11th International Workshop on Ship and Marine Hydrodynamics
Hamburg, Germany, September 22-25, 2019

Investigation of Training Data Selection in the Black-box Modeling of Ship Maneuvering Motion

Zihao Wang[1,2], C. Guedes Soares[2], Zaojian Zou[1,3*]

[1] School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, China
[2] Centre for Marine Technology and Ocean Engineering (CENTEC), Instituto Superior Técnico, Universidade de Lisboa, Lisboa, Portugal
[3] State Key Laboratory of Ocean Engineering, Shanghai Jiao Tong University, Shanghai, China

*Corresponding author, zjzou@sjtu.edu.cn

ABSTRACT

For the identification modeling of ship maneuvering motion, comparisons between various training data are conducted to select appropriate excitation signal with maximum dynamic information, thereby ensuring the generalization ability of the identified model. The identification framework is black-box modeling based on the ν ("nu")-support vector machine algorithm with radial basis function kernel, which automatically controls the number of support vectors and keeps sparsity. A Mariner class ship is taken as the study object, and the training data is generated from the reliable simulation model, including 10º/10º, 20º/20º, 30º/30º zigzag maneuvers and 35º turning circle maneuver. The generalization performance of the identified model under different training data is compared by predicting other standard zigzag and turning maneuvers. The results indicate that the 20º/20º and 30º/30º zigzag maneuvers contain more dynamic information and can be used to train the model when the data is pure. The present work provides guidance for the subsequent experiment research to update the ship model quickly in the field.

1 INTRODUCTION

The accuracy of mathematical models is an essential foundation for the study of ship maneuvering and control, involving applications such as real-time prediction and the model-based path planning and controller design. The dynamic characteristics of the ship are affected by changes in operating conditions such as payload, and water depth, which makes the original mathematical model not always accurate. Therefore, it is desirable to have an easy-to-operate way to update the model. System identification (SI) approach based on free-running dataset provides a practical and efficient way which required lower experiment time and cost. As a data-driven method, it can extract information from the measurements of input and state variables. The measured data can be easily obtained from some onboard sensors. This feature makes it available both for the ship model and full-scale vessel.

A common approach is to set the model structure as prior knowledge, also known as white-box modeling, which represents hydrodynamic force/moment as a Taylor expansion of the state items or as a modular model [1-3]. In this case, the identification problem can be converted to multiple linear regression and has been solved by many algorithms, such as least square, extended Kalman filter, ridge regression, support vector regression, to name but a few. A foundation issue is to pre-determine the model structure. Different series of ships often correspond to different model structures. The choice of an appropriate model structure is a trade-off [4]: a complex model makes it easier to describe the characteristics of a dynamic system. On the other hand, too many parameters are more likely to cause ill-condition and overfitting, which increases the variance of the results and makes the identified model less stable. In general, more than one expression may meet the

requirements of ship maneuvering forecast, but this situation may change if the operating condition changed. Therefore, there is always a problem of choosing an appropriate model structure.

An alternative approach is black-box modeling, which does not require any knowledge of the model structure. The only prior knowledge is the input and output dataset of the system. This property makes the identification more flexible and easier to implement. For all its advantages, black-box technical is something of a double-edged sword. The generalization ability of the model is more vulnerable to be influenced by the training data. The black-box structure may change if the training data is under noise disturbance, causing overfitting or underfitting. Thus, there are more stringent requirements on the selection of training data and identification algorithm.

Recently, there has been considerable interest in the so-called kernel-based method, which is derived from statistical learning theory, such as kernel least square, kernel ridge regression or support vector machines (SVM). The mathematical form of the basis function is embedded into the structure of the kernel. Some studies have shown good results in black-box identification for marine hydrodynamic. Luo [5] applied least square support vector regression (LSSVM) to model the maneuvering motion of a catamaran under disturbances induced by current and wind, with dataset from a turning circle test. The implicit model was established but only the turning circle maneuver was simulated. Moreno-Salinas [6] modeled maneuvering motion of a surface marine vehicle with kernel ridge regression and kernel ridge regression confidence machine. The dataset from the 20/20 zigzag experiment was used as training data, and results show satisfactory prediction accuracy. Hou [7] used the ε-SVM to identify the linear coefficients and the nonlinear function of ship roll motion in irregular waves. Sclavounos [8] employed LSSVM into the roll motion modeling and the short-term wave elevation forecast. For above SVM method, a drawback of LSSVM is the lack of sparsity as all the data points become support vectors (SVs). ε-SVM can control the number of SVs by parameter ε, and then solves the quadratic problem to obtain the global optimal solution. However, there is still the problem of choosing an appropriate parameter ε, which is related to the degree of disturbance in the training data. To improve the applicability of identification framework, the algorithm applied in the present work is an improved algorithm of ε-SVM, namely ν ("nu")-SVM algorithm [9], which can adjust the parameter ε optimally and automatically control the number of SVs even under different level-noise disturbance.

Training data plays a key factor to guarantee the generalization ability by providing sufficient dynamic information. Theoretically, the training data should be as rich as possible by involving different maneuvers. However, to investigate the applicability of the black-box identification framework and explore the simplest possible implementation, the present work compares the effects of the different standard maneuvers as the training data in the black-box modeling framework based on ν-SVM. The selected maneuvers include 10º/10º, 20º/20º, 30º/30º zigzag test and 35º turning circle test. A Mariner class vessel is taken as the object of research and simulation. It is well known that in the case where the algorithm is effective in the simulation, it can be implemented more confidently in the subsequent experimental part. The current work is intended to provide theoretical reference and guidance for practical applications.

2 BLACK-BOX MODELING

The goal of modeling is to establish an accurate forecasting model which is able to predict the normal maneuvers under design speed which rudder angle range from zero to full rudder and can be applied to simulators, controller simulations, and planning algorithms in the future.

2.1 ν-Support vector machine regression

Support Vector Machines (SVM) for regression (also known as ε-SVM [10]) aim to solve a quadratic programming problem. Based on the principle of structural risk minimization, ε-SVM has good generalization performance and the ability to avoid parameter over-fitting. The parameter ε in ε-SVM controls the sparseness property, which directly affects the performance of the algorithm. Therefore, the prior determination of the desired accuracy ε is a crucial issue. To better solve this problem, ν ("nu")-SVM is applied which can automatically adjust ε and predetermine the fraction of training sample as support vectors [10]. It is not necessary to re-determine the parameter ε if the training dataset is polluted by different levels of noise, which makes the algorithm more intelligent and robust.

Suppose the training dataset are given as $\{(x_1, y_1),...,(x_i, y_i),...,(x_N, y_N)\}, x_k \in R^n, y_k \in R$. The regression problem is to find a function $f(x)$ to approximate the training dataset, taking the form:

$$f(x) = w^T \phi(x) + b \tag{1}$$

where w is the weight vector; $\phi(\cdot)$ is the nonlinear function; b is the bias term. For ε-SVM algorithm, according to the structural risk minimization principle, the risk bound is minimized by the following optimization problem:

$$\min_{w,\xi_k} \frac{1}{2} w^T w + C \cdot \frac{1}{l} \sum_{i=1}^{l} (\xi_i + \xi_i^*) \tag{2}$$

Subject to

$$((w \cdot x_i) + b) - y_i \le \varepsilon + \xi_i \tag{3-a}$$

$$y_i - ((w \cdot x_i) + b) \le \varepsilon + \xi_i^* \tag{3-b}$$

$$\xi_i^{(*)} \ge 0 \tag{3-c}$$

Equation (2) is also called the ε-insensitive cost function, where $|\xi|_\varepsilon$ is called ε-insensitive training error, described by $|\xi|_\varepsilon = \max\{0, |\xi| - \varepsilon\}$. Only the errors above some $\varepsilon > 0$ will be penalized, the subset of these data is also called support vectors (SVs). The regularization constant C determines the trade-off between model complexity and ε-insensitive training error. l is the dimension of the corresponding variable.

In υ-SVM algorithm, the ε-insensitive cost function is used, but ε serves as a variable of the optimization problem, becoming an additional term in the cost function that attempts to minimize ε. Hence the optimization problem becomes:

$$\min_{w,\xi_k} \frac{1}{2} w^T w + C \cdot \left(\upsilon \varepsilon + \frac{1}{l} \sum_{i=1}^{l} (\xi_i + \xi_i^*) \right) \tag{4}$$

Subject to

$$y_i - (w^T \phi(x_i) + b) \le \varepsilon + \xi_i \tag{5-a}$$

$$(w^T \phi(x_i) + b) - y_i \le \varepsilon + \xi_i^* \tag{5-b}$$

$$\xi_i^{(*)} \ge 0, \quad \varepsilon \ge 0 \tag{5-c}$$

In the cost function, ε is a variable optimized by constant υ. The Lagrange formulation of this optimization problem is:

$$\begin{aligned} & L(w, b, \alpha^{(*)}, \beta, \xi^{(*)}, \varepsilon, \eta^{(*)}) \\ & = \frac{1}{2} w^T w + C\upsilon\varepsilon + \frac{C}{l} \sum_{i=1}^{l} (\xi_i + \xi_i^*) - \beta\varepsilon - \sum_{i=1}^{l} (\eta_i \xi_i + \eta_i^* \xi_i^*) \\ & - \sum_{i=1}^{l} \alpha_i (\xi_i + y_i - w^T \phi(x_i) - b + \varepsilon) \\ & - \sum_{i=1}^{l} \alpha_i^* (\xi_i^* + w^T \phi(x_i) + b - y_i + \varepsilon) \end{aligned} \tag{6}$$

where $\alpha_i^{(*)}, \beta, \eta_i^{(*)}$ are Lagrange multipliers. The optimal solution is given by the saddle point of the Lagrangian. The υ-SVM optimization formula is obtained in a quadratic programming form:

$$\min_{\alpha_i} \frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j) \phi(x_i)^T \phi(x_j) - \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) y_i \tag{7}$$

Subject to

$$\sum_{i=1}^{l} (\alpha_i^* - \alpha_i) = 0 \tag{8-a}$$

$$\alpha_i^{(*)} \in [0, \frac{C}{l}] \tag{8-b}$$

$$\sum_{i=1}^{l} (\alpha_i^* - \alpha_i) \le C \cdot \upsilon \tag{8-c}$$

The dot product of $\phi(x)$ can be substituted by kernel functions, which can avoid the dimensionality curse. In this study, the radial basis function kernel is used:

$$k(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \quad i \in \{1, ..., N\}, j \in \{1, ..., N\} \tag{9}$$

Then the regression function Eq.(1) can be rewritten as follows:

$$f(x) = \sum_{i=1}^{l} (\alpha_i^* - \alpha_i) k(x_i, x) + b \tag{10}$$

By using the υ-SVM algorithm, the insensitivity tube width is automatically adjusted by automatically determining the hyper-parameter ε. The relative number of SVs is determined in advance by the constant ν, and only the training data belonging to the SVs will affect the regression results.

2.2 Training data preparation

The selection and processing of training data are critical to the generalization performance of the identification model. Since most previous studies typically applied a single standard maneuvering data as training data, for research purposes, the comparison will be taken between the widely used standard maneuvers, i.e. the zigzag test and the turning circle test. The collected data includes input and output variables, i.e. rudder angle, surge and sway velocity and yaw rate. Referring to the Taylor expansion, the hydrodynamic forces/moments are expressed as nonlinear mapping of the above four variables and other items are considered as constants. Therefore, the inputs of training data in each time step are surge speed, sway speed, yaw rate and rudder angle, and the outputs are the surge and sway velocity and yaw rate in the next time step of input data. The dataset is divided into a large and a small group for training and evaluation, called training set and holdout set respectively.

To avoid the larger range variables dominating smaller range variables, the data are standardized and centralized.

2.3 Validation

The validation of the modeling method is focused on the generalization performance, which is the ability to predict maneuvering motion not included in the training data. Before verifying the generalization performance, the first step is to evaluate the degree of overfitting, which has a significant impact on the generalization performance. Comparing the prediction results on the training and holdout sets, if the performance on the training set is significantly better, it suggests overfitting. Then, the hyper-parameter C and γ need to fine-tune and repeat the process to get the best performance.

3 COMPARISON OF TRAINING DATA

Standard maneuvering datasets are used as training data, including the standard 10º/10º, 20º/20º, 30º/30º zigzag maneuvers and 35º turning maneuver. These datasets are relatively easy to obtain because they are the tests required by the Standards for Ship Maneuverability (IMO 2002). A Mariner class vessel is taken as the object of research. The training data is generated by computer simulation programs. The hydrodynamic coefficients in the simulation model are obtained from the PMM tests, and more details can be found in the literature [12-13]. It should be mentioned that the object ship is full-scale with the length between perpendiculars of 160.93 meters, design displacement of 16005 tons, approach speed of 15 knots. The training data has a total of 1200 data points with a time interval of 0.5 seconds. The identification framework based on nu-SVM is used to model the ship maneuvering motion from the free-running test dataset. The hyper-parameter C and γ is set by the same method of grid search and cross-validation, with the search range of $C \in 2^{\{-2,0,2,4,6,8,10,12\}}$ and $\gamma \in 2^{\{-12,-10,-8,-6,-4,-2,0,2\}}$. The hyper-parameter υ is set manually considering the computing efficient and sample data structure, which is selected as 0.3.

The comparisons of predictive ability are made between the identified models trained by different datasets. Figure 1-4 show the speed predictions over 500 seconds and compare them to the raw data. Each model can accurately predict the motion which has been used to train the model. As regards the generalization ability, the model trained by 10º/10º zigzag maneuver is relatively poor than that of 20º/20º and 30º/30º zigzag maneuvers, especially in the prediction of surge motion. Moreover, for the model trained by 10º/10º zigzag, the prediction deviation for the 30º/30º zigzag is significantly larger than the prediction deviation for the 20º/20º zigzag. This can be considered that the larger the difference between the rudder angles of motions, the larger the difference of dynamic information may be contained. The results about the turning circle test show that this maneuver is not suitable as the training data since the generalization performance is terrible, which indicates the insufficient of dynamic information. By contrast, models trained by 20º/20º and 30º/30º zigzag maneuver datasets have satisfactory predictive ability. Regarding the model trained by 20º/20º zigzag dataset, there is only a deviation in the surge motion of 35º turning circle. As for the model trained by 30º/30º zigzag

training data, it has cumulative deviations in the prediction of 10º/10º zigzag maneuvers. Anyway, their generalization ability is acceptable, even from the perspective of the overall 500s duration forecast.

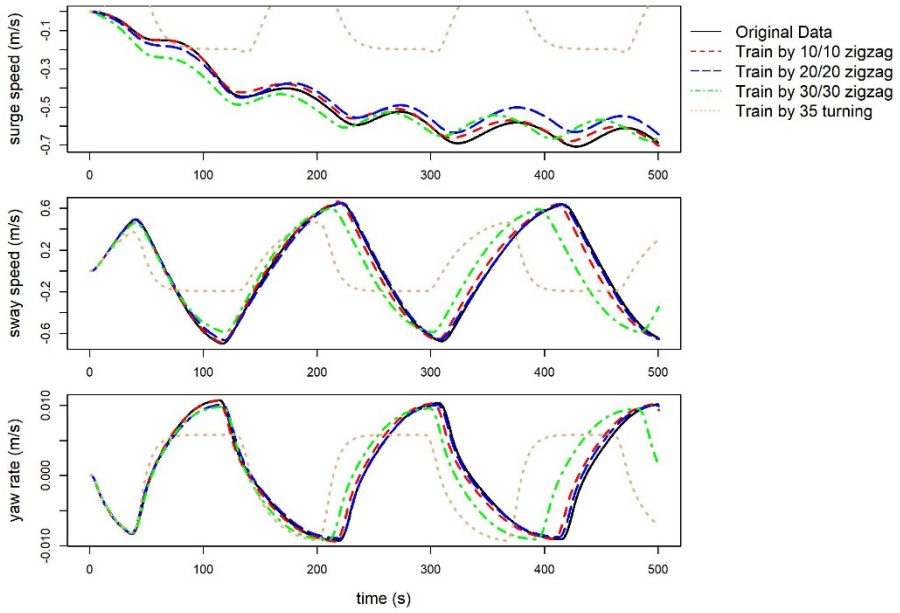

Figure 1: Prediction of 10º/10º zigzag maneuver
by model identified under different training data.

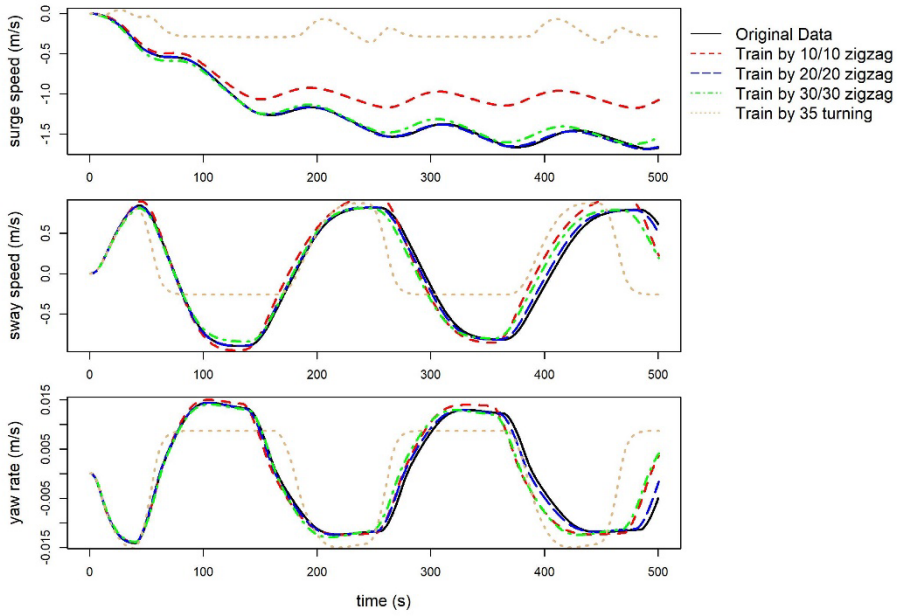

Figure 2: Prediction of 20º/20º zigzag maneuver
by model identified under different training data.

The comparison result suggests that the 20º/20º and 30º/30º zigzag maneuver are more suitable as training data for the requirement of predicting the normal maneuvers under design speed, which rudder angle range from zero to full rudder. The better generalization performance indicates that the dataset provides more dynamic information. It should be noted that the above investigation of training data is based on the specific black-box identification frameworks, including the ν-SVM with radial basis function kernel and the tuning method for the hyper-parameters. Although it may also be effective for other algorithms, further verification is needed due to the influence of other factors. Moreover, the training data is expected to be as clean as possible, otherwise problems such as overfitting will be more likely to occur. Through the current research, it is found that the

adjustment of hyper-parameters has a significant influence on the results, and problems such as over-fitting can be solved or alleviated by tuning. The simulation initially validates the identification framework and provides recommendations for the selection of training data. Despite the simplification and idealization, these results can serve as theoretical support for the subsequent experiment research.

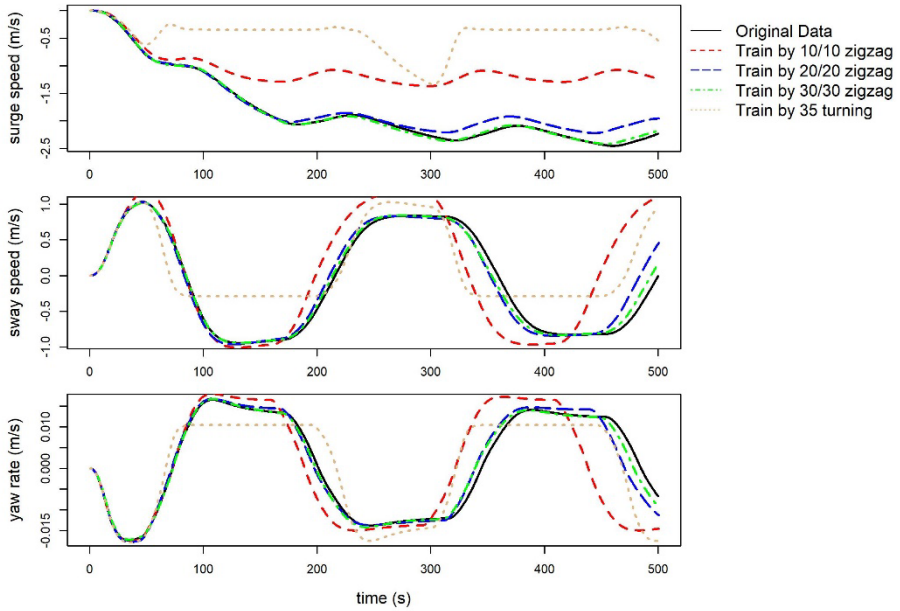

Figure 3: Prediction of 30º/30º zigzag maneuver by model identified under different training data.

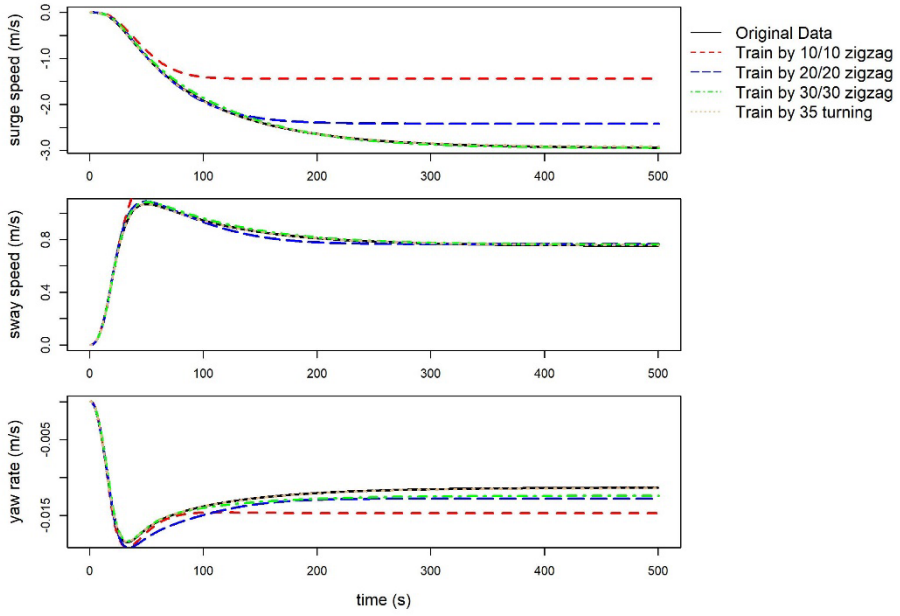

Figure 4: Prediction of 35º turning circle maneuver by model identified under different training data.

4 CONCLUSION

The analysis of the trainings data and their selection have a significant impact on identification modeling. By getting as much information as possible from limited data, the identified model will be more stable and have better generalization capabilities. This work describes a black-box identification framework for ship

maneuvering modeling which based on the ν ("nu")-SVM algorithm. On this basis, the comparison of training data is made to explore which excitation signal can provide more information, hence improving the generalization ability. On the premise that the modeling goal is to predict the maneuvers under design speed which rudder angle range from zero to full rudder, the dataset of 20º/20º and 30º/30º zigzag maneuvers have a better performance which guarantees the prediction of other maneuvers. By contrast, the turning circle dataset presents limited dynamic information, which is not suggested to be applied for the maneuvers modeling alone. The dataset of 10º/10º zigzag maneuver seems to lack information in the surge motion, which causes the accumulated error in other two DOFs. To sum up, the 20º/20º and 30º/30º zigzag maneuvers are suggested as the training data in ship maneuvering modeling. Meanwhile, the operator should obtain clean data as much as possible, in which case even the training data of a single standard maneuver may achieve good generalization performance and obtain an accurate mathematical model.

ACKNOWLEDGEMENTS

This work is financially supported by the National Natural Science Foundations of China [Grant number 51609132 and 51779140] and the CSSC Joint Fund Project 2017 [Grant number K10402]. The work of the first author has been supported by the scholarship from China Scholarship Council (CSC) under Grant No. 201806230201.

REFERENCES

[1] Sutulo, S. and C. Guedes Soares. "An algorithm for offline identification of ship manoeuvring mathematical models from free-running tests". *Ocean Engineering*, 2014. 79: p. 10-25.

[2] Luo, W.L., C. Guedes Soares. and Z.J. Zou. "Parameter Identification of Ship Maneuvering Model Based on Support Vector Machines and Particle Swarm Optimization". *Journal of Offshore Mechanics and Arctic Engineering*, 2016, 138(3):031101.

[3] Xu, H.T, M. Hinostroza, and C. Guedes Soares. "Estimation of hydrodynamic coefficients of a nonlinear manoeuvring Mathematical model with free-running ship model tests". *International Journal of Marine Engineering*, Vol. 160. 2018. A-213.

[4] Ljung, L., T.S. Chen, and B.Q. Mu. "A Shift in Paradigm for System Identification". *International Journal of Control*, 2019. (1): p. 1-7.

[5] Luo, W.L., L. Moreira, and C. Guedes Soares. "Manoeuvring simulation of catamaran by using implicit models based on support vector machines". *Ocean Engineering*, 2014. 82: p. 150-159.

[6] Moreno-Salinas, D., R. Moreno, A. Pereira, J. Aranda and J. M. de la Cruz. "Modelling of a surface marine vehicle with kernel ridge regression confidence machine". *Applied Soft Computing*, 2019. 76: p. 237-250.

[7] Hou, X.R., Z.J. Zou, and C. Liu. "Nonparametric identification of nonlinear ship roll motion by using the motion response in irregular waves". *Applied Ocean Research*, Vol. 73. 2018. 88-99.

[8] Sclavounos, P.D. and Y. Ma. "Artificial intelligence machine learning in marine hydrodynamics". In: *ASME 2018 37th International Conference on Ocean, Offshore and Arctic Engineering*, Madrid, Spain, 2018.

[9] Wang, Z.H., Z.J. Zou. and C. Guedes Soares. "Identification of Ship Manoeuvring Motion Based on nu-Sup port Vector Machine". *Ocean Engineering*, 2019. https://doi.org/10.1016/j.oceaneng.2019.04.085.

[10] Cortes, C. and V. Vapnik. "Support-vector networks". *Machine learning*, 1995. 20(3): p. 273-297.

[11] Schölkopf, B., A.J. Smola, R.C. Williamson and P.L. Bartlett. "New support vector algorithms". *Neural computation*, 2000. 12(5): p. 1207-1245.

[12] Chislett, M.S. and J. Strom-Tejsen. "Planar motion mechanism tests and full-scale steering and maneuvering predictions for a mariner class vessel". *International Shipbuilding Progress*, 1965. 12(129): p. 201-224.

[13] Fossen, T.I. *Guidance and control of ocean vehicles*. Wiley New York, 1994.