Indian Journal of Geo Marine Sciences
Vol. 49 (11), November 2020, pp. 1729-1741

Use of different modeling approach for sensitivity analysis in predicting the Catch per Unit Effort (CPUE) of fish

V K Yadav*[a,b], S Jahageerdar[b] & J Adinarayana[a]

[a]Centre of Studies in Resource Engineering (CSRE), Indian Institute of Technology, Bombay – 400 076, India
[b]Central Institute of Fisheries Education (CIFE), Panch Marg, Off Yari Road, Versova, Andheri (W), Mumbai – 400 061, India
*[E-mail-vinodkumar@cife.edu.in]

Received 14 December 2018; revised 09 December 2019

The contribution (Sensitivity analysis) of four variables, namely chlorophyll-*a* (Chl-*a*), sea surface temperature (SST), photosynthetically active radiation (PAR) and diffuse attenuation coefficient (Kd_490 or Kd) in predicting the Catch per Unit Effort (CPUE) of fish was evaluated using simple General Linear Model, Generalized Linear Model (GLM), Generalized Additive Model (GAM) and different explanatory methods of Artificial Neural Networks (ANN) technique. The models were assessed for their accuracy in determining the relative importance of the four variables in predicting the CPUE. GAM was an improvement over the General Linear Model, while ANN was found better than GAM. The six explanatory methods which can give the relative contribution or importance of variables were compared using ANN modeling techniques: (i) Connection weights algorithm, (ii) Garson's algorithm (iii) Partial derivatives (PaD) (iv) Profile method (v) Perturb method, and (vi) Classical stepwise (forward and backward) method. Our results showed that the PaD method, Profile method, Input perturbation (50 % noise), and Connection weight approaches were only consistent in identifying the two most important variables (Chlorophyll-*a* and Kd) in the network. The distribution of profile plot & partial derivative helped indirectly in finding the other three variables in decreasing order of importance (PAR > fishing hour > SST). It was observed that the significance (sensitivity) of independent variables under GAM and explanatory methods of ANN were similar.

[**Keywords**: Artificial Neural Networks, Catch Per Unit Effort, Generalized Additive Model, Generalised Linear Model, Relative importance, Sensitivity analysis]

Introduction

Fish catch rates are expressed as Catch per Unit Effort (CPUE) which is used as the relative measure of the abundance of a fish[1]. This is widely used in fisheries management and marine conservation efforts. The estimation of the total catch of fish per hour (in kg per hour) is represented as CPUE. In ecology, normally, the prediction models are based on linear relationships with environmental variables[2] as the error in the data followed a normal distribution. But there is always a concern of satisfying this normal assumption[3], so the new non-linear modeling methods such as generalized linear models (GLM) and generalized additive models (GAM) are being fostered and are in wide use[4-7].

Contrary to, Artificial Neural Networks (ANNs) is more unique and widely used in ecology due to its ability to model non-linear relationships[8]. The special features include: (i) to store the knowledge and use whenever required, (ii) ability to recognize patterns, in spite of noise presence, (iii) ability to take the past observation into consideration, and (iv) make a conclusion, and discernment about new situations. There are highly non-linear and complex relationships between the environment variables and fishery and ANNs is very strong and powerful to deal the non-linear relationships[9,10] and has been widely chosen by many authors over linear statistical models[2,11-13]. This method has become increasingly popular in the analysis of ecological phenomena[14-16].

In ANN, the output value is generated with entered input variables without knowing the process that occurs within the network[17]. The description of how explanatory or independent variables (input) and dependent variables (output) are associated is unaccounted in the network. So, ANNs are normally appraised as black boxes, and so it is enthralling to study from their explanatory point of view[18].

The significance of variables in ANN models has been explored by many authors, and algorithms have been proposed. In the majority of works, pruning

methods are used to eliminate irrelevant input [19-21] that reduces the size of the network and minimizing the redundancy in the training data. Nevertheless, although the good prediction is required in fish catch or assemblage, knowing what beneficence each environmental variable makes is of most important. This informative and interpretive aspect of ANNs with different explanatory methods was discussed here. These methods were used to ascertain the significance and relative contribution of each explanatory variable to the output.

There is numerous literatures[18,22,23] pertaining to contribution study of variables in the different domains, but the contribution made by said environmental factors using explanatory methods of ANN in the fishery is scarce (limited).

The contribution analysis has been performed by using six different methods: (i) Connection weights algorithm, (ii) Garson's algorithm (iii) Partial derivatives (PaD)- calculates the partial derivatives of the dependent variables (output) with respect to the input variables; (iv) Profile method- is a variation of one input variable while the others are kept constant at a fixed value; (v) Perturb method- is input variables perturbation; and (vi) classical stepwise (forward and Backward) method) – is the change in the error value when forward (adding) or backward (elimination) step of the input variables (independent variables) is operated.

Although the GLM, the GAM, and ANNs have been used in many different domains involving many different variables, the three approaches have seldom been compared in the context of fisheries with such variables as chlorophyll-*a* (Chl-*a*), sea surface temperature (SST), photosynthetically active radiation (PAR) and diffuse attenuation coefficient (Kd_490 or Kd). The importance of these variables to fisheries is discussed later. Damalas *et al*.[6] used GAM and GLM model to examine the relative influence of different environmental variables on swordfish catch. Madhavan *et al*.[24] used ANN to predict the Mackerel landing using the environmental variables- SST, chlorophyll-*a*, and PAR.

The present study sought to rank, using GLM, GAM, and explanatory methods of ANNs, the above four variables in terms of their contribution to predicting the CPUE, and also assessed the models for their accuracy in ascertaining the relative significance of the four variables in predicting the CPUE.

Materials and Methods

Data used

The validated data of potential fishing zone advisory of Gujarat coastal region was obtained from Indian National Centre for Ocean Information Services (INCOIS), Hyderabad, India, from December 2007 to December 2009. The data included the fishing hour (duration of the trip), latitude and longitude of each fishing set, date of fishing and total catch. CPUE values were estimated as the total catch of fish (in kg per hour) and (ii) daily or composite days chlorophyll-*a* (Chl-*a*), sea surface temperature (SST), photosynthetically active radiation (PAR) and diffuse attenuation coefficient (Kd_490 or Kd) corresponding to fish catch location or area were obtained from Moderate Resolution Imaging Spectroradiometer (MODIS) sensor with a spatial resolution of 4*4 km. Retrieval of daily or composite days value of the above said variables depends upon the fishing activity (single day or multiple consecutive days). The L1A MODIS images of said parameters were processed in SeaDAS (version 7.3.1) software.

The L1A MODIS images data were processed from level L1A to L1 Geo for geometric corrections and similarly L1Geo to L1B for radiometric corrections to extract L2 products of chlorophyll-a, Kd, PAR, SST, etc. The corresponding to fish catch point (Latitudes and Longitudes), ASCII file containing the value of the product was acquired and also from MODIS level 3 standard binned images archived by the Ocean Biology Processing Group (OBPG), composite 8 days or fortnightly data was obtained as an ASCII file.

The Trawl net & Gillnet gear were used in the sampling and had no significant effect on the fish catch (CPUE); also, latitude and longitude had no significant effect on CPUE (using GLM & GAM model). So in this study, our attention was on the impact of environmental variables (chlorophyll-*a*, Kd, PAR and SST) along with the fishing effort (fishing hour) on CPUE.

Importance of environmental variables

a. Chlorophyll-a concentration (Chl-a)

Chlorophyll-*a* is the primary phytoplankton pigment for photosynthesis of marine algae in the ocean. The concentration of Chl-*a* is often considered as an index of biological productivity, and in an oceanic environment, it can be related to fish production. The concentration of Chl-*a* (mg/m^3) was taken as one of the inputs into the prediction models.

b. Sea Surface Temperature (SST)

Sea surface temperature (°C) affects the activity, movement, feeding, and reproduction of fish, especially of tropical fish, and therefore formed one of the inputs into the models.

c. Photosynthetically Active Radiation (PAR)

Photosynthetically active radiation is the amount of light available for photosynthesis and is defined as the quantum energy flux from sunlight in the 400–700 nm wavelength band. Since some fish species (such as mackerels) are herbivores[24], PAR, that is the number of photons received by a unit area over a specified amount of time, or the photosynthetic photon flux density (PPFD, expressed per square meter per day) is considered one of the significant biophysical parameters.

d. Diffuse attenuation coefficient (Kd)

Diffuse attenuation coefficient (Kd) is a measure (m^{-1}) of the transparency of a column of water and important because some species of fish (such as tuna) need light to locate their prey and thus affect the amount of food for such species.

The above four independent environmental variables were taken as inputs into the models to predict the catch of fish more reliably.

Methods

The association of independent variables with the dependent variable (CPUE) was examined using Generalized linear model (GLM) and Generalized Additive Model (GAM) and Artificial Neural Network (ANN) technique. The CPUE data had skewed distribution; the logarithmic transformation was applied to make a normal distribution (Fig. S1).

Generalized Linear Model (GLM) & Generalized Additive Model (GAM)

The generalization of linear regression models that allow non-linearity and non-constant variance structures in the data[25] is called GLM. There is an assumed relationship, called a link function, which tells how the expected value of the response (output) variable is related to a linear combination of a set of explanatory (input) variables[5]. Data are assumed to fall within one of the several families of probability distributions, including normal, binomial, Poisson, negative binomial, or gamma[26].

A GAM also uses a link function to establish a relationship between the mean of the output variable and a 'smoothed' function of the input variable(s). The strength of GAMs is their ability to deal with highly non-linear relationships between the output and the set of input variables[5]. The analysis under GLM and GAM was done using R software. The details about GLM and GAM are beyond the scope of this paper, and the same can be seen in Guisan *et al*.[4].

Log (CPUE) was modeled in three steps. A simple general linear model was applied to gain insight into how the independent variables related to the prediction of Log (CPUE) in the datasets. And then different generalized linear model (GLM) using different distributions and link functions were used, and afterward, switched to generalized additive models (GAMs) and tested whether they have an improvement over the linear approach (GLM).

Artificial Neural Network (ANN) modeling

The feed-forward multi-layer neural network architecture was used (Fig. S2). The back propagation error training algorithm was used to train the network. The weights were adjusted using the back-propagated error computed between the observed and the estimated results. The network consisted of three layers-Input, hidden and output layers.

The selection for the number of nodes/neurons in the hidden layers is an important aspect of neural networks. This is determined by taking different possible configurations of network, and the best one is selected based on good generalization ability of networks along with the best compromise between bias and variance[27]. A network with one hidden layer of eight neurons had been selected in this study. The different methods were applied to analyze the importance/contribution of the various input variables on the calibrated ANN mode. The analysis was done in MATLAB (R2012a) software.

We used the k (=10) fold cross-validation method[25] to check the superior model between ANN and GAM model as both deal with the non-linear relationship.

Methods for testing the contributions of the different variables in ANN

PaD (Partial Derivative) method

This method is giving two results. The first is a profile, which tells how the changes in variations of the output variable are affected by small changes in input variables, and second is the classification of the relative contribution of each input to the output[18]. The partial derivatives of the ANN output with respect to the input were computed to obtain the profile of the variations of the output for small changes of one input variable[18], for a network with n_i inputs, one hidden layer with n_h neurons, and one output (i.e. $n_o = 1$), the

partial derivatives of the output y_j with respect to input x_j (with j=/1, . . ., N and N the total number of observations) are:

$$d_{ji} = S_j * \sum_{h=1}^{n_h} w_{ho} * I_{hj}(1 - I_{hj}) * w_{ih}$$

(Gevrey *et al.*[18]) … (i)

Where, $S_j = y_j*(1-y_j)$ is the derivative of the output neuron with respect to its input, I_{hj} is the output of neuron h (h = 1 to n_h) of the hidden layer, which is connection weights between h^{th} hidden neuron and the output neuron and w_{ih} is the connection weights between the i^{th} input neuron and the h^{th} hidden neuron.

The graphs of the partial derivatives with respect to each corresponding input variable were plotted, and the effect on output variables by input variable was determined. If the partial derivative is negative, the output variable will decrease when the input variable increases, and inversely if the partial derivative is positive, the output variable increase when input variable also increases[18].

The second result of PaD was used to find the significance of the ANN output with respect to an input on the given set of data. It is calculated by a sum of the square partial derivatives obtained per input variable:

$$SSD_i = \sum_{j=1}^{N} (d_{ji})^2$$ (Gevrey *et al.*[18]) … (ii)

SSD (Sum of Square Derivatives) values were obtained for all input variables. The SSD values allow ranking of the variables according to their increasing contribution to the output variable in the model. The input variable which influences the output variable most has the highest SSD value.

Perturb method

This method evaluates the change in the mean square error of output by adding a small amount of noise increased in steps of 10-50 % of the input value to each input variable[18] while holding all other input variables at their observed values. The relative significance of the input variables was determined by finding the change in mean square error for each input perturbation[18].

Connection weights algorithm

In this algorithm, the product of weights between input-hidden and hidden-output connection through each input neuron and output neuron is calculated and then sums the products across all hidden neurons are computed[22].

The relative importance of a given input variable can be defined as:

$$RI_x = \sum_{y=1}^{m} w_{xy} w_{yz}$$ (Ibrahim[23]) … (iii)

Where, RI_x is the relative importance/contribution of input neuron *x*,

and $\sum_{y=1}^{m} w_{xy} w_{yz}$ is the sum of the product of the final weights of the connection from input neuron to hidden neurons with the connection from hidden neurons to output neuron (where y is the total number of hidden neurons, and z is output neurons).

Garson's algorithm

This algorithm partitions hidden-output connection weights into components associated with each input neuron using absolute values of connection weights[28]. The direction of the relationship between the input and output variables is not taken care of in this algorithm.

Profile method

This method was suggested by Lek *et al.*[9,29]. Here, the input variable is studied consecutively, keeping the values of the remaining variables fixed. Each variable is divided into a fixed number of equal intervals between its minimum and maximum values. All variables except one are set initially, at their minimum values, then successively at their first quartile, median, third quartile, and maximum[18]. The median predicted response value across the five summary statistics is calculated, and the relative importance of each input variable is determined by the magnitude of its range of predicted response values (i.e., maximum-minimum).

Stepwise method

This method consists of adding or rejecting successively one input variable and the effect on the output variable is estimated. The input variables can be ranked according to their importance based on the changes in Mean Square Error (MSE), in several different ways depending on different arguments[18]. The two stepwise (forward and backward) modeling approaches were adopted and the detail of this method can be seen in Mair *et al.*[30.]

Results and Discussion

Generalized Linear Model (GLM) & Generalized Additive Model (GAM)

Initially, a simple general linear model was applied, and results are shown in Table 1 & 2. Type III analysis under the general linear model revealed that 2 out of five main effects were significant (Table 1 & 2). The simple general linear model appeared to be inadequate because the plots of residual values versus predicted values were not randomly scattered, i.e., no pattern appears in the residual values (Fig. S3). Also, the QQ plot does not confirm the normal distribution of residual as most points fell on either side of the line (Fig. S3). Scatter plots of transformed catches against the independent variables showed indications of non-linearity for all of the variables and depicted in Figure S4 of supplementary data.

The influence of each variable can be assessed by the regression coefficients (B). The two significant variables have a negative relationship with log (CPUE). The coefficient of determination was 0.24 (24 %). The catch was not increased proportionally with an increase of fishing hours that resulted in a decrease of CPUE and, thus, a negative relationship with CPUE. PAR has an inverse relationship with Chl-*a*[24] (Table S1), which is indirectly related to fish assemblage, and therefore PAR has a negative correlation with CPUE.

Generalized Linear Model (GLM)

Akaike information criterion (AIC), Bayesian information criterion (BIC), and Akaike weights (w_i) were used to test the different dimensions of a model under GLM. An application of this method on fishery data can be found in Dick (2004)[31]. Comparison of AIC, BIC, and Akaike weights (w_i) gave massive evidence for the Gamma distribution, relative to all the other candidate models (Table 3). GLM with the Gaussian distribution and 'identity' link function is

Table 1 — Fitting of general linear model, relating log (CPUE) to the two significant predictive factors in the Log (CPUE) prediction

Source	Sum of square (Type III)	d.f	Mean square	F-ratio	*p*-value
fish_hour	1.161	1	1.161	27.942	**.000**
Chl_*a*	.001	1	.001	.031	.861
KD_490	.037	1	.037	.881	.350
PAR	.318	1	.318	7.646	**.007**
SST	.000	1	.000	.012	.914
Error	5.482	132	.042		
Total	327.563	138			
Corrected Total	7.211	137			

R^2 = 24 % (Adjusted R Squared = 21.1 %), Akaike information criterion (AIC) = 200.04
Dependent variable: log(CPUE)

Table 2 — Parameter estimate under general linear model

Parameter Estimates

Dependent Variable: log_cpue

Parameter	B	Std. Error	t	Sig.	95 % Confidence Interval	
					Lower Bound	Upper Bound
Intercept	-1.233	30.923	-.040	.968	-62.400	59.935
fish_hour	-.003	.001	-5.286	.000	-.004	-.002
Chl_*a*	.141	.801	.176	.861	-1.443	1.725
KD_490	-1.107	1.179	-.939	.350	-3.438	1.225
PAR	-.006	.002	-2.765	.007	-.010	-.002
SST	.133	1.232	.108	.914	-2.305	2.571

Table 3 — AIC, BIC, Wi (Akaike weight) and significant variables for several distributions in the GLM model for factor affecting CPUE abundance (Higher CPUE value)

Distribution	AIC value	BIC value	Δi	Wi	Significant variables
Gaussian	200.04	220.53	1.39	0.231469	Fishing hour and PAR
Gamma	198.65	219.141	1	0.463796	Fishing hour and PAR
Log-Normal	199.49	219.98	0.84	0.304736	Fishing hour and PAR

equivalent to a simple general linear model with all continuous predictor variables.

Generalized Additive Model (GAM)

GAM model building was applied to the data, with different error distributions and spline functions and found that cubic regression splines were favorable over other (as it has highest adjusted R^2 (32.1 %) and lowest AIC value (183.76; Table 4)

GAM model was written in the following way-

log(cpue) ~ s(fish_hour, bs = "cr") + s(Chl-a, bs = "cr") + s(KD_490, bs = "cr") + s(PAR, bs = "cr") + s(SST, bs = "cr")

cr = Cubic regression splines, bs = B-splines

Effects of explanatory variables

The best model (GAM with gamma distribution and cubic regression splines function) explained 32.1 % of the variance in predicting CPUE (Table 5). Model analysis indicated that Fishing hour, KD_490, and PAR have a significant effect on CPUE/ (log_cpue). Solid lines are cubic regression spline smoothers (Fig. 1).

Abundance related to fishing hour and PAR, no unique pattern were seen as CPUE fluctuate throughout their range (Figs. 1a & c). Abundance related to SST (Fig. 1b) fluctuated throughout the temperature range studied, however higher CPUE values were observed in temperatures from around 24 °C to less than 25 °C. Most catches were made, or CPUE were abundant where Kd_490 is less than 0.2 m^{-1} (Fig. 1d) and Chl-*a* is less than 2 mg/m^3 (Fig. 1e).

GAM had an improvement over the linear approach (Simple General Linear Model), as it explained an additional 11 % of the variance (Table 5). Our results indicated that Fishing hour, Kd_490 & PAR played the most significant role in the model substantially affecting catches, while the remaining features: Chl-*a* and SST, were subsequent constituents. But as there is a high degree of correlation between Kd_490 and Chl-*a* (Table S1), so Chl-a would be equally important as Kd_490. Smooth function (graph) created using GAM Models were shown in Figure 1.

Artificial neural network models

Predictive capacity

Average recognition and prediction percentages vary quickly with the number of neurons in the hidden layer. Considering the values of MSE (mean square error) and R (correlation coefficient) with all the three samples (training, validation and testing data) obtained, 8 hidden neurons were selected, where there was marginal variation in MSE and correlation coefficient among all three samples to take care of over fitting and poor generalization (Table S2 & Fig. 2). The overall adjusted R^2 in ANN is 33 % (Table S2), which is slightly higher than the adjusted

Table 4 — Comparison of different GAM models

Distribution	Spline function	AIC value	R^2 adjusted	Significant variables
Gamma	Cubic Regression spline	**183.76**	**32.1 %**	Fishing hour, Kd_490 and PAR
	Duchon splines	196.52	27.1 %	Fishing hour and PAR
	Thin plate regression spline	199.19	26.6	Fishing hour and PAR
Gaussian Distribution	Cubic Regression spline	**200.04**	**29.1**	Fishing hour, Kd_490 and PAR
	Duchon splines	196.72	28.6	Fishing hour and PAR
	Thin plate regression spline	197.97	26.8	Fishing hour and PAR
Log-Normal	Cubic Regression spline	**184.29**	**30.9**	Fishing hour, Kd_490 and PAR
	Duchon splines	200.06	26.7	Fishing hour and PAR
	Thin plate regression spline	198.23	27.2	Fishing hour and PAR

Table 5 — Comparison between General Linear Model, GLM and GAM Model

Model	Adjusted R^2	AIC value	Significant variables
General Linear Model	21.1	200.04	Fishing hour and PAR
GLM with Gaussian distribution with an identity link function	21.1	200.04	Fishing hour and PAR
GLM with gamma distribution	---------	198.65	Fishing hour and PAR
GAM (Cubic regression spline function with gamma distribution)	32.1 %	183.76	Fishing hour, Kd_490 and PAR,

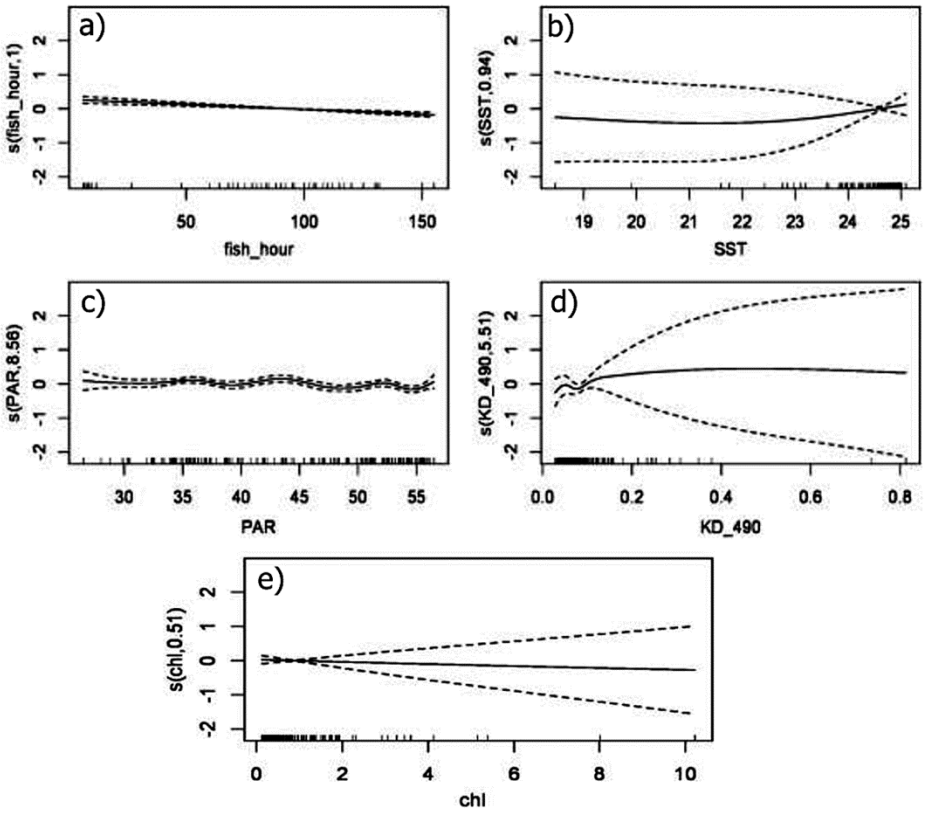

Fig. 1 — Effects of five predictor variables [(a) Fishing hour, (b) SST, (c) PAR, (d) Kd_490, and (e) Chl-*a*] on log-transformed CPUE. Dashed lines (or upper and lower brackets) indicate centered 95 % confidence intervals. A relative density of data points is shown by the 'rug' on the *x*-axis

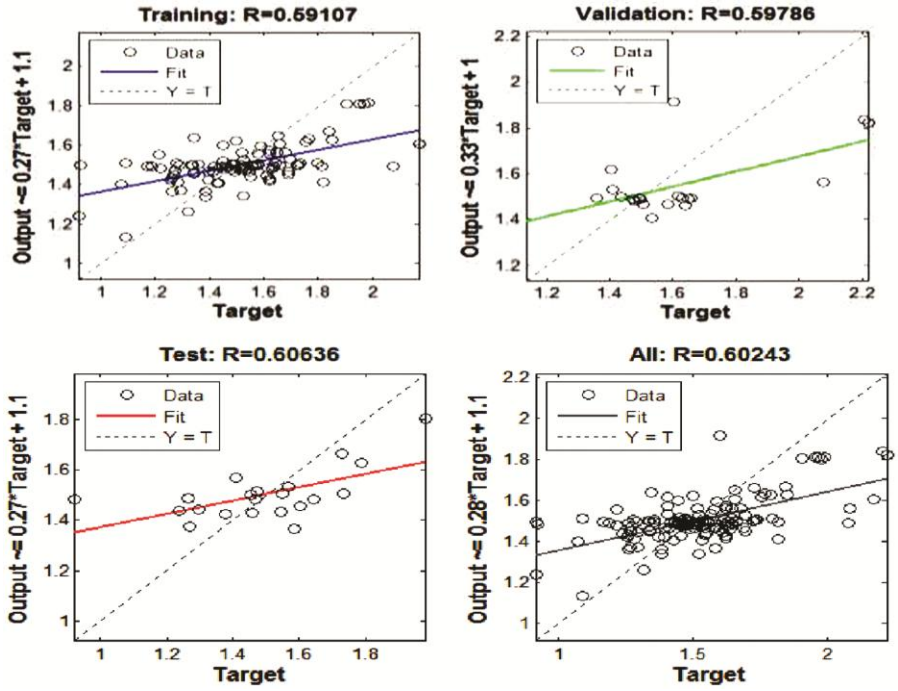

Fig. 2 — The relationship between output and the target variable in ANN model

R^2 (= 32.1 %) in the GAM model (Table 5). Damalas *et al.*[6] observed 36.1 % and 46.7 % variance in swordfish CPUE under GLM and GAM model, respectively, where input variables were- gear types, month, year, latitude, longitude, SST, Lunar index, and bathymetry. Usman *et al.*[32] observed 26 % variance in CPUE of skipjack due to Chl-*a*. The results of the correlation coefficient obtained in ANN are very close in both the learning set and the testing set. Hence, the obtained ANN structure can be used for the second step, using the complete database for sensitivity analysis.

Comparison between modeling techniques - ANN and GAM

K (= 10) fold cross-validation method (Hastie *et al.*)[25] was implemented to confirm the superiority of the ANN approach on GAM. Data (N = 138) were partitioned into ten almost equal-sized subsets, the "training" set comprised of the nine subsets while the remaining subset was used as the "test" set. After models were fitted, prediction errors were used to compute the Average prediction accuracy in terms of mean square error (MSE). Results showed that ANN was slightly better as compared to GAM approach (ANN, MSE = 0.0018; GAM, MSE = 0.0026)

Contributions of input variables in ANN

PaD method

With reference to Figure 3(A-E)

a. The partial derivative values of log(CPUE) with respect to the fishing hour are negative for low values of the fishing hour and near zero for the higher values. Log(CPUE) decreases with the increase of fishing hours till it becomes constant at high values of the fishing hour.
b. The partial derivative values of log(CPUE) with respect to chlorophyll-*a* (are negative for low values of chlorophyll-*a* and near zero for higher values. Log(CPUE) increased rapidly in a positive direction and then decreased with an increase of Chl-*a* value.
c. The partial derivative values of log(CPUE) with respect to Kd are all negative. Log(CPUE) decreases with the increase of Kd having a negative slope.
d. The partial derivative values of log(CPUE) with respect to PAR are all negative: an increase of PAR leads to a decrease of log(CPUE). For high values PAR, the partial derivative values approach zero; thus, log(CPUE) tends to become constant.

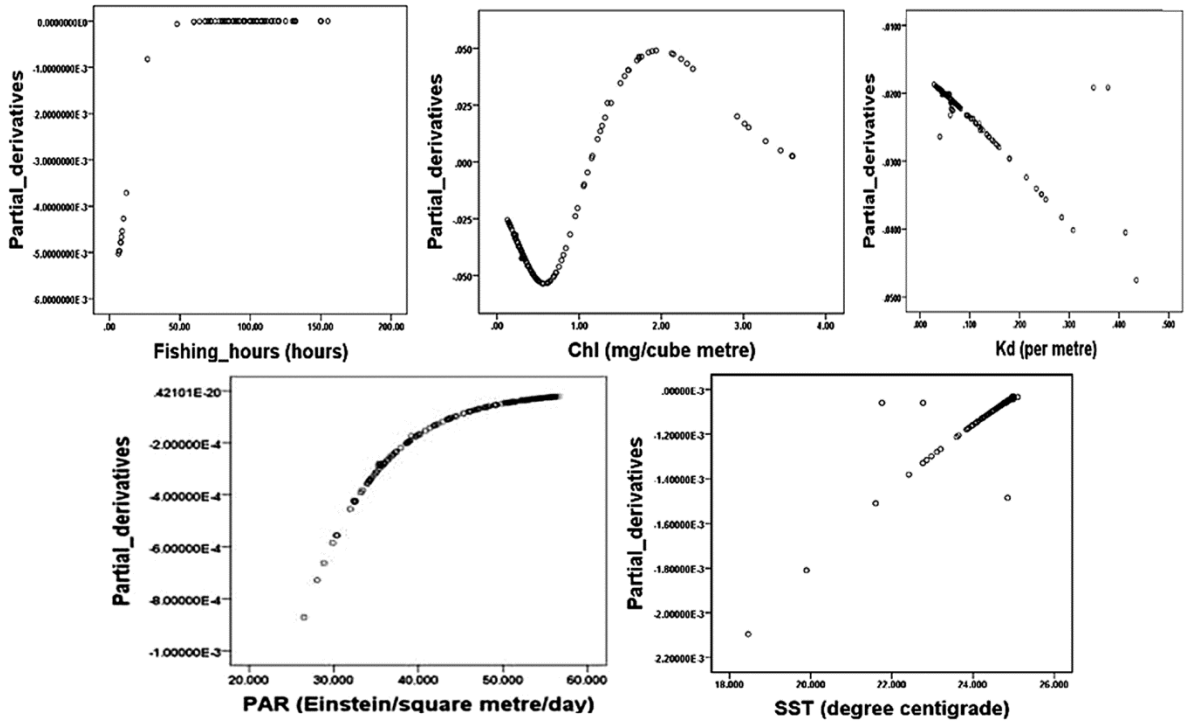

Fig. 3 — Partial derivatives of the ANN model response of CPUE with respect to each independent variable (PaD algorithm, Derivatives profile); (A) fishing hour; (B) Chl-*a*; (C) Kd; (D) PAR; and (E) SST

e. All the partial derivative values of log(CPUE) with respect to SST are negative: an increase of SST leads to a decrease of log(CPUE)

Table S3 shows the relative contributions resulting from the application of the PaD method. Chl-*a* was the highest contributed variable (73.95 %), followed by Kd (25.90 %). However, the contribution of other variables was very low.

Profile method

The influence of the independent variables on predicting log(CPUE) may exhibit any number of relationships. The summary of the response curve observed in our example with reference to Figure 4(A-E) is given below-

a. Influence of SST on the log(CPUE) – input variables contributes greatest at intermediate values, and exhibit decreasing influence at low and high values.

b. Influence of fishing hour on log(CPUE) – decreasing response curve-input variable contributes decreasingly at increasing values: Influence of fishing hour on the log(CPUE).

c. Influence of chlorophyll-*a* greatest at a low value and exhibits minimal influence at intermediate and high values.

d. Influence of Kd on the log(CPUE) – right skewed response curve-Input variable contributes greatest at low values and exhibits minimal influence at intermediate and high values.

e. Influence of PAR on the log(CPUE) – input variable contributes greatest at low and intermediate value but decreases influence at a high value when all other variables are at an intermediate level (Q2).

Relative importance and ranking of the variable based on the range of predicted response value (lower the range, better the variables in terms of rank

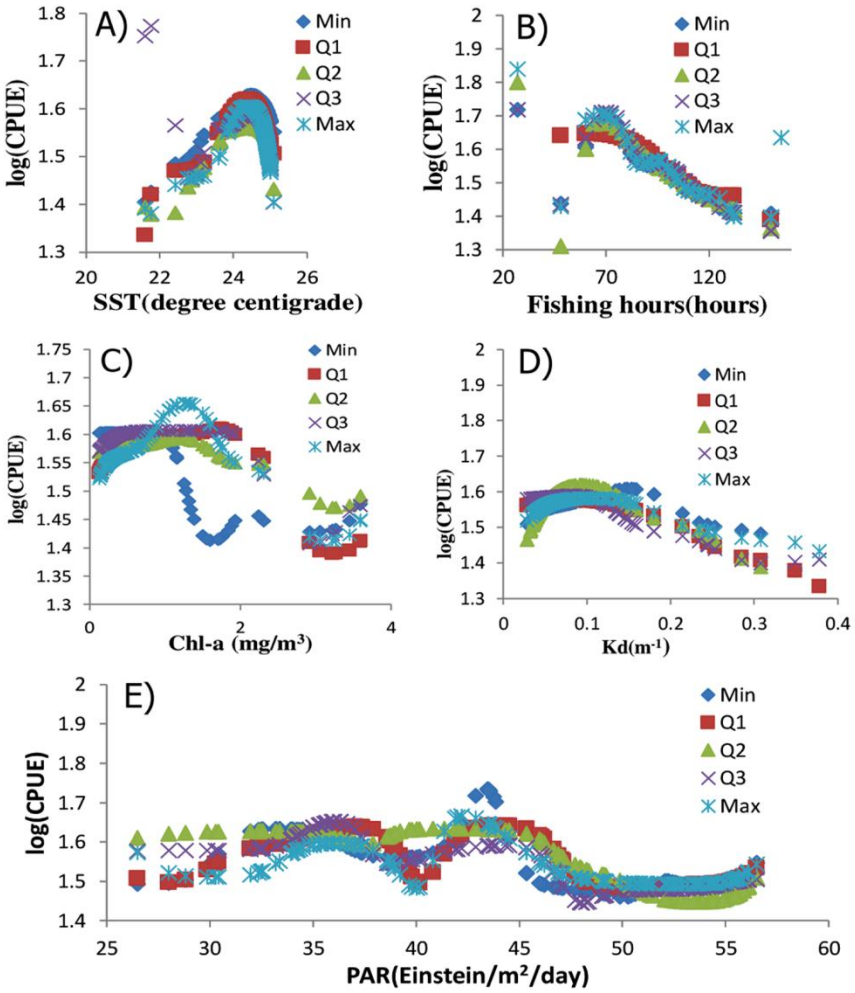

Fig. 4 — A-E: Contribution plots from the sensitivity analysis illustrating the neural network response curve to changes in each variable with all other variables held at a minimum, first quartile, median, the third quartile, and maximum

or importance) showed in Table S4 of supplementary data.

Perturb method

The responses of the output variable in terms of mean square error (MSE) against increased in steps of 10 % noise of the input value up to 50 % (commonly used values) showed in Table S5 of supplementary data. The rank of input variables under every incremental noise shown in brackets

As commonly used noise value is 50 % change in input value[22], the most important variable is chlorophyll-*a*, followed by Kd. The least important variable is SST, followed by the fishing hour.

Connection weight & Garson algorithms

The critical difference between the result of Connection weight approach[22] (Table S6 – S8) and Garson's algorithm[28] (Table S9 & S10) is in their differential ability to identify variable importance in neural networks correctly. The inability of Garson's algorithm to accurately estimate true variable importance could be simply illustrated for input variable Kd which was incorrectly ranked the least important variable in the network (Table S10) which contradicts the statement of the high degree of correlation between Kd and most important variable Chl-*a*[33] (R^2 = 0.983). The connection weight product matrix (Table S7) shows that although input neuron three positively influences the output neuron *via* hidden neurons 1, 6, and 8, it also negatively influences the output neuron via hidden neurons 2, 3, 4, 5, 6 and 7. As Garson's algorithm uses absolute connection weights in its calculations, it fails to account for the determining the influences of input neuron 3 (Kd) through different hidden neurons, resulting in an incorrect estimation of variable importance. In contrast, the Connection Weight approach uses raw connection weights, which account for the direction of the input-hidden-output relationship and results in the correct identification of variable contribution[22].

Forward & Backward stepwise method

The forward and backward stepwise method assesses the change in the mean square error of the output by sequentially adding and removing respectively input neurons to the neural network (rebuilding the neural network at each step). The resulting change in mean square error for each variable addition illustrates the relative importance of the predictor variables[22].

The ranking of the variables under this method showed in Table S11. The similar performances had been seen in forward & backward stepwise approaches with minor changes in rank (1 & 2) of Kd and PAR.

Comparative ranking of variable importance in log (CPUE) prediction – A comparison of methods

Comparative of ranking's relative importance by all methods showed in Figure 5, which clearly indicates the 1st two important variables are Chl-*a* and Kd. The input variables SST and fishing hour were getting the same ranks (4) by two methods and ranked 3 & 5 by the other two methods in said figure. But the profile plot of fishing power showed the decreasing response curve where all other input variables were kept constant at a different level (minimum, first quartile, second quartile, third quartile, and maximum values) (Fig. 4B). Also, the partial derivative values of log (CPUE) with respect to the fishing hour were negative for low values of the fishing hour and near zero for the higher values (Fig. 3A). Combining the profile plot and partial derivative distribution approach (Figs. 4A & B and Figs. 3E & A), it was clear that SST is the less important variable as compared to the fishing hour. The input variable PAR was getting rank 5 (least important) by two methods and rank 3 by the other two methods. But the distribution of profile plot and partial derivative (Figs. 3E & D) could not clearly indicate it, the least important variables (rank 5), and hence it could get rank 3. So we conclude that fishing hour is the least important variable (rank 5), SST, and PAR will be placed at rank 4 and 3, respectively. The catch is not

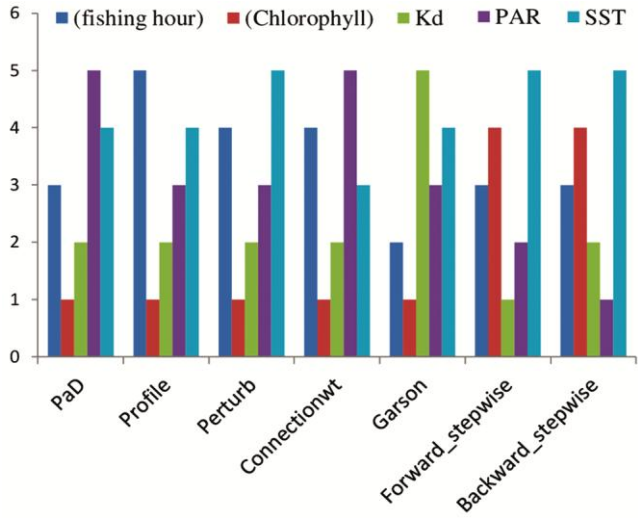

Fig. 5 — Comparative ranking of variables by all methods

increasing proportionally with an increase of fishing hour that results from the decrease of CPUE and hence justifying the fishing hour as 2nd last least important variable. SST & PAR were placed at a very close rank as they have a linear relationship[24].

Conclusion & Summary

Log(CPUE) was modeled in three steps. Initially, we applied a simple general linear model to gain insight into how the independent variables related to the prediction of Log(CPUE) in our dataset. And then different generalized linear model (GLM) using different distributions and link functions were used afterward we shifted to non-parametric generalized additive models (GAMs) and tested whether they are an improvement over the linear approach. GAM was an improvement over the linear approach (Simple General Linear Model), as it explained an additional 11 % of the variance. Our results indicated that Fishing hour, Kd_490 & PAR played the most significant role in the model.

By contrast, Artificial Neural Networks (ANNs), characterized by their ability to model non-linear relationships, is more novel in ecology. To confirm the superiority of the ANN approach on GAM, a ten-fold cross-validation method was implemented. After models had fitted, prediction errors were used to compute the Average prediction accuracy in terms of mean square error (MSE). Results showed that ANN was slightly better as a comparison to the GAM approach.

The explanatory methods in ANN helped in identifying the environmental factors affecting CPUE prediction and also how these factors contribute to CPUE prediction. Several methods used in this study. Our study provides a robust comparison of the performance of six different methodologies for assessing variable contributions in artificial neural networks. The results observed for each method are not always the same. Their different computation leads to different results. Our results showed that the PaD method, Profile method, Input perturbation (50 % noise), and connection weight approaches were only consistent in identifying the two most important variables (Chlorophyll-*a* and Kd) in the network. The orders of the importance of the other three variables-SST, fishing hours, and PAR had not been clearly identified. But based on the profile plot and partial derivative distribution approach of these variables, it was clear that SST is the least important variable (rank 5), Fishing hour and PAR will be placed at rank 4 and 3 respectively.

Similar performances were observed in forward & backward stepwise approaches, but contributions were not sufficiently expressed. The Garson's Algorithm was the poorest performing approach, as the variable contribution has been determined by using absolute connection weights and does not account for counteracting connection weights linking input and output neurons.

To the best of the authors' knowledge, this is the first time that various explanatory methods, which determine the contributions of the independent variables under ANNs, were used to analyze CPUE. This approach seems to be stable since all the methods (except Garson Algorithm and Stepwise method) used to assign a very similar hierarchy of importance to the variables, especially to the first two important variables (Chl-*a* and Kd). These results concur with those from other studies related to the subject, which supports the validity of the results[34]. The reason for the first two important variables is obvious as Chl-*a* is the direct indicator of the food source of fish and Kd gives a fair idea of the transparency of the water column and assumes importance, as distribution or assemblage of some species depends on availability (or sighting) of prey. Also, it was observed from the profile and the derivative plot that most of the fish catch obtained where Chl-*a* and Kd values were less than 2 mg m^{-3} and 0.2 m^{-1}, respectively.

It is our belief that this paper provides a comparison of GLM, GAM, and ANN in terms of sensitivity or importance of independent variables, and it was found that ANN was equally good in dealing with the nonlinear relation with commonly used GAM. Also, GLM performs better as a comparison to the general linear model. In GAM, three significant variables were fishing hour, Kd, and PAR, but there was a very high degree of correlation between Kd & Chl-*a* so that Chl-*a* would be equally important as Kd. Hence four important variables were fishing hour, Chl-*a*, Kd, and PAR. SST was the least important variable, as it was not significant. Similar sensitivity or importance of variables was observed under different explanatory method (PaD method, profile method, input perturbation (50 % noise) & connection weight) under ANN. So we concluded that the above said different explanatory methods under ANN could be used to find the sensitivity or

importance or contribution of variables in place of GAM, which is an established model in ecology for the same.

As far as the limitation of this study is concerned, it is the availability of sufficient data of fish habitat (species wise) with environmental variables. Large availability of data would be served a better purpose to reach a concrete decision on variables importance on fish catch along with model superiority of ANN over GAM and GLM.

Regarding future work, distribution or assemblage of individual fish species will be studied with respect to environmental variables and variables will be ranked (sensitivity analysis) species-wise so that fishers can target the species for catch based on the prevalence of the most important environmental variables and their ranges in the coastal marine area.

Supplementary Data

Supplementary data associated with this article is available in the electronic form at http://nopr.niscair.res.in/jinfo/ijms/IJMS_49(11)1729-1741_SupplData.pdf

Acknowledgments

The authors are grateful to the Director, ICAR-CIFE, Mumbai, and Director, IIT Bombay, for providing the facilities to carry the work. The authors are also thankful to Indian National Centre for Ocean Information Services (INCOIS) Hyderabad, India, for providing the fish advisory validated data.

Conflict of Interest

The authors would like to declare that there are no conflicts of interest to publish this research papers in the journal.

Author Contributions

The authors like to certify that, the first author (VK) of this paper had contributed towards the preparation of the paper such as conceptualization, data collection, data analysis, drafting of manuscript; second author (SJ) contributed in guidance, editing the contents of writing; and third author (JA) contributed in over all supervision, guidance on manuscript revision.

References

1 Yadav V K, Jahageerdar S, Ramasubramanian V, Bharti V S & Adinarayana J, Use of different approaches to model catch per unit effort (CPUE) abundance of fish, *Indian J Geo-Mar Sci*, 45 (12) (2016) 1677-1687

2 Manel S, Dias J M & Ormerod S J, Comparing discriminant analysis, neural networks, and logistic regression for predicting species distributions: a case study with a Himalayan river bird, *Ecol Model*, 120 (1999) 337-347

3 Austin M P & Meyers J A, Current approaches to modeling the environmental niche of eucalyptus: implications for management of forest biodiversity, *Forest Ecol Management*, 85 (1996) 95–106.

4 Guisan A, Edwards T C, Thomas C & Hastie T, Generalized linear and generalized additive models in studies of species distributions: setting the scene, *Ecol Model*, 157 (2002) 89-100.

5 Maunder M N & Punt A, Standardizing catch and effort data: a review of recent approaches, *Fish Res*, 70 (2004) 141–159.

6 Damalas D, Megalofonou P & Apostolopoulou M, Environmental, spatial, temporal, and operational effects on swordfish (*Xiphias gladius*) catch rates of eastern Mediterranean Sea longline fisheries, *Fish Res*, 84 (2007) 233–246.

7 Anuja A & Yadav V K, Use of a different approach in finding catch effort relationship in hook and line fishery in Kombuthurai village of Thoothukudi district of Tamil Nadu, *J Appl Nat Sci*, 10 (2) (2018) 648-654.

8 Mastrorillo S, Lek S, Dauba F & Belaud A, The use of artificial neural networks to predict the presence of small-bodied fish in a river, *Freshw Biol*, 38 (1997) 237–246.

9 Lek S, Delacoste M, Baran P, Dimopoulos I, Lauga J & Aulagnier S, Application of neural networks to modelling nonlinear relationships in ecology, *Ecol Model*, 90 (1996b) 39-52.

10 Yadav V K, Krishnan M, Biradar R S, Kumar N R & Bharti V S, A comparative study of neural-network & fuzzy time series forecasting techniques — Case study: Marine fish production forecasting, *Indian J Geo-Mar Sci*, 42 (6) (2013) 707-716.

11 Paruelo J M & Tomasel F, Prediction of functional characteristics of ecosystems: a comparison of artificial neural networks and regression models, *Ecol Model*, 98 (1997) 173-186.

12 Ramos-Nino M E, Ramirez-Rodriguez C A, Clifford M N & Adams M R, A comparison of quantitative structure-activity relationships for the effect of benzoic and cinnamic acids on Listeria monocytogenes using multiple linear regression, artificial neural network, and fuzzy systems, *J Appl Mirobiol*, 82 (1997) 168-176.

13 Özesmi S L & Özesmi U, Artificial neural network approaches to spatial habitat modeling with interspecific interaction, *Ecol Model*, 116 (1999) 15-31.

14 Xu M, Zeng G M, Xu X Y, Huang G H, Sun W, *et al.*, Application of Bayesian regularized BP neural network model for the analysis of aquatic ecological data - A case study of chlorophyll-*a* prediction in the Nanzui water area of Dongting Lake, *J Environ Sci*, 17 (6) (2005) 946-952.

15 Tirelli T & Pessani D, Use of decision tree and artificial neural network approaches to model presence/absence of *Telestes muticellus* in piedmont (North-Western Italy), *River Res Appl*, 25 (2009) 1001–1012.

16 Bharti V S, Inamdar A B, Purusothaman C S & Yadav V K, Soft Computing and Statistical Technique - Application to Eutrophication Potential Modelling of Mumbai Coastal Area, *Indian J Geo-Mar Sci*, 47 (2) (2018) 365-377.

17 Ripley B D, *Pattern recognition, and neural networks*, (Cambridge University Press), 1996, pp. 416.

18 Gevrey G, Dimopoulos I & Lek S, Review and comparison of methods to study the contribution of variables in artificial neural network models, *Ecol Model*, 160 (2003) 249-264.

19 Zurada J M, Malinowski A & Cloete I, Sensitivity analysis for minimization of input data dimension for feed forward neural network, *ISCAS'94, IEEE International Symposium on Circuits and Systems*, Vol 6, (IEE Press, London), 1994, pp. 447-450.

20 Kim S H, Yoon C & Kim B J, A structural monitoring system based on sensitivity analysis and a neural network, *Comput Aided Civ Inf*, 155 (2000) 309-318.

21 Liong S Y, Lim W H & Paudyal G N, River stage forecasting in Bangladesh: a neural network approach, *J Comput Civ Eng*, 14 (2000b) 1-8.

22 Olden J D & Jackson D A, Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks, *Ecol Model*, 154 (2002) 135–150.

23 Ibrahim O M, A comparison of methods for assessing the relative importance of input variables in artificial neural networks, *J Appl Sci Res*, 9 (11) (2013) 5692-5700.

24 Madhavan N, Thirumalai V D, Ajith J K & Sravani K, Prediction of Mackerel Landings Using MODIS Chlorophyll-*a*, Pathfinder SST, and SeaWiFS PAR, *Indian J Nat Sci*,5 (29) (2015) 4858-4871.

25 Hasti, T, Tibshirani R & Friedman J, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, (Springer-Verlag, New York), 2001, pp. 764.

26 Nelder J A & Wedderburn R W M, Generalized linear models, *J R Statist Soc A*, 137 (1972) 370–384.

27 Geman S, Bienenstock E & Doursat R, Neural networks and the bias/valance dilemma, *Neural Comput*, 4 (1992) 1-58.

28 Garson G D, Interpreting neural network connection weights, *Artif Intell Expt*, 6 (1991) 47-51.

29 Lek S, Belaud A, Dimopoulos I, Lauga J & Moreau J, Improved estimation, using neural networks, of the food consumption of fish populations, *Mar Freshw Res*, 46 (1995) 1229-1236.

30 Maier H R, Dandy G C & Burch M D, Use of artificial neural networks for modelling Cyanobacteria Anabaena spp. in the river Murray, South Australia, *Ecol Model*, 105 (1998) 257-272.

31 Dick E J, Beyond 'lognormal vs. gamma': discrimination among error distributions for generalized linear models, *Fish Res*, 70 (2004) 347–362.

32 Usman T, Ersti Y S & Syaifuddin A, The relationship between the concentration of chlorophyll-*a* with skipjack (*Katsuwonus pelamis*, Linnaeus 1758) production at West Sumatera waters, Indonesia, *IOP Conf. Series: Earth Environ Sci*, 54 (2017), pp. 012072.

33 Morel A, Huot Y, Gentili B, Werdell P J, Hooker S B, *et al.*, Examining the consistency of products derived from various ocean color sensors in open ocean (Case 1) waters in the perspective of a multi-sensor approach, *Remote Sens Environt*, 111 (2007) 69–88.

34 Juan D O & Concepción G, Extracting the contribution of independent variables in neural network models: a new approach to handle instability, *Neural Comput App*, 25 (2014) 859–869.