaussian mixture model. *Journal of Software*, 26, 21–32.

Rodrigue, J.-P., Comtois, C. and Slack, B. (2016). *The Geography of Transport Systems*. New York, USA: Routledge.

Series, M. (2010). Technical characteristics for an automatic identification system using time-division multiple access in the VHF maritime mobile band. Recommendation ITU-R M.1371-5, February 2014, Geneva, Switzerland.

Sharif, M. and Alesheikh, A. A. (2017). Context-awareness in similarity measures and pattern discoveries of trajectories: a context-based dynamic time warping method. *GIScience & Remote Sensing*, 54, 426–452.

Sharif, M. and Alesheikh, A. A. (2018). Context-aware movement analytics: implications, taxonomy, and design framework. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8, e1233.

Sharif, M., Alesheikh, A. A. and Tashayo, B. (2019). CaFIRST: A context-aware hybrid fuzzy inference system for the similarity measure of multivariate trajectories. *Journal of Intelligent & Fuzzy Systems*, 36(6), 5383–5395.

Sheng, P. and Yin, J. (2018). Extracting shipping route patterns by trajectory clustering model based on automatic identification system data. *Sustainability*, 10, 2327.

Sidibé, A. and Shu, G. (2017). Study of automatic anomalous behaviour detection techniques for maritime vessels. *The Journal of Navigation*, 70, 847–858.

SOLAS (2003). *International Convention for the Safety of Life at Sea*. London: International Maritime Organization.

Tang, H., Yin, Y. and Shen, H. (2019). A model for vessel trajectory prediction based on long short-term memory neural network. *Journal of Marine Engineering & Technology*, 1–10, https://doi.org/10.1080/20464177.2019.1665258.

Valsamis, A., Tserpes, K., Zissis, D., Anagnostopoulos, D. and Varvarigou, T. (2017). Employing traditional machine learning algorithms for big data streams analysis: The case of object trajectory prediction. *Journal of Systems and Software*, 127, 249–257.

Wijaya, W. M. and Nakamura, Y. (2013). Predicting Ship Behavior Navigating Through Heavily Trafficked Fairways by Analyzing AIS Data on Apache HBase. *2013 First International Symposium on Computing and Networking*. IEEE, pp. 220–226.

Xiao, Z., Ponnambalam, L., Fu, X. and Zhang, W. (2017). Maritime traffic probabilistic forecasting based on vessels' waterway patterns and motion behaviors. *IEEE Transactions on Intelligent Transportation Systems*, 18, 3122–3134.

Zhao, L. and Shi, G. (2019). Maritime anomaly detection using density-based clustering and recurrent neural network. *The Journal of Navigation*, 72, 894–916.

Zorbas, N., Zissis, D., Tserpes, K. and Anagnostopoulos, D. (2015) Predicting Object Trajectories from High-Speed Streaming Data. *2015 IEEE Trustcom/BigDataSE/ISPA*. IEEE, pp. 229–234.

https://doi.org/10.1017/S0373463320000442 Published online by Cambridge University Press