RESEARCH ARTICLE Indian Journal of Animal Research, Volume 55 Issue 11: 1364-1370 (November 2021)

Spatial and Temporal Distribution of *Portunus trituberculatus* in the Northern East China Sea based on Different Modelling Approaches

Xiaodong Li, Jing Wang, Ya Liu, Yingbin Wang

10.18805/IJAR.B-1365

ABSTRACT

Background: *Portunus trituberculatus* is an important economic crab in the East China Sea. With the increase of tonnage and power of offshore fishing vessels, fishing intensity also increases, which has caused great pressure on *P. trituberculatus* resources. Protecting *P. trituberculatus* and achieving sustainable utilisation of resources are urgent problems that need to be solved. Therefore, protection and rational development of *P. trituberculatus* resources are important to accurately understand its spatial and temporal distribution. **Methods:** In this study, the temporal and spatial distribution predictive models of *P. trituberculatus* in the northern East China Sea were built on the basis of three analysis methods (generalised additive model [GAM], random forest [RF] and artificial neural network [ANN]) using bottom trawl survey data and environmental data from 2006 to 2007. The fitting and prediction performances of these three models were compared.

Result: Season and sea bottom temperature were the most important factors on the distribution of *P. trituberculatus*. The fitting performance of ANNs was better than those of GAMs and RFs, but its predictive performance was worse than those of GAMs and RFs. Therefore, RFs was the appropriate model in predicting the distribution of *P. trituberculatus* in the northern East China Sea. The abundance of *P. trituberculatus* was significantly higher in summer than in other seasons ($P<0.01$) and generally higher in the northern part of the study area than in the southern part in all seasons.

Key words: Artificial neural network, East China Sea, Generalised additive model, *Portunus trituberculatus*, Random forest, Species distribution models.

INTRODUCTION

Portunus trituberculatus is a large crab with a wide geographical distribution, which extends from the Northwest Pacific Ocean to the Indian Ocean. China is the main fishing country for *P. trituberculatus.* With the increase of fishing pressure, *P. trituberculatus* may face the risk of overfishing. Environmental factors greatly affect the spatial and temporal distribution of *P. trituberculatus* (Liao *et al.,* 2008). Therefore, predicting the spatial and temporal distribution and change trend of *P. trituberculatus* by using environmental factors is important for the management of fishery resource.

Species distribution models (SDMs) are numerical tools that combine observations of species occurrence or abundance with environmental factors. In the field of fishery research, the commonly used SDMs are generalised linear models (GLMs) and generalised additive models (GAMs). Nonparametric methods such as machine learning (ML) are also used and they have good predictive performance (Chen *et al.* 2017; Olden and Jackson 2002). Therefore, ML methods were widely used in many fields (Bhimanpallewar and Narasingarao 2021; Rana *et al.,* 2021; Zaborski and Grzesiak 2019).

In this study, traditional GAM, RF and ANN based on ML were built on the basis of the fishery resource survey data and several environmental factors and the spatial and temporal distributions of *P. trituberculatus* in the northern East China Sea were simulated on the basis of the information of fishery surveys. The environmental factors with significant effects on the distribution of *P. trituberculatus* were screened out. The fitting ability and prediction ability of the three models were compared to determine their reliability in predicting *P. trituberculatus* distribution. The predictive performance of two robust models, RFs and ANNs, was compared in different seasons and different number of stations.

School of Fishery, Zhejiang Ocean University, Zhoushan, 316000, China.

Corresponding Author: Yingbin Wang, School of Fishery, Zhejiang Ocean University, Zhoushan, 316000, China.
Email: yingbinwang@126.com

How to cite this article: Li, X., Wang, J., Liu, Y. and Wang, Y. (2021). Spatial and Temporal Distribution of *Portunus trituberculatus* in the Northern East China Sea based on Different Modelling Approaches. Indian Journal of Animal Research. 55(11): 1364-1370. DOI: 10.18805/IJAR.B-1365.

Submitted: 29-03-2021 **Accepted:** 01-06-2021 **Online:** 14-08-2021

MATERIALS AND METHODS

Data collection

Fishery surveys were carried out in the northern East China Sea (29°45'N-31°15'N, 121°46'E-124°15'E) in August and January 2006 and May and November 2007 (Fig 1). A single

otter trawl vessel with a main engine power of 184 kW was used. The vessel was towed for 1 h at a speed of 2 knots. The sea bottom salinity (SBS), sea bottom temperature (SBT), water depth (WD), chlorophyll and pH at each station were measured and recorded. In this study, modelling and data analysis were performed in the laboratory of School of Fisheries, Zhejiang Ocean University in 2020.

Statistical methods

GAM is a non-parametric extension of GLM, which is widely used to analyse the nonlinear relationship amongst variables (Yee and Mitchell 1991). RF is a tree-based ensemble ML tool that generates multiple regression or decision trees and it has good prediction ability (Breiman 2001). ANN is a computer model that performs well in nonlinear modelling involving a multitude of variables. The model is implemented by mgcv package, random forest package and nnet package in R (V 3.6.3).

Factor screening and model evaluation

We used the variance inflation factor (VIF) to examine the collinearity amongst predictive variables and deleted highly collinear variables before modelling (Sethi *et al.* 2012). The critical value of VIF was set to 3. Variables with VIF greater than 3 were considered collinear with other variables and excluded.

Predictive variables were selected by stepwise variable selection and variables were added to the model step by step from a null model. The predictive variables of GAMs, RFs and ANNs were selected using the Akaike information criterion, variance explained (VE) and saliency analysis, respectively. VE was used to compare the fitting capability amongst different models:

$$VE = \left(1 - \frac{Var\,(residual)}{Var\,(y)}\right) \times 100\%$$

Where

Var (*residual*) is the residual variance and *Var* (*y*) is the variance of original data. The higher the VE is, the better the model fitting.

Sensitivity analysis was performed to determine the relationship between predictive variables and response variable. Finally, the plot of the predicted response to each separate variable was drawn. For each GAMs and RFs, one response curve was drawn for each predictive variable. For ANNs, 50 response curves were drawn for each predictive variable because the relationship generated by ANNs based on the initial values was not constant.

The predictive performances of different models were evaluated using the cross-validation approach. In this study, we ran cross-validation 100 times to evaluate the predictive performances of the models. The relative root-mean-square error (RRE) and coefficient of determination (R^2) were used to evaluate the accuracy and precision of model prediction, respectively (Olden and Jackson 2001). R^2, which ranges from 0 to 1, measures the correlations between the observations and predictions. The higher the R^2 is, the stronger the correlation. RRE indicates the deviation between the observed value and prediction and a smaller value implies a better predictive performance.

$$RRE = \frac{\sqrt{\sum_{i=1}^{n} (Oi-Pi)^2}}{\frac{n}{O_{max}-O_{min}}} \times 100\%$$

Where

O_i is the ith observed value; P_i is the ith predicted value; n is the size of data in the cross-validation and O_{max} and O_{min} are the maximum and minimum values of observation, respectively.

The difference of R^2 between model fitting and model prediction indicated the overfitting degree of the model.

Prediction of distribution

Based on the environmental data at each station in each season, the study sea area was meshed with the grid size of 0.05° × 0.05° to obtain the coordinates of each grid centre point. The Kriging interpolation method was used to obtain the environmental data of the grid centre point. The best prediction model was selected by comparing the three models and the environmental data obtained by interpolation were substituted into the model to predict the distribution of *P. trituberculatus*. Ocean data view software was used to draw the *P. trituberculatus* distribution map.

Effect of season and number of stations

We analysed the predictive performance of RF and ANN in different seasons and different number of stations. Nine levels (8-16) were randomly drawn from 20 stations, from which the data were used to train the models. In different seasons, each level was simulated 10,000 times; R^2 and RRE were used as indicators to measure the predictive performance of the model under different seasons and number of stations. Larger R^2 and smaller RRE indicated better predictive performance of the model.

RESULTS AND DISCUSSION

Selection of the prediction variables and model fitting

The results of the VIF test showed no multicollinearity amongst the predictive variables. Therefore, we considered season, SBS, SBT, WD, chlorophyll concentration and pH as candidate predictive variables for modelling.

The predictive variables and their relative importance are different amongst the three models (Table 1). ANNs retained more predictive variables and their fitting degree was the highest amongst the three models. The fitting results of the three models showed that season and SBT were the two most important factors of the relative abundance of *P. trituberculatus* (Table 1). GAMs showed a relatively simple relationship between *P. trituberculatus* and environmental factors, whereas RFs and ANNs showed more complex relationships (Fig 2). The cross-validation results showed that the median R^2 of RFs was the largest and the median RRE of RFs was the smallest, which indicated that the predictive performance of RFs was the

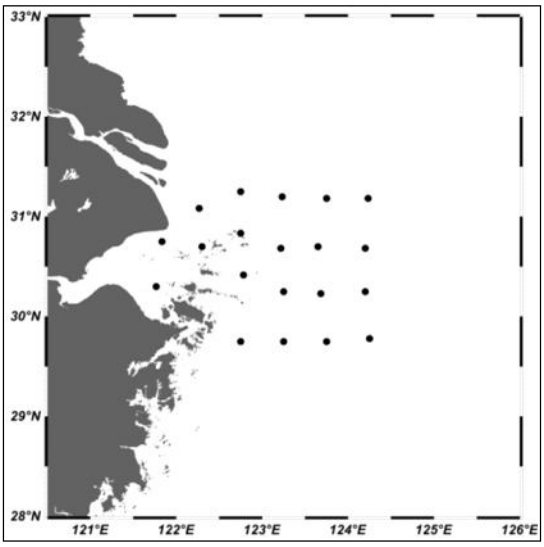


Fig 1: Survey station map of north East China Sea in 2007.

best amongst the three models (Fig 3a). BY contrast, the median R^2 of ANNs was the smallest and the RRE of ANNs was the largest, which indicated that the predictive performance of ANNs was poor (Fig 3b). ΔR^2 of GAMs, RFs and ANNs were 0.21, 0.15 and 0.46, respectively. RFs and ANNs had the lowest and highest overfitting degree, respectively. Considering the predictive performance and overfitting degree of the models, RFs were the best amongst the three models.

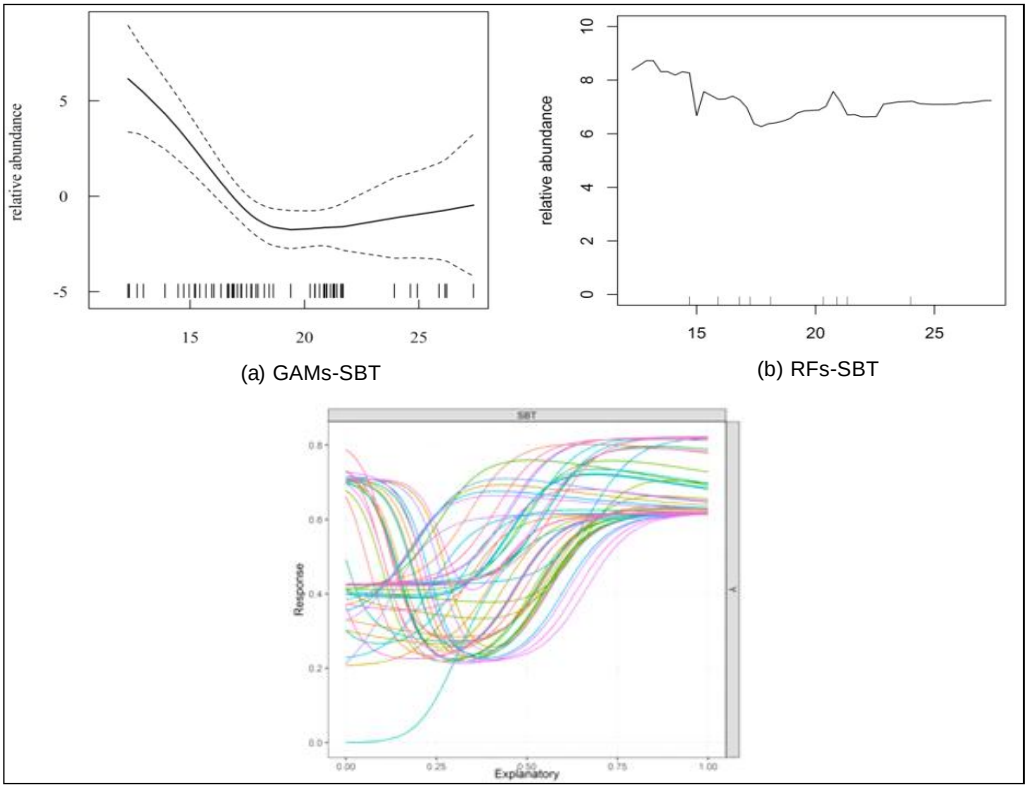
Mapping of *P. trituberculatus* distributions

Differences in *P. trituberculatus* distributions in different seasons were observed with regard to time and the relative abundance of *P. trituberculatus* in summer was significantly higher than that in the other seasons (Fig 4). With regard to space, the relative abundance of *P. trituberculatus* in the northern sea area was higher than that in the southern sea area, which was most apparent in spring and winter (Fig 4).

Table 1: Fitting results of three models.

Model	Relative importance (%)	Variance explained (%)
GAM	SBT (23.4) > season (12.8)	42.8
RF	Season (17.4) > SBT (16.5) > pH (14.9)	37.0
ANN	SBT (29.1) > pH (23.2) > depth (21.6) > Season (14.2) > SBS (12.0)	59.6

Note: Relative importance refers to the contribution of each prediction variable to the model (GAM was based on 'deviation explained', RF was based on the percentage of IncMSE; ANN was based on Garson-Goh algorithm).



(a) GAMs-SBT (b) RFs-SBT

(c) ANNs-SBT

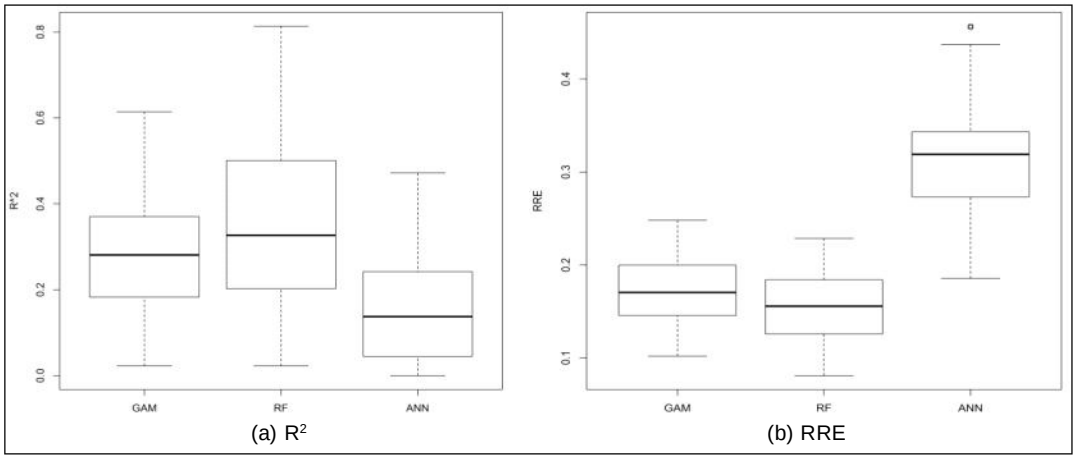
Fig 2: Relationship between predictive variables and response variables of different models.

Comparison between RFs and ANNs in different seasons and number of stations

Based on the variation of RRE and R^2, the predictive performance of RFs and ANNs was significantly affected by the number of stations and differences were observed amongst different seasons (Fig 5 and Fig 6). With the increase of the number of stations, R^2 of RFs and ANNs gradually increased (Fig 5) and the corresponding RRE gradually decreased (Fig 6). Therefore, with the increase of the number of stations, the predictive performances of the two models were gradually improved and the predictive performances of the two models in spring and winter were better than those in summer and autumn (Fig 5 and Fig 6). The predictive performance of RFs was better than that of ANNs.

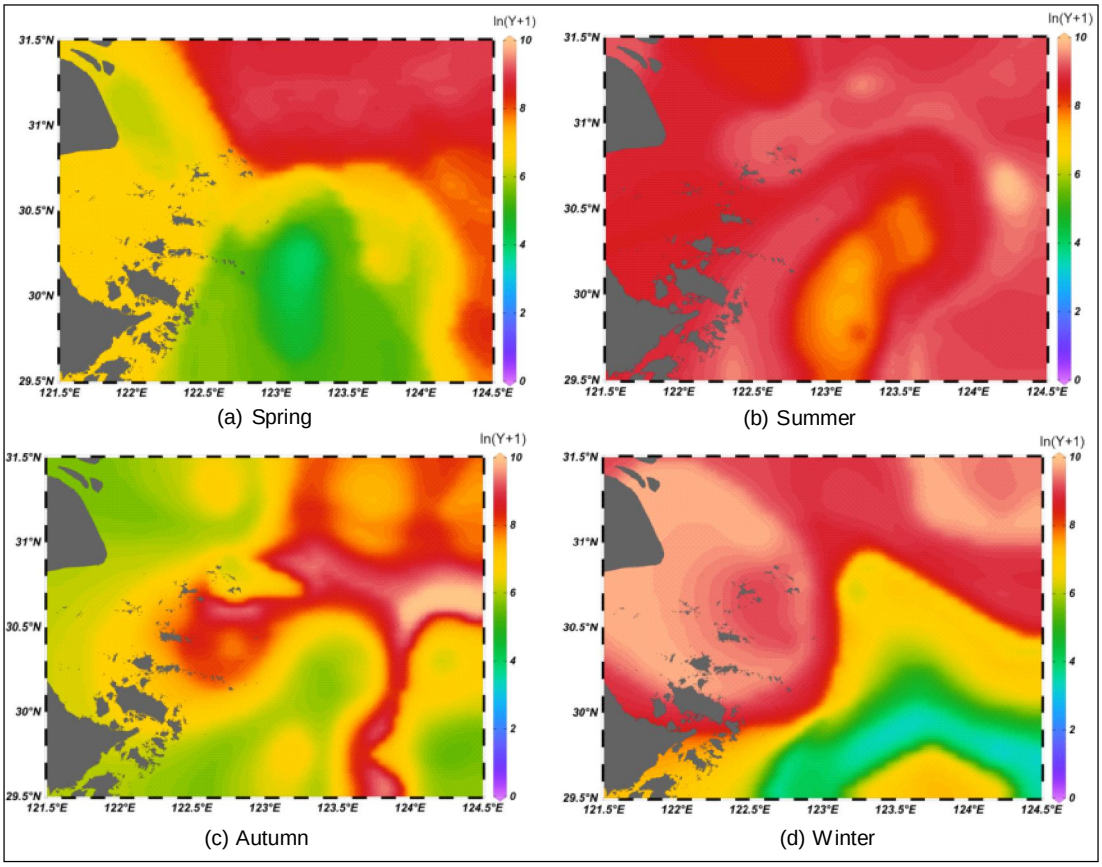
Comparison of models

Amongst the three models used in this study, ANNs have the best fitting performance and RFs have the best predictive performance. The results show that the fitting effect of the



(a) R^2 (b) RRE

Fig 3: Comparison of R^2 and RRE of the three models.

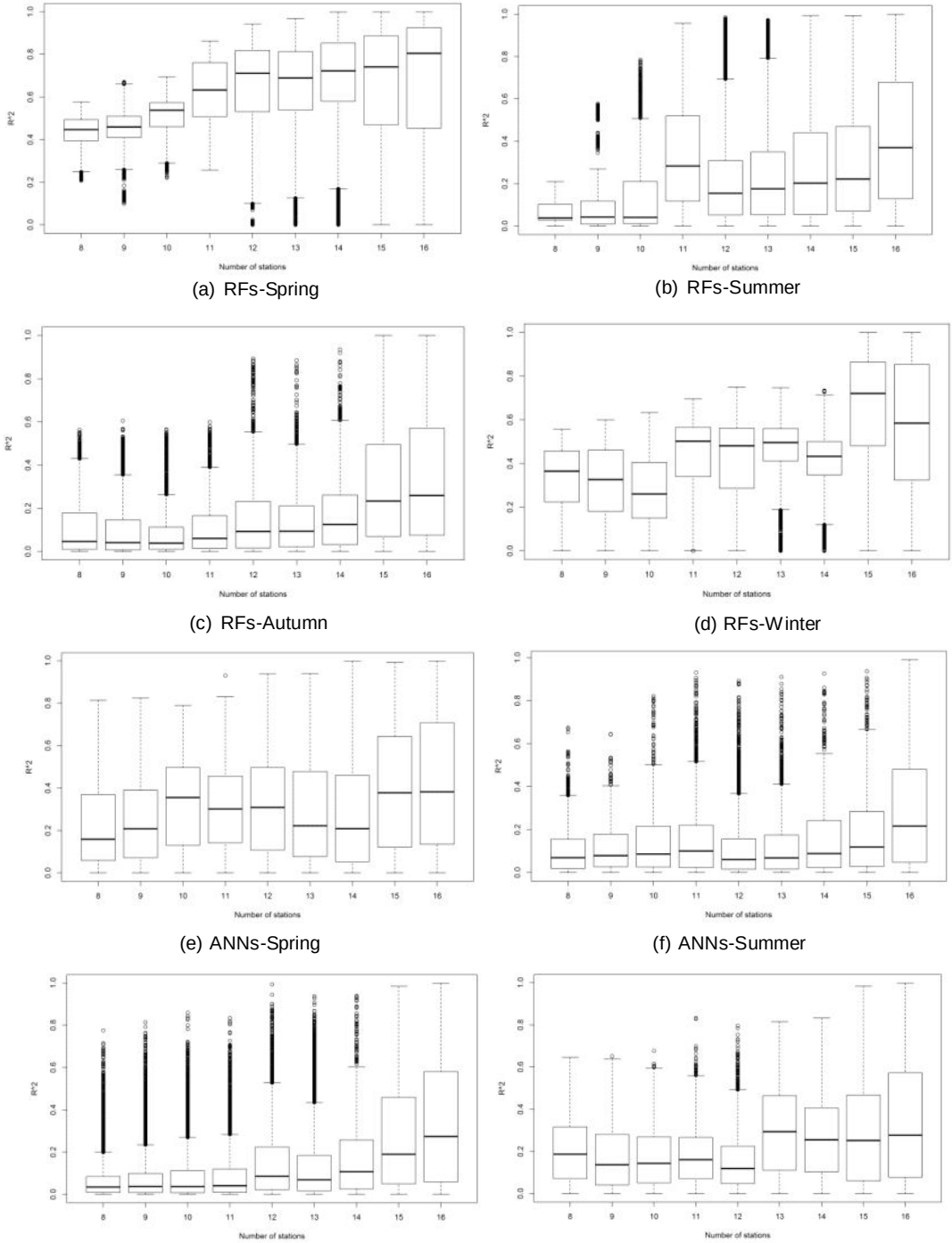


(a) Spring (b) Summer

(c) Autumn (d) Winter

Fig 4: Predicted spatial *P. trituberculatus* relative abundance distribution in different seasons: A: Spring, B: Summer, C: Autumn and D: Winter.

models on the training data set cannot guarantee the same predictive effect for the test data set and the performances of the three models are better on the training data set than on the test data set, which indicates an overfitting phenomenon. The models interpreted the sample noise but deviated from the interpretation of the real value; thus, the training data had a good fitting effect, but the prediction ability outside the training data set was not as good as that of training data (Luan *et al.,* 2018). Based on the value of ΔR^2, RFs have slight overfitting because of its ensemble learning, which improves the predictive accuracy by aggregating the results of multiple regression trees (Cai 2012).



(a) RFs-Spring (b) RFs-Summer

(c) RFs-Autumn (d) RFs-Winter

(e) ANNs-Spring (f) ANNs-Summer

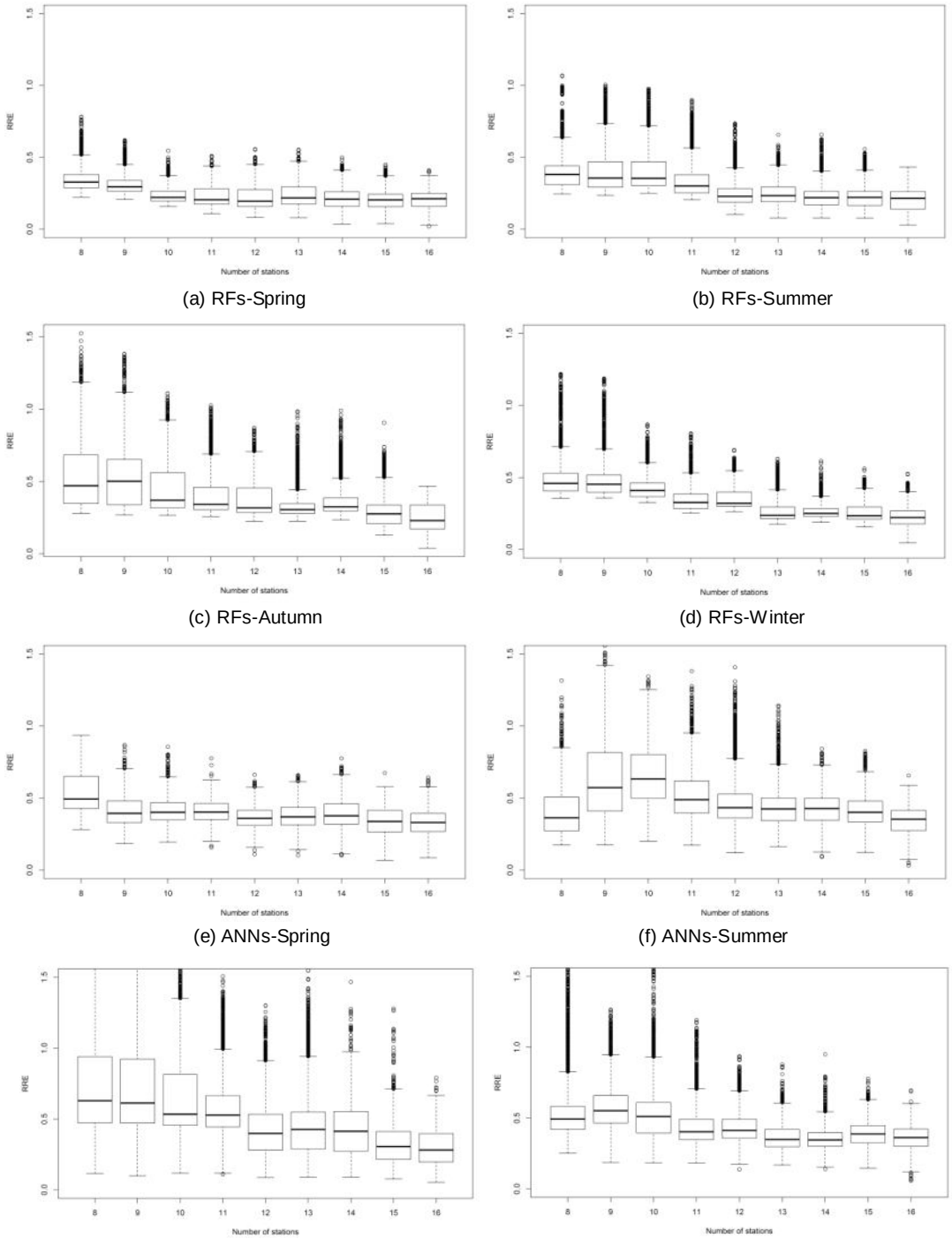
(g) ANNs-Autumn (h) ANNs-Winter

Fig 5: R^2 of RFs and ANNs in different seasons and number of stations.

Distribution of *P. trituberculatus* and influencing factors

The distribution of *P. trituberculatus* in the northern East China Sea has a clear seasonal variation because of the comprehensive effects of the environmental physicochemical factors, ocean currents and water masses in different seasons and the change of temperature amongst seasons (Yuan *et al.,* 2016).

Spring is the peak spawning season of *P. trituberculatus*. In spring, *P. trituberculatus* are less distributed along the Yangtze River Estuary, which may be affected by the runoff of the Yangtze River and no suitable spawning environment for *P. trituberculatus* is available. Summer is the foraging season of *P. trituberculatus* and the young *P. trituberculatus* hatched in that year fatten in the coastal shallow waters



(a) RFs-Spring (b) RFs-Summer

(c) RFs-Autumn (d) RFs-Winter

(e) ANNs-Spring (f) ANNs-Summer

(g) ANNs-Autumn (h) ANNs-Winter

Fig 6: RRE of RFs and ANNs in different seasons and number of stations.

(Yuan *et al.,* 2016). The relative abundance of *P. trituberculatus* is significantly higher in summer than in other seasons, which is closely related to the high SBT in summer. In autumn, the juvenile *P. trituberculatus* gradually grow and move to deep water. The water temperature along the coast gradually drops with the cold air going south. *P. trituberculatus* also migrate from north to south and from shallow water to deep water. In winter, the density of P. trituberculatus is evidently higher in the north than in the south and a banded area with less *P. trituberculatus* is observed in the south. This observation is consistent with the location of the Taiwan warm current entering the northern East China Sea from south to north in winter, which indicates that the Taiwan warm current may affect the distribution of *P. trituberculatus.*

In this study, SBT is considered as an important environmental factor affecting the distribution of *P. trituberculatus*. Most *P. trituberculatus* inhabit the sea floor and they are greatly affected by the bottom environmental factors. Thus, SBT is a factor affecting *P. trituberculatus* distribution.

In the four seasons, the predictive performance of RFs and ANNs showed a gradually increasing trend with the increase of the number of stations. The predictive performance of RFs and ANNs in spring and winter was significantly higher than that in summer and autumn, which may be related to the distribution of *P. trituberculatus*. In summer and autumn, the resource of *P. trituberculatus* is evenly distributed, whereas in spring and winter, the resource density is high in the north and low in the south. In general, fishery data with high contrast are suitable for stock assessment models to obtain accurate results. The degrees of difference of the distributions of *P. trituberculatus* are higher in spring and winter than those in summer and autumn. Therefore, the modelling of RFs and ANNs is more reliable in spring and winter, which may need relatively few survey stations. This result also indicates that setting different number of survey stations according to the resource distribution of *P. trituberculatus* is necessary to save costs.

The three models established in this study do not directly involve the ecological process and their interpretations depend on the existing understanding of the life history characteristics of *P. trituberculatus*. The environmental requirements of *P. trituberculatus* in different growth stages are also different. Therefore, in our future research, the life history of *P. trituberculatus* will be combined with the environmental factors to explore the comprehensive effect of *P. trituberculatus* at different growth stages.

ACKNOWLEDGEMENT

The authors gratefully thank the anonymous reviewers for their comments on the draft of this study. This research was supported by the National Key Research and Development Programme of China (2019YFD0901304 and 2017YFA0604902) and the Public Welfare Technology Application Research Project (LGN21C190009).

REFERENCES

Bhimanpallewar, R.N. and Narasingarao, M.R. (2021). Evaluating the influence of soil and environmental parameters in terms of crop suitability using machine learning. Indian Journal of Agricultural Research. DOI: 10.18805/IJARe.A-4942.

Breiman, L. (2001). Random Forests. Machine Learning. 45: 5-32.

Cai, L.L. (2012). Model selection of random forest and its parallelization. MESc Thesis, Harbin Institute of Technology.

Chen, Y., Chen, X., Guo, L., Fang, Z. and Wang, J. (2017). Comparison of Fishing Ground of Skipjack Based on BP Neural Network in the Western and Central Pacific Ocean. Journal of Guangdong Ocean University. 37(6): 65-73.

Geman, S., Bienenstock, E. and Doursat, R. (1992). Neural networks and the bias/variance dilemma. Neural Computation. 4: 1-58.

Liao, Y.Y., Xiao, Z.P. and Yuan, Y.Y. (2008). Temperature tolerance of larva and juvenile of Portunus trituberculatus. Acta Hydrobiologica sinica. 32(4): 534-543.

Luan, J., Zhang, C., Xu, B., Xue, Y. and Ren, Y. (2018). The relationship between habitat distribution characteristics and environmental factors of Charybdis bimaculatus in Haizhou Bay. Journal of Fisheries of China. 42(6): 889-901.

Olden, J.D. and Jackson, D.A. (2002). A comparison of statistical approaches for modelling fish species distributions. Freshwater Biology. 47: 1976-1995.

Olden, J.D. and Jackson, D.A. (2001). Fish-habitat relationships in lakes: gaining predictive and explanatory insightby using artificial neural networks. Transactions of the American Fisheries Society. 130: 878-897.

Rana, E., Gupta, A.K., Singh, A., Ruhil, A.P., Malhotra, R., Yousuf, S. and Ete, G. (2021). Prediction of first lactation 305-day milk yield based on bimonthly test day milk yield records in murrah buffaloes. Indian Journal of Animal Research. 55(4): 486-490.

Sethi, S.A., Dalton, M. and Hilborn, R. (2012). Quantitative risk measures applied to Alaskan commercial fisheries. Canadian Journal of Fisheries and Aquaticences. 69: 487-498.

Yee, T.W. and Mitchell, N.D. (1991). Generalized additive models in plant ecology. Journal of Vegetation Science. 2: 587-602.

Yuan, W., Jin, X. and Shan, X. (2016). Population biology and relationship with environmental factors of swimming crab in the Changjiang river estuary and adjacent waters. Fisheries Science. 35(2): 105-110.

Zaborski, D. and Grzesiak, W. (2019). Utilization of boosted classification trees for the detection of cows with conception difficulties. Indian Journal of Animal Research. 55(3): 359-364.