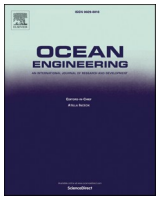
Ocean Engineering 267 (2023) 113251

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng



Identification modeling of ship maneuvering motion based on local Gaussian process regression

Zi-Lu Ouyang [a], Gang Chen [b,**], Zao-Jian Zou [a,c,*]

[a] School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China
[b] China Aerodynamics Research and Development Center Low Speed Institute, Sichuan, 621000, China
[c] State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China

ARTICLE INFO

Keywords:
Ship maneuvering
Nonparametric modeling
Local Gaussian process regression
Clustering analysis
k-means algorithm

ABSTRACT

A fast and accurate nonparametric modeling method based on local Gaussian process regression (LGPR) is proposed for the identification modeling and prediction of ship maneuvering motion. The training dataset collected from the free-running model tests of ship maneuvering is automatically divided into a number of clusters according to the similarity criterion by clustering analysis using k-means algorithm. Utilizing the data in each cluster, the corresponding local nonparametric model is identified. The computational cost of training and prediction based on LGPR is reduced compared to that based on the classic Gaussian process regression (CGPR) using the whole training dataset. Taking the KVLCC2 tanker and an unmanned surface vehicle (USV) as study objects, the nonparametric models are identified based on the experimental data of zigzag maneuvers of the KVLCC2 model and random maneuver of the USV. Using the identified models, the zigzag maneuvers of the KVLCC2 model and the random maneuver of the USV, which are not involved in the training data, are predicted. The results show that LGPR has higher computational efficiency than CGPR with acceptable prediction accuracy.

1. Introduction

With the development of Maritime Autonomous Surface Ships (MASS), an accurate and fast modeling method for ship maneuvering motion has become more and more important for the intelligent navigation of MASS, such as the design of model-based controller and collision avoidance strategy (Sutulo and Guedes Soares, 2014; Wang et al., 2021). System identification (SI) is regarded as a promising method for establishing a reliable ship dynamic model due to its advantages of low cost and high efficiency (Cao et al., 2015). Compared with other modeling methods, such as the system-based method with the help of physical captive model tests or conducting virtual captive model tests by means of computational fluid dynamics (CFD), SI only needs the easy-to-measure ship motion data collected during free-running model tests or full-scale trials (Miyauchi et al., 2022). Moreover, it can update the identified model according to the ship motion data collected in real time, which is beneficial for the design of the adaptive controller of MASS under varying operating conditions.

As an important branch of SI, nonparametric modeling has achieved significant progresses in the modeling of ship maneuvering motion. It needs almost no prior knowledges and only relies on the input data and output data to establish the nonlinear mapping relationship. The representative methods of nonparametric modeling include artificial neural network (ANN) and kernel-based method. Theoretically, neural network can approximate any nonlinear function (Hornik, 1991), and has shown satisfactory modeling and prediction ability for ship maneuvering motion (Woo et al., 2019; Xu et al., 2021; Xu et al., 2022). However, due to the complexity of the topological structure, it has the disadvantages of easy to fall into local optimum solution, overfitting, and requirement on training dataset with large size (Schölkopf et al., 2002).

Recently, kernel-based method has made considerable progresses in the modeling of ship maneuvering motion. Its core idea is to map the training data into a high-dimensional feature space by kernel trick, thereby to better capture the nonlinear relationship between the input data and output data (Ljung et al., 2020). Moreno-Salinas et al. (2019) proposed kernel ridge regression confidence machine (KRRCM) to establish the black-box model of a surface marine vehicle. Bai et al. (2019a) explored locally weighted learning (LWL) modified by genetic optimization to identify the nonparametric model of a full-scale ship.

* Corresponding author. School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China.
** Corresponding author.
E-mail addresses: chengang@cardc.cn (G. Chen), zjzou@sjtu.edu.cn (Z.-J. Zou).

https://doi.org/10.1016/j.oceaneng.2022.113251
Received 4 August 2022; Received in revised form 27 October 2022; Accepted 19 November 2022
Available online 2 December 2022

Wang et al. (2020) proposed nu-support vector machine (SVM) assisted by the kernel trick to establish a robust model for KVLCC2 ship. Xue et al. (2020) employed Gaussian process regression (GPR) with input noise to identify the dynamic model of a container ship. Among the methods mentioned above, GPR has sparked concern in the field of ship maneuverability due to its advantages of universality and interpretability, and has achieved accurate modeling results. However, the computational cost of the training process of GPR is of order $O(n^3)$ due to the need of inverting a potentially large covariance matrix, and the costs of calculating the predictive mean and variance are of order $O(n)$ and $O(n^2)$ respectively for one test sample, where n is the sample size of the training dataset. Considering that the prediction of ship maneuvering motion is a continuous iterative process, if n is large or the prediction duration is long, the whole training and prediction process of GPR will become computationally expensive, and this shortcoming will restrict its engineering practicability, especially in the design of adaptive controller for MASS. Melkumyan and Ramos (2009) pointed out that the tasks with thousands of training samples are intractable for the classic GPR. However, to identify a robust and accurate ship dynamic model by using nonparametric modeling technology generally needs a rather large amount of training samples due to the strong nonlinearity of ship maneuvering motion.

To reduce the computational cost of GPR, some researchers have proposed a number of effective methods. Candela and Rasmussen (2005) recast the approximations to the prior and introduced inducing variables. Snelson and Ghahramani (2005) designed a new GPR method, whose covariance is parameterized by the locations of several pseudo-input points. The prediction performances reveal that the proposed method can match full Gaussian process performances with sparse solutions. Melkumyan and Ramos (2009) presented a new stationary covariance function that can provide a sparse covariance matrix at a lower computational cost, and the experimental results showed that the proposed method can achieve faster inference and less memory usage. Titsias (2009) introduced a variational formulation for sparse approximations that jointly infers the inducing inputs and the kernel hyperparameters by maximizing a lower bound of the true log marginal likelihood. Wilson and Nickisch (2015) introduced kernel interpolation for scalable structured Gaussian processes, which is more scalable than inducing point alternatives and can be used for fast and expressive kernel learning.

Although considerable progresses have been made in reducing the computational cost of GPR, there are few researches on improving the computational efficiency and real-time performance in the modeling and prediction of ship maneuvering motion based on GPR. Chen et al. (2021a) employed sparse GPR with similarity for the dynamic modeling of a ship. Xue et al. (2022) proposed a novel online nonparametric identification method that combines noisy input Gaussian process and fully independent training conditional algorithm for identification modeling of ship maneuvering motion. As for reducing the computational cost of other kernel-based methods for identifying the ship dynamic model, Bai et al. (2019b) used grid index subspace constructed algorithm to reduce the computational complexity of LWL. Zhang and Ren (2021) applied sparse Gaussian Process to reduce the data requirements of LWL. The results obtained by the methods mentioned above show that these methods can effectively reduce the time-consuming of predicting the ship maneuvering motion, and the modeling accuracy is guaranteed.

Inspired by the previous studies, a fast and accurate nonparametric modeling method based on local Gaussian process regression (LGPR) is proposed in this paper. With the aid of k-means algorithm, the whole training dataset of ship motion is divided into a number of clusters automatically according to the similarity criterion by clustering analysis. Utilizing the data in each cluster, the corresponding local nonparametric model is identified. The time-consuming of training and prediction process of LGPR is reduced compared with that of the classic GPR (CGPR).

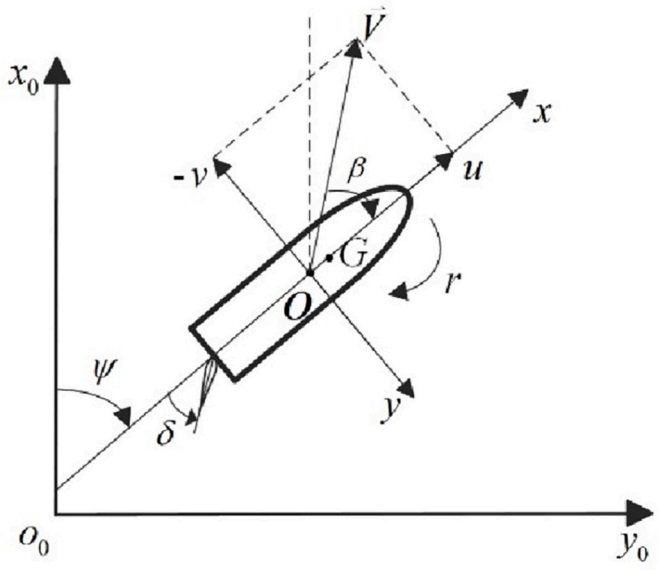


Fig. 1. Coordinate systems.

The rest of this paper is organized as follows: Section 2 presents ship kinematic model and dynamic model. In Section 3, LGPR based on clustering analysis using k-means algorithm is introduced. In Section 4, taking the KVLCC2 tanker and an unmanned surface vehicle (USV) as study objects, the nonparametric models are identified based on the experimental data of zigzag maneuvers of the KVLCC2 model and random maneuver of the USV. Using the identified models, the zigzag maneuvers of the KVLCC2 tanker model and the random maneuver of the USV, which are not involved in the training data, are predicted. The prediction results are compared with the measured data to verify the effectiveness of the proposed method. Section 5 is devoted to conclusions and perspectives for the further research.

2. Mathematical model of ship maneuvering motion

2.1. Ship kinematic model

This paper focuses on the modeling and prediction of 3-DOF ship maneuvering motion, i.e., surge, sway and yaw, in the horizontal plane. Two right-handed coordinate systems are adopted, as shown in Fig. 1, where $O_0 - x_0y_0z_0$ is the Earth-fixed frame, $O - xyz$ is the body-fixed frame; $\vec{V}$ is the translational speed with the components u and v (the surge speed and the sway speed); r is the yaw rate; ψ is the heading angle, β is the drift angle, δ is the rudder angle; G is the center of gravity of the ship. The kinematic model of 3-DOF ship maneuvering motion is given as

$$\begin{cases} \dfrac{dx_0}{dt} = u\cos\psi - v\sin\psi \\ \dfrac{dy_0}{dt} = u\sin\psi + v\cos\psi \\ \dfrac{d\psi}{dt} = r \end{cases} \tag{1}$$

where x_0 and y_0 are the position coordinates of the origin O in the Earth-fixed frame.

2.2. Ship dynamic model

In this paper, nonparametric modeling method is applied to model the ship maneuvering motion. Thus, black-box model is used when

identifying the ship dynamic model. According to the Abkowitz model (Abkowitz, 1964), the black-box model can be derived, as given in the following form:

$$\begin{cases} (m - X_{\dot{u}})\dot{u} = f_1(u, v, r, \delta) \\ (m - Y_{\dot{v}})\dot{v} + (mx_G - Y_{\dot{r}})\dot{r} = f_2(u, v, r, \delta) \\ (mx_G - N_{\dot{v}})\dot{v} + (I_z - N_{\dot{r}})\dot{r} = f_3(u, v, r, \delta) \end{cases} \tag{2}$$

where m is the mass of the ship, I_z is the moment of inertia about z-axis; x_G is the longitudinal coordinate of the ship's center of gravity; $\dot{u}$, $\dot{v}$ and $\dot{r}$ are the accelerations; X, Y, N with the accelerations as subscript are the inertial hydrodynamic derivatives; $f_i(u, v, r, \delta)(i = 1, 2, 3)$ are the functions related to the ship motion state variables and control variable.

Eq. (2) can be rewritten as

$$\begin{cases} \dot{u} = \dfrac{f_1(u, v, r, \delta)}{(m - X_{\dot{u}})} \\ \dot{v} = \dfrac{(I_z - N_{\dot{r}})f_2(u, v, r, \delta) - (mx_G - Y_{\dot{r}})f_3(u, v, r, \delta)}{(m - Y_{\dot{v}})(I_z - N_{\dot{r}}) - (mx_G - N_{\dot{v}})(mx_G - Y_{\dot{r}})} \\ \dot{r} = \dfrac{(m - Y_{\dot{v}})f_3(u, v, r, \delta) - (mx_G - N_{\dot{v}})f_2(u, v, r, \delta)}{(m - Y_{\dot{v}})(I_z - N_{\dot{r}}) - (mx_G - N_{\dot{v}})(mx_G - Y_{\dot{r}})} \end{cases} \tag{3}$$

As can be seen from Eq. (3), $\dot{u}$, $\dot{v}$ and $\dot{r}$ can be expressed as the functions of u, v, r, δ in the following form:

$$\begin{cases} \dot{u} = h_1(u, v, r, \delta) \\ \dot{v} = h_2(u, v, r, \delta) \\ \dot{r} = h_3(u, v, r, \delta) \end{cases} \tag{4}$$

Eq. (4) is the black-box model used in the identification algorithm. In the nonparametric modeling of ship maneuvering motion, $h_i(u, v, r, \delta)(i = 1, 2, 3)$ are the unknown mapping functions to be identified from the input and output data. The input of the identification algorithm is the motion state variables and control variable, while the output is the acceleration variable.

3. Nonparametric modeling based on local Gaussian process regression (LGPR)

3.1. Gaussian process regression (GPR)

Gaussian process regression (GPR) is a typical nonparametric modeling algorithm that uses Gaussian process (GP) prior to perform regression analysis (Rasmussen et al., 2004). It contains the regression residual and a prior of GP solved by Bayesian inference.

A GP is a collection of random variables and can be used to describe a distribution over functions. A real process $f(\boldsymbol{x})$ can be completely determined by a mean function $m(\boldsymbol{x})$ and a covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$; thus, it can be written as

$$f(\boldsymbol{x}) \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \tag{5}$$

For a given training dataset $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$, where $\boldsymbol{x}_i$ is the input vector of i-th sample, and y_i is the observation value of i-th sample, the goal of GPR is to identify the mapping relationship between the input vector $\boldsymbol{x}$ and the observation $\boldsymbol{y}$:

$$\boldsymbol{y} = f(\boldsymbol{x}) + \zeta \tag{6}$$

where ζ is the additive noise which follows an independent Gaussian distribution with zero mean and variance σ_n^2. The covariance of the noisy observations $\boldsymbol{y}$ can be calculated as $K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}_n$, where $\boldsymbol{y} = [y_1, y_2, ..., y_n]^{\mathrm{T}}$, $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]^{\mathrm{T}}$, K is the covariance matrix, $\boldsymbol{I}_n$ is a n-dimensional identity matrix. Then, the joint distribution of the observations at the test sample $\boldsymbol{x}_*$ under the prior can be calculated as

$$\begin{bmatrix} \boldsymbol{y} \\ f(\boldsymbol{x}_*) \end{bmatrix} \sim N\left(0, \begin{bmatrix} K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}_n & K(\boldsymbol{X}, \boldsymbol{x}_*) \\ K(\boldsymbol{x}_*, \boldsymbol{X}) & K(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{bmatrix}\right) \tag{7}$$

Based on Eq. (7), the mean of $f(\boldsymbol{x}_*)$ and the covariance of $f(\boldsymbol{x}_*)$ can be expressed as

$$m(f(\boldsymbol{x}_*)) = E[f(\boldsymbol{x}_*)|\boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*] = K(\boldsymbol{x}_*, \boldsymbol{X})\left[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}_n\right]^{-1} \boldsymbol{y} \tag{8}$$

$$\mathrm{cov}(f(\boldsymbol{x}_*)) = K(\boldsymbol{x}_*, \boldsymbol{x}_*) - K(\boldsymbol{x}_*, \boldsymbol{X})\left[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}_n\right]^{-1} K(\boldsymbol{X}, \boldsymbol{x}_*) \tag{9}$$

The covariance function defines the similarity between the samples. Squared exponential (SE) function k_{SE} that has been proven to be reliable (Bai et al., 2019a; Moreno et al., 2019; Wang et al., 2020; Ouyang and Zou, 2021) is used in this paper. The mathematical expression is given as

$$k_{SE} = \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{x}_i - \boldsymbol{x}_*)^{\mathrm{T}}(\boldsymbol{x}_i - \boldsymbol{x}_*)\right] \tag{10}$$

where σ^2 is the hyperparameter that represents the width of SE covariance function. In this paper, the hyperparameters of the covariance function are tuned by maximizing the log of marginal likelihood, whose mathematical expression is given as

$$\log(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})) = -\frac{1}{2}\boldsymbol{y}^{\mathrm{T}}\left(K + \sigma_n^2 \boldsymbol{I}_n\right)^{-1}\boldsymbol{y} - \frac{1}{2}\log\left|K + \sigma_n^2 \boldsymbol{I}_n\right| - \frac{n}{2}\log 2\pi \tag{11}$$

where $\boldsymbol{\theta}$ is the vector of all hyperparameters. As can be seen from Eq. (11), the marginal likelihood has three terms: the first term stands for the data fit, the second term accounts for the complexity penalty, and the last term is a normalization constant. To maximize the log marginal likelihood, the partial derivatives of the marginal likelihood is calculated as

$$\begin{cases} \boldsymbol{\lambda} = \left[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}_n\right]^{-1} \boldsymbol{y} \\ \dfrac{\partial}{\partial \theta_j}\log(p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})) = \dfrac{1}{2} tr\left\{\left[\boldsymbol{\lambda}\boldsymbol{\lambda}^{\mathrm{T}} - \left(K + \sigma_n^2 \boldsymbol{I}_n\right)^{-1}\right]\dfrac{\partial\left(K + \sigma_n^2 \boldsymbol{I}_n\right)}{\partial \theta_j}\right\} \end{cases} \tag{12}$$

In the classic GPR (CGPR), the whole training dataset $D = \{\boldsymbol{x}_i, y_i\}_{i=1}^n$ is used to train the nonparametric model and calculate the prediction results for a given test sample, the dimension of the covariance matrix K in CGPR is $n \times n$. Therefore, as can be seen from Eq. (12), the training of nonparametric model based on CGPR requires $O(n^3)$ times due to the inversion of the covariance matrix. Once the inversion of $[K(\boldsymbol{X}, \boldsymbol{X}) + \sigma_n^2 \boldsymbol{I}_n]$ is done, as can be seen from Eqs. (8) and (9), the computational burden of prediction based on CGPR is of order $O(n)$ for calculating the predictive mean, and $O(n^2)$ for calculating the predictive variance per each new test case. If the training dataset has a large sample size, both the training and prediction processes will become time-consuming (Rasmussen, 1996).

3.2. Local Gaussian process regression (LGPR) based on clustering analysis

In order to improve the computational efficiency of GPR, clustering analysis is introduced to divide the whole training dataset into a number of clusters. The core idea is to calculate and update the cluster center points according to similarity between the samples, so that the similarities of samples in the same cluster are as large as possible, while the differences of samples in the different clusters are as large as possible (Ester et al., 1996). For the input matrix $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n]^{\mathrm{T}}$ in the training dataset, the n samples in $\boldsymbol{X}$ can be divided into k clusters by clustering analysis algorithm:

$$\begin{aligned} &\boldsymbol{X} \rightarrow [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_s, ..., \boldsymbol{X}_k]^{\mathrm{T}} \\ &size(\boldsymbol{X}_1) + size(\boldsymbol{X}_2) + ...size(\boldsymbol{X}_s) + ...size(\boldsymbol{X}_k) = n \end{aligned} \tag{13}$$

where $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_s, ..., \boldsymbol{X}_k$ are k different clusters divided from $\boldsymbol{X}$.

According to Rasmussen et al. (2004), the mean of $f(\boldsymbol{x}_*)$ can be written as

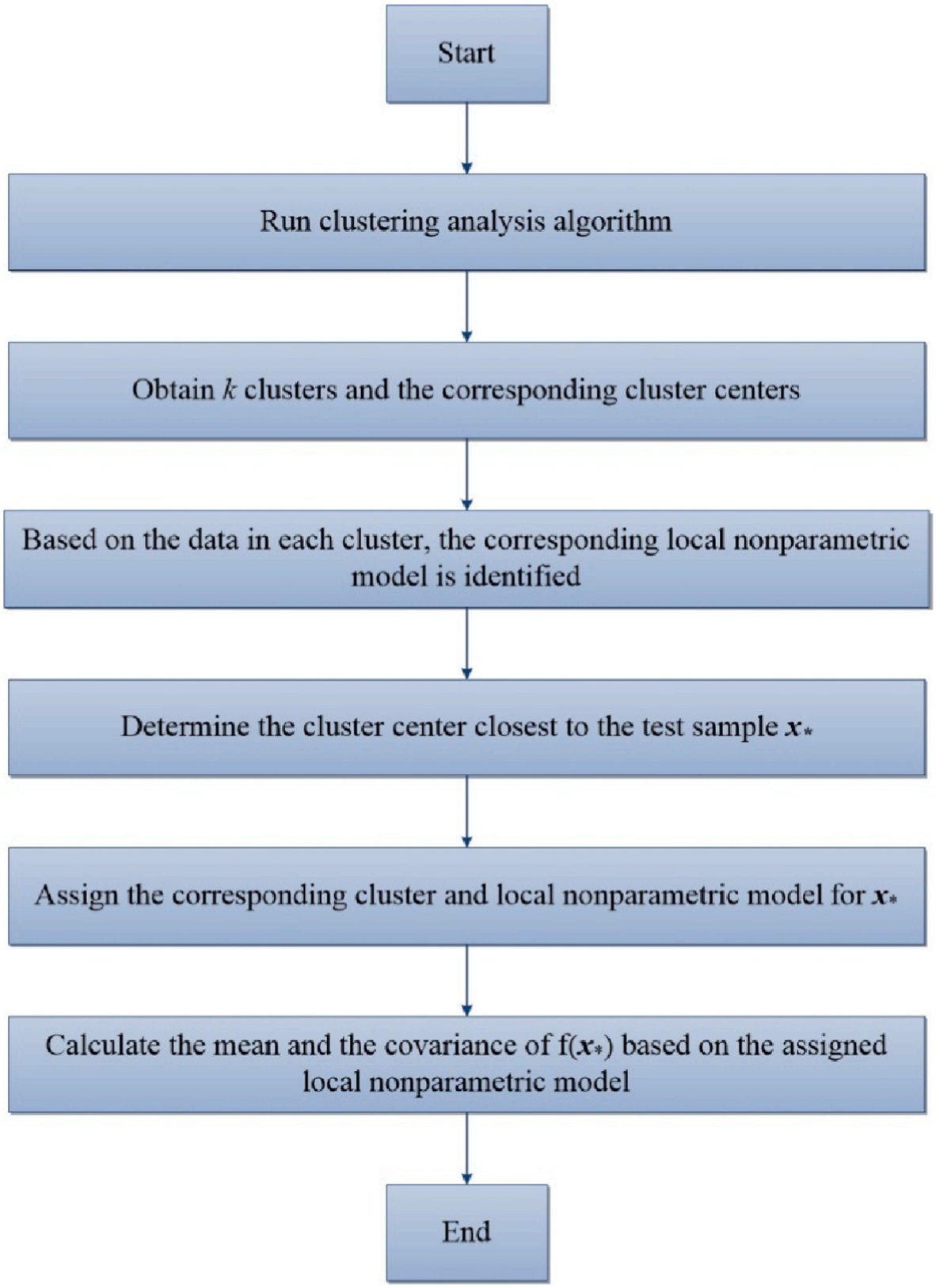


Fig. 2. The workflow of local GPR.

$$m(f(\boldsymbol{x}_*)) = \sum_{i=1}^{n} \alpha_i k_f(\boldsymbol{x}_i, \boldsymbol{x}_*) \tag{14}$$

where $\boldsymbol{\alpha} = (K + \sigma_n^2 \boldsymbol{I}_n)^{-1} \boldsymbol{y}$, k_f is the adopted covariance function. As can be seen from Eq. (14), the prediction results of the test sample $\boldsymbol{x}_*$ are determined by the coefficients α_i and the output of n covariance functions $k_f(\boldsymbol{x}_i, \boldsymbol{x}_*)$. For the latter, as can be seen from the mathematical expression of SE covariance function, Eq. (10), for a given test sample $\boldsymbol{x}_*$,

the larger the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_*$ is, the closer the output of k_{SE} approaches to zero. Therefore, according to Eq. (14), it can be concluded that the prediction results of the test sample $\boldsymbol{x}_*$ are mainly determined by the samples in the training dataset close to $\boldsymbol{x}_*$. The clustering process that utilizes the natural characteristics of the training dataset does little harm to the smoothness of the identified model, and even if the classic GPR with the whole training dataset is used, the degree of the sparsity and aggregation of the training data affects and even determines the performances of the identified model naturally.

Based on the above analysis, once the clustering analysis for $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^{\mathrm{T}}$ is done, the prediction results of test sample $\boldsymbol{x}_*$ can be calculated using the local nonparametric model established on the cluster $\boldsymbol{X}_s = [\boldsymbol{x}_{s1}, \boldsymbol{x}_{s2}, \ldots, \boldsymbol{x}_{sm}]^{\mathrm{T}} (m < n)$, whose center point is closest to $\boldsymbol{x}_*$, instead of using the nonparametric model established on the whole training dataset $\boldsymbol{X}$. This can not only accelerate the prediction process, but also guarantee the prediction accuracy. Considering that GPR is used to identify the local nonparametric model based on the data in the cluster, the proposed method is named as local GPR (LGPR).

The mean and the variance of $f(\boldsymbol{x}_*)$ in LGPR are calculated as

$$m(f(\boldsymbol{x}_*)) = E[f(\boldsymbol{x}_*)|\boldsymbol{X}_s, \boldsymbol{y}_s, \boldsymbol{x}_*] = K(\boldsymbol{x}_*, \boldsymbol{X}_s)\left[K(\boldsymbol{X}_s, \boldsymbol{X}_s) + \sigma_n^2 \boldsymbol{I}_m\right]^{-1} \boldsymbol{y}_s \tag{15}$$

$$\mathrm{cov}(f(\boldsymbol{x}_*)) = K(\boldsymbol{x}_*, \boldsymbol{x}_*) - K(\boldsymbol{x}_*, \boldsymbol{X}_s)\left[K(\boldsymbol{X}_s, \boldsymbol{X}_s) + \sigma_n^2 \boldsymbol{I}_m\right]^{-1} K(\boldsymbol{X}_s, \boldsymbol{x}_*) \tag{16}$$

where $\boldsymbol{y}_s = [y_{s1}, y_{s2}, \ldots, y_{sm}]^{\mathrm{T}} (m < n)$ is the corresponding observation vector of $\boldsymbol{X}_s$. As can be seen from Eqs. (15) and (16), the computational burden of prediction based on LGPR for calculating the predictive mean and variance of $\boldsymbol{x}_*$ are of order $O(m)$ and $O(m^2)$, respectively. As for the training process, since the whole training dataset is divided into k clusters, there are k local nonparametric models to be trained. The computational burden is of order $O(m_i{}^3)$ for training the i-th ($i = 1,2, \ldots, k$) nonparametric model based on the data in the i-th cluster (m_i is the size of the i-th cluster, $m_i < n$).

The detailed workflow of LGPR is shown in Fig. 2.

3.3. k-means algorithm for clustering analysis

k-means algorithm is adopted for clustering analysis in this paper. As a classic unsupervised method, it has the advantages of fast operation and convergence speed, strong interpretability, and good scalability. It is a partition-based clustering algorithm and has been successfully applied to image segmentation (Liew and Yan, 2003), recognition of structures (Chen et al., 2021b), chemical engineering (Sancho et al., 2022), etc. Since the SE covariance function given in Eq. (10) uses Euclidean distance as the basic component, it is also used as the criterion for similarity measurement between the data samples in k-means algorithm. The smaller the distance between the data samples is, the more likely they are in the same cluster. The implementation steps are given as follow:

Step 1. Determine the number of clusters k of the normalized training dataset; select k sample points randomly as the initial centers of clusters; calculate the initial squared error value E_0. The squared error E is defined as

$$E(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k) = \frac{1}{2}\sum_{j=1}^{k}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}_j)^2 \tag{17}$$

where $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$ are the centers of clusters.

Step 2. For each sample point $\boldsymbol{x}_i$, compute the distances between $\boldsymbol{x}_i$ and $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_k$. Euclidean distance is used to measure the similarity between the sample points. After the calculation, each sample point $\boldsymbol{x}_i$ is assigned into the cluster whose center is closest to $\boldsymbol{x}_i$.

Step 3. Recalculate the center of each cluster, which is the mean vector of the cluster.

Step 4. Recalculate E based on the new cluster and cluster center; if $|E - E_0|/E_0 \leq \sigma$(where σ is the critical rate of change), output the clustering results, otherwise turn to Step 2.

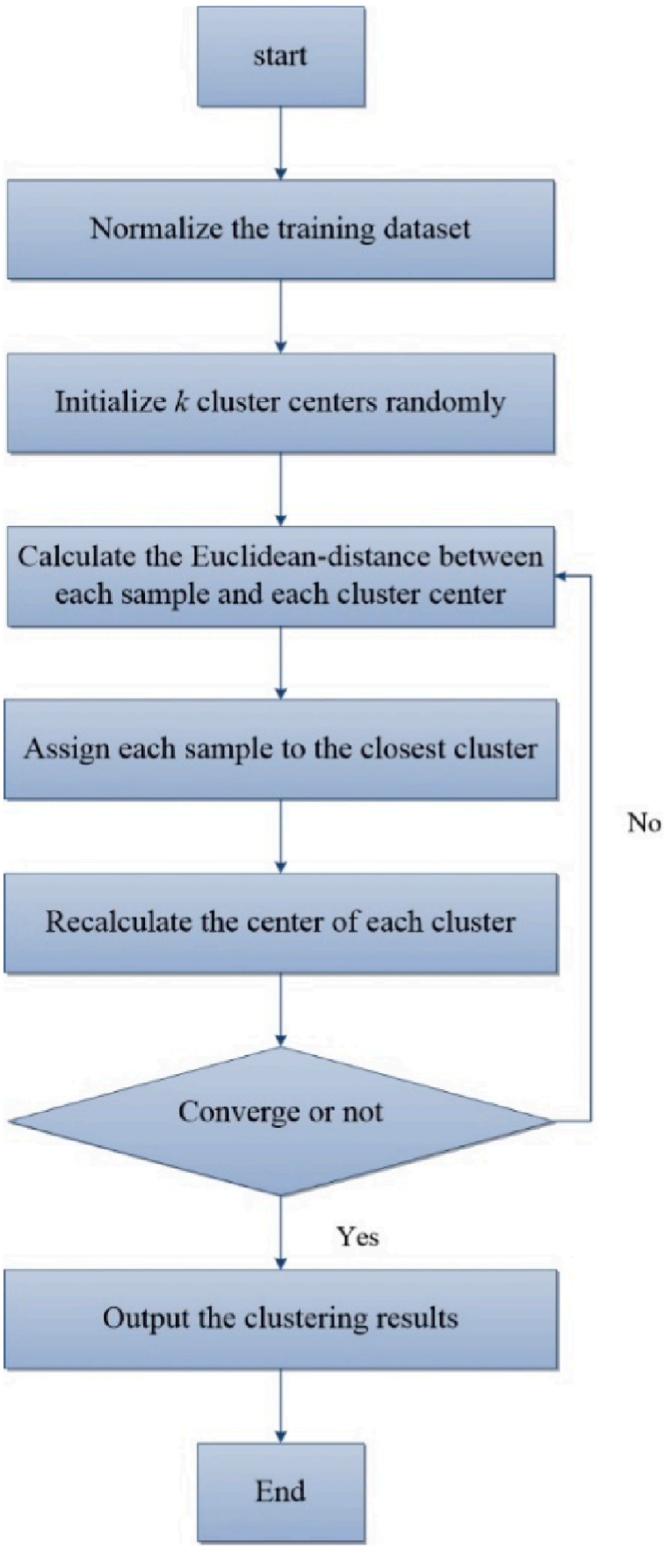


Fig. 3. The workflow of k-means algorithm.

The detailed workflow of k-means algorithm is shown in Fig. 3.

Table 1
Principal particulars of KVLCC2 tanker.

Parameter	Full scale	Ship model
Length, L (m)	320	7
Breadth, B (m)	58	1.269
Depth, D (m)	30	0.656
Draft, d (m)	20.8	0.455
Block coefficient, C_B	0.810	0.810

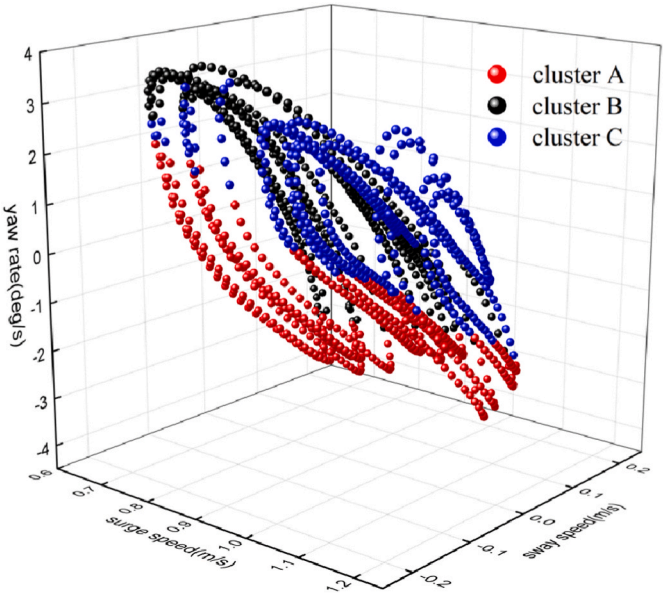


Fig. 4. The results of clustering analysis based on k-means algorithm, KVLCC2.

Table 2
Sizes of the clusters, KVLCC2.

	Cluster A	Cluster B	Cluster C	Total
Size	471	420	383	1274

4. Modeling of ship maneuvering motion utilizing measured test data

4.1. Modeling for KVLCC2 based on zigzag maneuvers

In order to evaluate the effectiveness of LGPR, the KVLCC2 tanker is taken as the study project. The principal particulars of the ship are listed in Table 1. The nonparametric modeling for this ship is carried out based on the experimental data of zigzag maneuvers of the free-running model provided by SIMMAN 2008 Workshop (SIMMAN, 2008). To evaluate the computational efficiency of LGPR, CGPR is also used to establish the nonparametric model for comparison.

Referring to Chen et al. (2021a), the data of 15°/5°, 20°/5°, 30°/5° and 35°/5° zigzag maneuvers are selected as the training dataset. The whole training dataset contains 1274 samples with a sampling time interval of 0.05 s. To avoid that the variables with larger range of values dominate the variables with smaller range of values, the min-max normalization is performed before clustering analysis; the mathematical expression is given as

$$x' = \frac{x - x_{\max}}{x_{\max} - x_{\min}} \tag{18}$$

where x' is the scaled variable, x is the original variable; $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the variable, respectively.

The number of cluster k in k-means algorithm is set as 3, and the critical rate of change σ is set as 5%. The results of clustering analysis for the whole training dataset based on k-means algorithm are shown in Fig. 4; the size of each cluster is given in Table 2. As can be seen from Fig. 4 and Table 2, the whole training dataset is divided into three clusters denoted as cluster A, cluster B, and cluster C, with the largest size of 471.

Table 3
Comparison of the total training time between LGPR and CGPR, KVLCC2.

	CGPR	LGPR
Time (s)	59.923	18.215

Table 4
Sizes of the validation datasets, KVLCC2.

	ZZ2010SB	ZZ2505SB	ZZ1005SB
Size	3761	3661	3461

Based on the data in each cluster, the corresponding local nonparametric model is established by LGPR, while CGPR uses the whole training dataset to identify the ship dynamic model. Table 3 presents the comparison of the total training time between LGPR and CGPR. As can be seen from Table 3, LGPR can save almost 70% of the training time compared with that of CGPR when identifying the ship dynamic model. Although there are three local nonparametric models to be identified by LGPR due to clustering analysis, for each local model, the dimension of λ in Eq. (12) based on LGPR is much smaller than that based on CGPR, thereby saving the time to calculate the inverse matrix during the training process.

To evaluate and compare the generalization ability and the computational efficiency of LGPR, the prediction of 20°/10°, 25°/5°, and 10°/5° zigzag maneuvers which are not involved in the training dataset are conducted for validation by using the nonparametric models identified by LGPR and CGPR. Table 4 presents the sizes of the validation datasets, where, taking 'ZZ2010SB' as example, 'ZZ' denotes zigzag maneuver, '20' denotes the specified rudder angle of the zigzag maneuver, '10' denotes the heading angle when switching the rudder, 'SB' denotes the zigzag maneuver starting from turning to starboard side.

Figs. 5–7 show the prediction results of heading angle and speed components of 20°/10°, 25°/5°, and 10°/5° zigzag maneuvers by CGPR and LGPR in comparison with the test data, respectively. Fig. 8 shows the comparison of time consumed for prediction. Table 5 presents the prediction accuracy evaluated by root-mean-square error (RMSE), where R_u, R_v and R_r denote the RMSE values of surge speed (m/s), sway speed (m/s) and yaw rate (rad/s), respectively. RMSE is defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{19}$$

where y_i is the experimental data, $\hat{y}_i$ is the predicted value by the identified model.

As can be seen from Figs. 5–7, overall the heading angle and the speed components predicted by LGPR and CGPR both fit with the experimental results well. No overfitting phenomenon occurs, which indicates that these two methods both have good generalization ability. The RMSE values of the speed components given in Table 5 show that the R_u, R_v and R_r values of LGPR are very close to and most are less than those of CGPR, indicating that the prediction results obtained by the local nonparametric models based on Eqs. (15) and (16) are not unfavorably affected by the smaller sample size of each cluster, and the prediction accuracy is well guaranteed by clustering analysis. Moreover, the time consumed for prediction shown in Fig. 8 reveals that LGPR can save almost 80% of the prediction time compared with that of CGPR. The improvement of the computational efficiency can be attributed to the reduction of computational burden in the prediction of ship

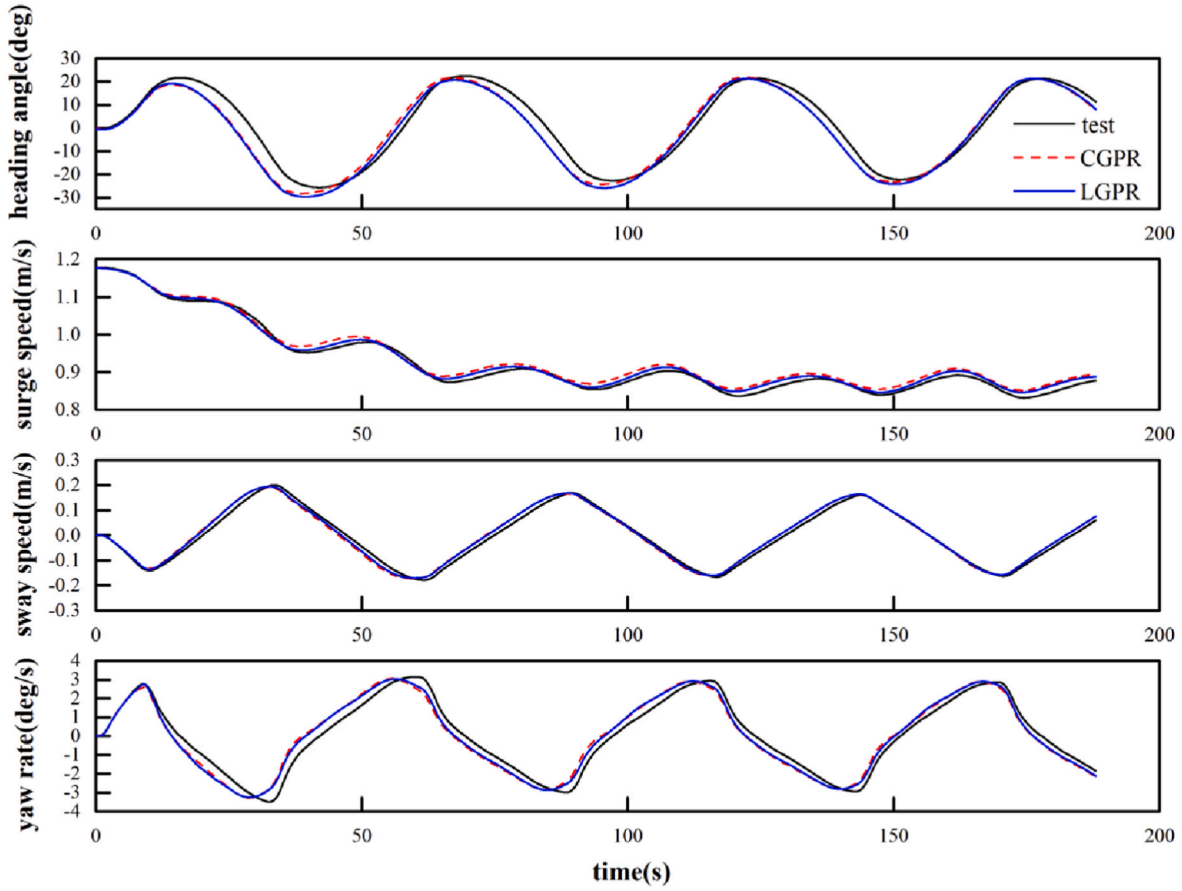


Fig. 5. Prediction results of 20°/10° zigzag maneuver by CGPR and LGPR in comparison with test data, KVLCC2.

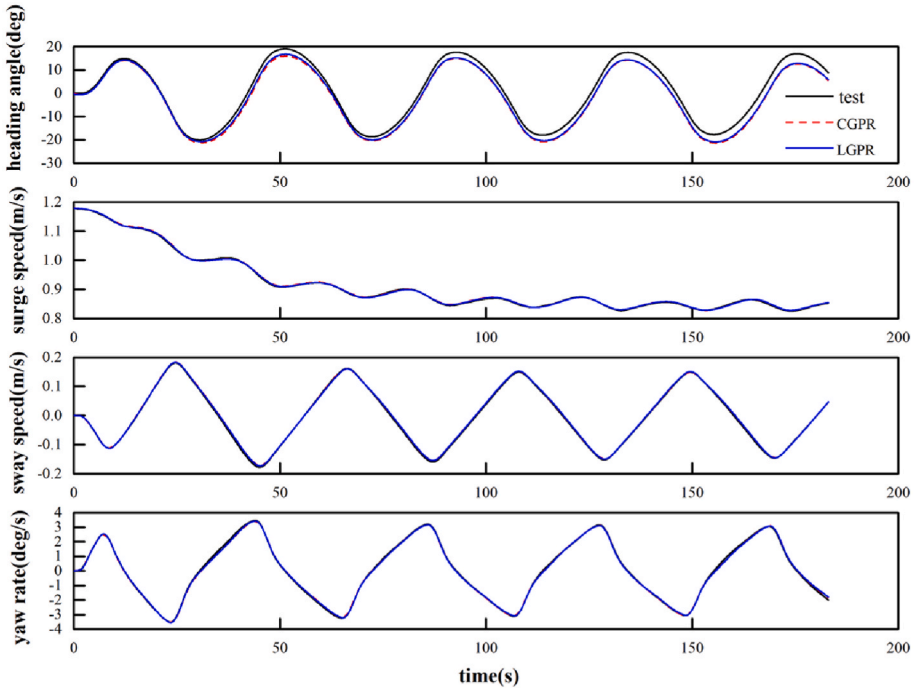


Fig. 6. Prediction results of 25°/5° zigzag maneuver by CGPR and LGPR in comparison with test data, KVLCC2.

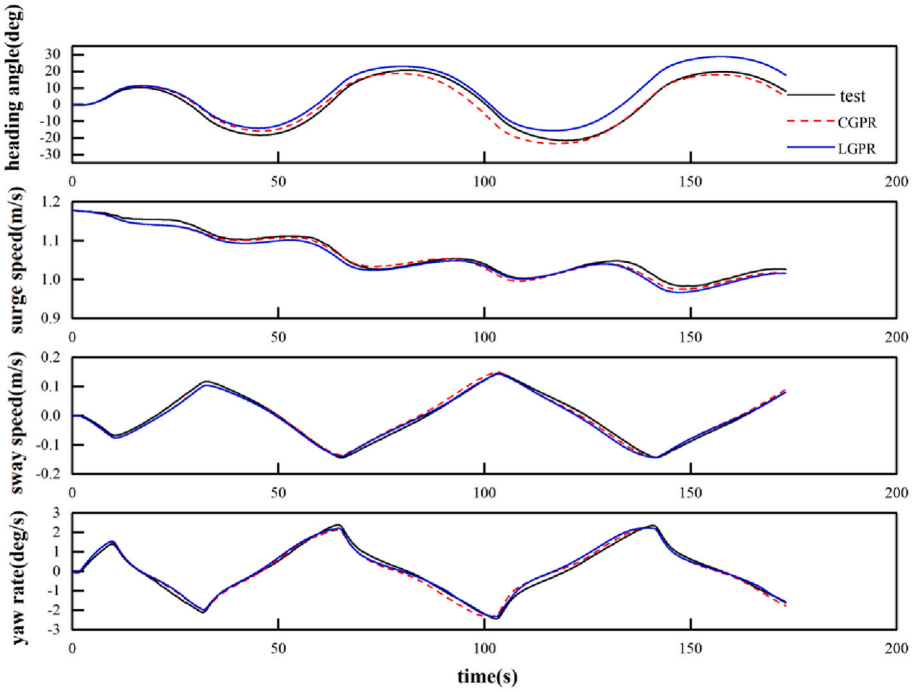


Fig. 7. Prediction results of 10°/5° zigzag maneuver by CGPR and LGPR in comparison with test data, KVLCC2.

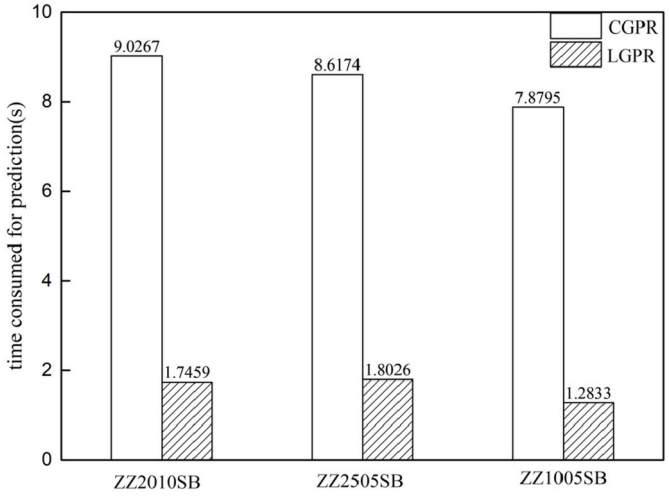


Fig. 8. Comparison of time consumed for prediction between CGPR and LGPR, KVLCC2.

maneuvering motion; even if the test samples are divided into the cluster with the largest sample size of 471, the computational burden of prediction is of order $O(471)$ for calculating the predictive mean, and $O(471^2)$ for calculating the predictive variance per each new test case, compared with $O(1274)$ and $O(1274^2)$ for those based on CGPR, the computational burden is reduced significantly.

It can be seen from Table 5 that most RMSE values of LGPR are slightly smaller than those of CGPR in the prediction of zigzag maneuvers. However, it does not mean that LGPR can improve the prediction accuracy. Two possible reasons for the smaller RMSE values of LGPR are as follows: First, sometimes the RMSE values are affected by some predictive outliers, especially in the case where the overall prediction accuracy of the two methods is very close to each other; only a few predictive outliers may lead to a larger RMSE value. In fact, as can be seen from the speed components predicted by these two methods shown in Figs. 5 and 6, the overall prediction results of LGPR are almost entirely consistent with those of CGPR. Second, considering that the dimension of $[K(\boldsymbol{X},\boldsymbol{X})+\sigma_n^2\boldsymbol{I}_n]$ in CGPR given in Eqs. (8) and (9) is large, the accuracy decrease will inevitably happen in CGPR when computing its inverse matrix. By contrast, each local nonparametric model is established based on the corresponding cluster in LGPR, the dimension of $\left[K(\boldsymbol{X}_s,\boldsymbol{X}_s)+\sigma_n^2\boldsymbol{I}_m\right]$ in LGPR is smaller; thus, the degree of accuracy decrease is lower. However, this does not mean that a smaller sample size of training data used in GPR is beneficial for modeling and prediction. As a data-driven method, nearly no prior knowledge of the dynamic system is available for GPR. Therefore, the training dataset should contain enough information that can characterize the unknown dynamic system, especially for the ship dynamic system with high nonlinearity.

Remarkably, in the prediction results of 10°/5° zigzag maneuver shown in Fig. 7, the heading angle predicted by LGPR demonstrates that LGPR has larger prediction deviation, especially after 100 s. However, as can be seen in Table 5, the R_r value of LGPR is slightly smaller than that of CGPR. Considering that the heading angle is the integration of yaw rate over time, the smaller R_r value of LGPR is a bit of a contradiction. This phenomenon can be explained as follows: As can be seen from

Table 5
Comparison of RMSE values of speed components between CGPR and LGPR, KVLCC2.

	20°/10° zigzag		25°/5° zigzag		10°/5° zigzag	
	CGPR	LGPR	CGPR	LGPR	CGPR	LGPR
R_u (m/s)	1.47×10^{-2}	9.18×10^{-3}	2.26×10^{-3}	1.90×10^{-3}	7.90×10^{-3}	1.06×10^{-2}
R_v (m/s)	1.56×10^{-2}	1.30×10^{-2}	3.47×10^{-3}	3.49×10^{-3}	1.01×10^{-2}	1.00×10^{-2}
R_r (rad/s)	8.10×10^{-3}	7.19×10^{-3}	1.20×10^{-3}	1.21×10^{-3}	3.00×10^{-3}	2.78×10^{-3}

Table 6
Principal particulars of the USV.

Length, L (m)	Breadth, B (m)	Draft, d (m)	Block coefficient, C_B
3.0	1.45	0.25	0.522

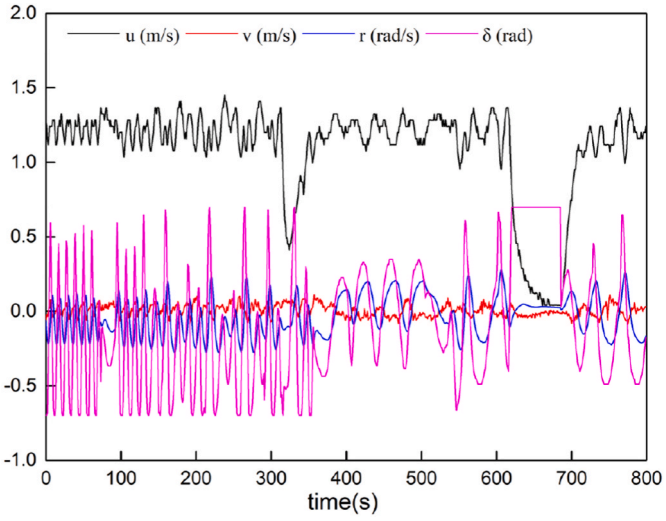


Fig. 9. Training data.

Fig. 7, the yaw rate predicted by CGPR is smaller than the experimental result during 75–100 s and is larger than the experimental result during 100–150 s. Therefore, when integrating over time, the heading angle predicted by CGPR is also smaller than the experimental result during 75–100 s. During 100–150 s, the larger yaw rate predicted by CGPR gradually makes up the smaller heading angle predicted during 75–100 s to a certain extent, thereby making the predicted heading angle gradually fit with the experimental result. By contrast, the yaw rate predicted by LGPR is consistent with the experimental result during 75–100 s and is larger than the experimental result during 100–150 s. Therefore, the heading angle predicted by LGPR gradually becomes larger than the experimental result after 100 s. The rather distinct deviation of LGPR in the prediction of 10°/5° zigzag maneuver may be attributed to the construction of the training dataset. The difference of the dynamic characteristics between 10°/5° zigzag maneuver and the zigzag maneuvers involved in the training dataset is rather large. Therefore, the prediction accuracy is not as high as those of 20°/10° and 25°/5° zigzag maneuvers.

4.2. Modeling for an unmanned surface vehicle based on random maneuver

The datasets of KVLCC2 tanker model provided by SIMMAN 2008 Workshop only include zigzag maneuvers. However, the rudder angle signals are always irregular for a ship in the actual voyage, and especially for a MASS during the path following and collision avoidance operation. To further evaluate the generalization ability and the computational efficiency of the proposed modeling method, the modeling for an unmanned surface vehicle (USV) is carried out based on the data of random maneuver collected during the free-running tests. The principal particulars of the USV are given in Table 6. The free-running tests were conducted in an open-air basin with dimension of 120 m × 60 m. The nominal speed of the USV was 2 m/s. Considering that the tests were conducted in a rather moderate wind condition, the environmental effect is not taken into account in this paper, same as in Xue et al. (2022).

The samples collected during 0–800 s with a sampling time interval of 1 s are used to identify the nonparametric models based on LGPR and CGPR, respectively. Using the identified nonparametric models, the prediction of ship maneuvering motion of 200 s not involved in the training dataset is conducted for validation. Thus, the sizes of the training dataset and the validation dataset are 800 and 200, respectively. Figs. 9 and 10 show the training data of 800 s and the random rudder angle signal of 200 s for validation, respectively.

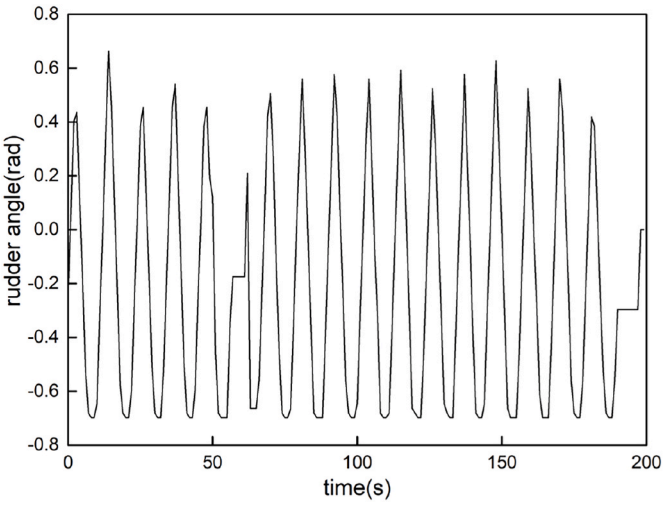


Fig. 10. Random rudder angle signal for validation.

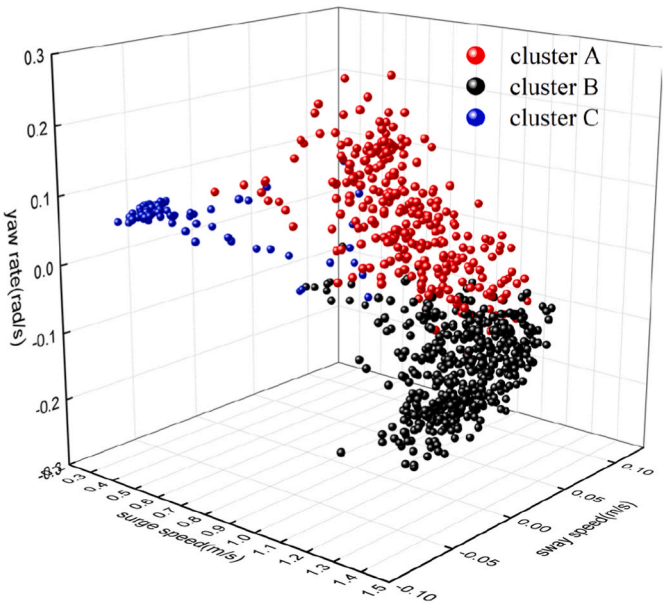


Fig. 11. The results of clustering analysis based on k-means algorithm, USV.

Table 7
Sizes of the clusters, USV.

	Cluster A	Cluster B	Cluster C	Total
Size	282	433	85	800

Table 8
Comparison of the total training time between LGPR and CGPR, USV.

	CGPR	LGPR
Time (s)	24.023	8.609

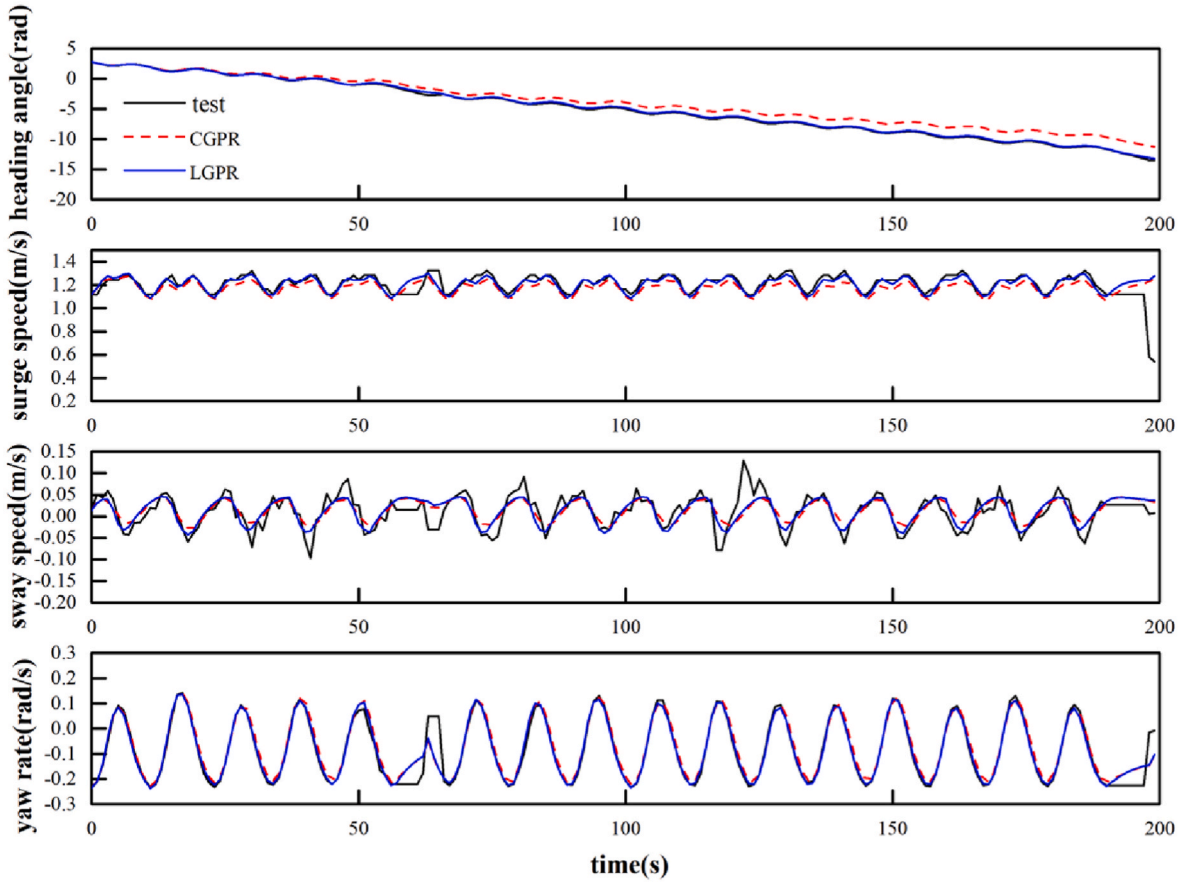


Fig. 12. Prediction results of random maneuver by CGPR and LGPR in comparison with test data, USV.

Table 9
Comparison of RMSE values of speed components and time consumed for prediction between CGPR and LGPR, USV.

	R_u (m/s)	R_v (m/s)	R_r (rad/s)	Time (s)
CGPR	8.76×10^{-2}	2.53×10^{-2}	3.27×10^{-2}	2.99×10^{-1}
LGPR	7.93×10^{-2}	2.31×10^{-2}	2.88×10^{-2}	1.17×10^{-1}

In LGPR, the number of cluster k in k-means algorithm is set as 3. The results of clustering analysis for the whole training dataset based on k-means algorithm are shown in Fig. 11, the sizes of the clusters are given in Table 7. Based on the data in each cluster, the corresponding local nonparametric model is identified. The comparison of the training time between LGPR and CGPR is given in Table 8.

As can be seen from Table 8, similar to the studied case of KVLCC2, the training time of LGPR is much smaller than that of CGPR. Based on the nonparametric models identified by LGPR and CGPR, the prediction of ship maneuvering motion is conducted, and the prediction results by CGPR and LGPR in comparison with the test data are shown in Fig. 12. Table 9 shows the comparison of RMSE values of speed components and time consumed for prediction between CGPR and LGPR. It can be seen from Fig. 12 that the yaw rate and the heading angle predicted by LGPR and CGPR both fit with the experimental results well, indicating that these two methods both have acceptable prediction accuracy. Some deviations can be observed in the prediction of surge speed and sway speed, which may be attributed to the measurement noise contained in the training data. As can be seen from Fig. 12, the signals collected by the sensors on board contain rather distinct high-frequency components. As can be seen from Table 9 that the RMSE values of LGPR are smaller than those of CGPR, indicating that the local nonparametric models bring no accuracy decrease. Moreover, it can be seen that the time consumed for prediction by LGPR is smaller than that by CGPR. These demonstrate that LGPR also has high computational efficiency when identifying the nonparametric model for the USV based on the random maneuver.

In general, the comparison results of the two studied cases reveal that LGPR has good generalization ability and high computational efficiency. Compared to CGPR, the training time is remarkably reduced due to the decrease of the dimensionality of the covariance matrix in each local nonparametric model. Meanwhile, the time consumed for prediction is also significantly reduced, while the prediction accuracy is guaranteed. The key step of the proposed method is clustering analysis, which can not only divide automatically the whole training dataset into a number of clusters before the training process, but also assign the training samples to the closest cluster to perform regression during the prediction process.

5. Conclusions

In this paper, a fast and accurate nonparametric modeling method based on local Gaussian process regression (LGPR) is proposed for the identification and prediction of ship maneuvering motion. To improve the computational efficiency, the whole training dataset of ship maneuvering motion is divided into a number of clusters according to the similarity criterion by clustering analysis based on k-means algorithm. Utilizing the data in each cluster, the corresponding local nonparametric model is identified.

Taking the KVLCC2 tanker and an unmanned surface vehicle (USV) as study objects, the nonparametric models are identified based on the experimental data of zigzag maneuvers of the KVLCC2 model and random maneuver of the USV. Using the identified models, the zigzag maneuvers of the KVLCC2 model and the random maneuver of the USV,

which are not involved in the training dataset, are predicted. The results demonstrate that much less time is consumed for training and prediction of LGPR than that of CGPR. Meanwhile, the RMSE values of LGPR are close to or even slightly smaller than those of CGPR, indicating that LGPR has high computational efficiency and satisfactory prediction accuracy.

It shows that LGPR is promising and has the potential to be an efficient tool for the online modeling and prediction of ship maneuvering motion. However, it should be noted that the number of clusters is set as 3 artificially in this paper, the optimal solution for this needs to be explored in the future study. In addition, the proposed method is to be combined with model predictive control (MPC) in attempting to explore an online modeling and model-based adaptive control algorithm for autonomous ships.

CRediT authorship contribution statement

Zi-Lu Ouyang: Conceptualization, Methodology, Simulation, Data curation, Formal analysis, Writing – original draft. **Gang Chen:** Data curation, Experiment, Data collection. **Zao-Jian Zou:** Supervision, Funding acquisition, Project administration, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work is financially supported by the National Natural Science Foundation of China (Grant Nos. 51979165, 52211530101). The authors are grateful to the Hamburg Ship Model Basin (HSVA) for providing the experimental data as benchmark data for SIMMAN 2008 Workshop.

References

Abkowitz, M.A., 1964. Lectures on Ship Hydrodynamics - Steering and Manoeuvrability, Hydro- and Aerodynamics Laboratory. Lyngby, Denmark. Report No. Hy-5.
Bai, W.W., Ren, J.S., Li, T.S., 2019a. Modified genetic optimization-based locally weighted learning identification modeling of ship maneuvering with full scale trial. Future Generat. Comput. Syst. 93, 1036–1045.
Bai, W.W., Ren, J.S., Li, T.S., 2019b. Grid index subspace constructed locally weighted learning identification modeling for high dimensional ship maneuvering system. ISA (Instrum. Soc. Am.) Trans. 86, 144–152.
Candela, J., Rasmussen, C.E., 2005. A unifying view of sparse approximate Gaussian process regression. J. Mach. Learn. Res. 6, 1939–1959.
Cao, J., Zhuang, J.Y., Xu, F., et al., 2015. Parametric estimation of ship maneuvering motion with integral sample structure for identification. Appl. Ocean Res. 52, 212–221.
Chen, G., Wang, W., Xue, Y.F., 2021a. Identification of ship dynamics model based on sparse Gaussian process regression with similarity. Symmetry 13, 1956.
Chen, L., Shan, W.B., Liu, P., 2021b. Identification of concrete aggregates using K-means clustering and level set method. Structures 34 (2), 2069–2076.
Ester, M., Kriegel, H.P., Sander, J., et al., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon, USA.
Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Network. 4 (2), 251–257.
Liew, A.W.-C., Yan, H., 2003. An adaptive spatial fuzzy clustering algorithm for 3-D MR image segmentation. IEEE Trans. Med. Imag. 22 (9), 1063–1075.
Ljung, L., Chen, T.S., Mu, B.Q., 2020. A shift in paradigm for system identification. Int. J. Control 93 (1), 1–7.
Melkumyan, A., Ramos, F., 2009. A sparse covariance function for exact Gaussian process inference in large datasets. In: Proceedings of the 21st International Joint Conference on Artificial Intelligence. Pasadena, California, USA.
Miyauchi, Y., Maki, A., Umeda, N., et al., 2022. System parameter exploration of ship maneuvering model for automatic docking/berthing using CMA-ES. J. Mar. Sci. Technol. 27, 1065–1083.
Moreno, R., Moreno-Salinas, D., Aranda, J., 2019. Black-box marine vehicle identification with regression techniques for random manoeuvres. Electronics 8 (5), 492.
Moreno-Salinas, D., Moreno, R., Pereira, A., 2019. Modelling of a surface marine vehicle with kernel ridge regression confidence machine. Applied Soft Computing Journal 76, 237–250.
Ouyang, Z.L., Zou, Z.J., 2021. Nonparametric modeling of ship maneuvering motion based on Gaussian process regression optimized by genetic algorithm. Ocean Eng. 238, 109699.
Rasmussen, C.E., 1996. Evaluation of Gaussian Processes and Other Methods for Non-linear Regression. PhD Thesis. University of Toronto.
Rasmussen, C.E., Bousquet, O., Luxburg, U.V., et al., 2004. Gaussian Processes in Machine Learning. Springer Berlin Heidelberg.
Sancho, A., Ribeiro, J.C., Reis, M.S., et al., 2022. Cluster analysis of crude oils with k-means based on their physicochemical properties. Comput. Chem. Eng. 157, 107633.
Schölkopf, B., Smola, A.J., Bach, F., 2002. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and beyond. The MIT Press, Cambridge, Mass.
SIMMAN, 2008. http://www.simman2008/dk/.
Snelson, E., Ghahramani, Z., 2005. Sparse Gaussian processes using pseudo-inputs. In: Proceedings of the 18th International Conference on Neural Information Processing Systems, NIPS'05. Vancouver. British Columbia, Canada, pp. 1257–1264.
Sutulo, S., Guedes Soares, C., 2014. An algorithm for offline identification of ship manoeuvring mathematical models from free-running tests. Ocean Eng. 79, 10–25.
Titsias, M.K., 2009. Variational learning of inducing variables in sparse Gaussian processes. In: Proceedings of the 12th International Conference on Artificial Intelligence and Statistics. Clearwater Beach, Florida.. USA.
Wang, T.T., Li, G.Y., Wu, B.H., et al., 2021. Parameter identification of ship manoeuvring model under disturbance using support vector machine method. Ships Offshore Struct. 16 (Suppl. 1), 13–21.
Wang, Z.H., Xu, H.T., Xia, L., et al., 2020. Kernel-based support vector regression for nonparametric modeling of ship maneuvering motion. Ocean Eng. 216, 107994.
Wilson, A.R., Nickisch, H., 2015. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In: Proceedings of the 32nd International Conference on Machine Learning. Lille, France.
Woo, J., Yu, C., Kim, N., 2019. Deep reinforcement learning-based controller for path following of an unmanned surface vehicle. Ocean Eng. 183, 155–166.
Xu, P.F., Han, C.B., Cheng, H.X., et al., 2022. A physics-informed neural network for the prediction of unmanned surface vehicle dynamics. J. Mar. Sci. Eng. 10, 148.
Xu, W.Z., Maki, K.J., Silva, K.M., 2021. A data-driven model for nonlinear marine dynamics. Ocean Eng. 236, 109469.
Xue, Y., Liu, Y., Ji, C., et al., 2020. System identification of ship dynamic model based on Gaussian process regression with input noise. Ocean Eng. 216, 107862.
Xue, Y., Chen, G., Li, Z., et al., 2022. Online identification of a ship maneuvering model using a fast noisy input Gaussian process. Ocean Eng. 250, 110704.
Zhang, Z., Ren, J., 2021. Locally weighted non-parametric modeling of ship maneuvering motion based on sparse Gaussian Process. J. Mar. Sci. Eng. 9, 606.