Natl. Acad. Sci. Lett. (May–June 2022) 45(3):245–248
https://doi.org/10.1007/s40009-022-01110-0

SHORT COMMUNICATION

Application of Machine Learning to Investigate the Impact of Climatic Variables on Marine Fish Landings

Labonnah Farzana Rahman[1] · Mohammad Marufuzzaman[2] · Lubna Alam[1] · Md Azizul Bari[1] · Ussif Rashid Sumaila[1,3] · Lariyah Mohd Sidek[2]

Received: 15 June 2021 / Revised: 27 December 2021 / Accepted: 10 January 2022 / Published online: 13 March 2022

Abstract The fisheries industry of Malaysia is known as the strategic sector that can help the country raise domestic food production and supply. This research proposed machine learning (ML) based prediction of marine fish landings to project fish supply and compare those projections with the observed data. Three ML models, i.e., linear regression (LR), decision tree (DT), and random forest (RF) regression, are applied to the dataset that contains 18 years of climatic variables and the marine fish landings (tonnes) information of 5 major states of Malaysia. The results suggest that the developed LR model shows an R^2 value of 0.60 and 0.64 in the validation and testing phases. The DT and RF model indicates a significant improvement as the R^2 values are 0.88 and 0.89 in the validation data and 0.89 and 0.86 in the testing data. Finally, we calculated the Nash–Sutcliffe efficiency (NSE) values, and the results indicated that RF based ML model has the highest NSE value of 0.86, which turns out to be the best fit for prediction. The developed ML models have utilized for the first time to predict the marine fish landing using environmental inputs collected from 5 different states of Malaysia.

Keywords Climatic variables · Marine fish landings · Decision tree · Random forest · Machine learning

Malaysia is highly vulnerable to the effects of climate change [1]. The selection of climatic variable indicators for regional analysis was fraught with constraints, assumptions, and availability of datasets. The previous studies demonstrated rainfall and temperature impacts on fish landings in the focus country [2, 3]. Similarly, sea surface temperature (SST) was an essential indicator for coastal upwelling events influencing fish production reported for the region [4]. Prior studies demonstrated that relative humidity was a significant climatic factor in fisheries studies because of its indirect impact on some environmental stressors [5, 6]. Therefore, the use of climatic variables such as rainfall and SST on Malaysia's marine fish landings must be investigated. Forecasting marine fish landings is highly dependent on the analysis of previous and current behaviors [6]. Autoregressive integrated moving averages (ARIMAs), seasonal ARIMAs, vector autoregression, neural networks, nonlinear autoregressive networks, and wavelets are a few well-known approaches that researchers have used to forecast short-term fish catches [6, 7]. However, these statistical models will not produce satisfactory results if the time series data have nonlinear components [8]. Machine learning (ML) models use only historical data to learn the stochastic dependency between the past and the future [9, 10]. Previous studies have used ecological variables in Malaysia to estimate marine fish catches [11, 12]. However, none of them

Significance Statement We presented the prediction of marine fish landings (tonnes) for five major states using machine learning (ML) method. The impact of climatic variables is investigated using 3 ML models and based on two error objective fuinctions the random forest model is suggested to be the best fit (NSE 0.86) for prediction.

✉ Lubna Alam
lubna@ukm.edu.my

[1] Institute for Environment and Development (LESTARI), Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

[2] Institute of Energy Infrastructure, Universiti Tenaga Nasional, 43000 Kajang, Selangor, Malaysia

[3] Institute for the Oceans and Fisheries, Faculty of Science, The University of British Columbia, Vancouver, Canada

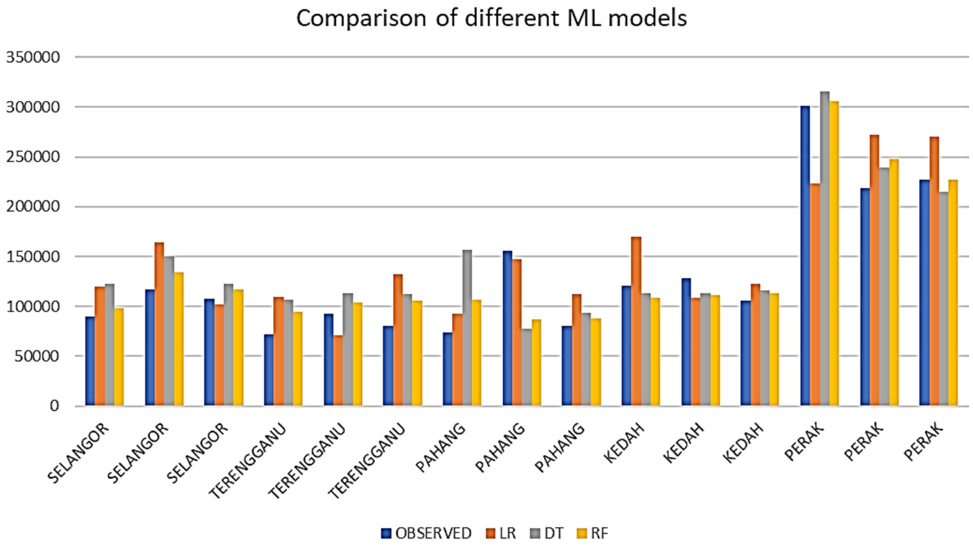

Fig. 1 Three ML-based outputs for marine fish landings prediction using the validation dataset

implemented ML models. Researchers typically use the ML-based linear regression (LR) technique for the prediction of time-series data. This modeling approach is excellent if we have a correlated dataset because the algorithm can accurately predict values. However, algorithms such as the decision tree (DT)-based regression technique can handle data with different measurement scales. DT-based algorithms do not influence outliers and missing values to a fair degree and simplify the building of rules for predictions about individual cases and complex relationships [13]. Moreover, the random forest (RF) algorithm can be used instead of a single DT to reduce overfitting, resulting in better results than with a single optimized DT [14].

In this research, we considered different ML-based predictive models to demonstrate the impact of climatic variables on marine fish landings in Malaysia's five central states: Kedah, Pahang, Perak, Selangor, and Terengganu. Two error objective functions, the coefficient of determination (R^2) and the Nash–Sutcliffe efficiency (NSE), were used to determine the performance of the ML model.

We considered the maximum and minimum air temperature, SST, and humidity to build models using ML. We collected data from 18 consecutive years (2000–2017); we obtained the temperature, rainfall, and humidity data from the Department of Statistics Malaysia and the SST data from the Malaysian Meteorological Department. Marine fish landing data were collected from the Department of Fisheries, Malaysia. For the interpretation of the ML model, individual states were combined into one dataset by mapping the states to numbers, where Selangor is 1, Terengganu is 2, Pahang is 3, Kedah is 4, and Perak is 5. We used the first 16 years (2000–2015) of data to validate the training of the model and the latest 2 years (2016–2017) of data to test the ML models. We used 65 random data points for training and 15 random data points for validation, and the data points were fragmented by the *stratifying* method so that all of the states exist in both datasets. We implemented the LR, DT and RF algorithms to generate predictions. For the DT and RF algorithms, the maximum depth was set to 7 to reduce data overfitting [15]. We used Python scikit-learn to implement the model and measured the R^2 and NSE values to determine the predictive accuracy [16]. Both of these error objective functions expressed values between 0 and 1, and a value closer to 1 indicated a more accurate prediction. The NSE was the best objective function for evaluating the overall fit between the predictive and observed values [17]. Figure 1 shows the graph comparing 15 data points after implementing the 3 ML algorithms. Three values (years) from each state were plotted on the x-axis, and the observed values and those predicted by the 3 ML models were plotted on the y-axis. The RF and DT-based ML regression models produced values closer to the observed values, and they had better R^2 (0.88 and 0.89, respectively) and NSE (0.7 and 0.8, respectively) values than the LR model (R^2 = 0.6 and NSE = 0.3).

Table 1 shows the predicted and observed results for the test dataset (2016–2017) as well as the error matrices. We found that the RF model output most closely resembles the observed dataset. Table 1 indicates that the LR model has a high bias, whereas DT and RF have comparatively improved prediction results with low bias. The results of the analysis of the 2017 data showed that LR resulted in negative values, indicating that the LR model has low predictive accuracy (R^2 = 0.64 and NSE = 0.082). We also found that in 2016 and 2017, the DT model predicted the same values for different states, which is one of the drawbacks of employing a single DT (R^2 = 0.89 and NSE = 0.84). Similar or identical inputs yielded a

Table 1 Observed and predicted fish landing (tonnes) values and the corresponding error

Year	State	Actual	LR	DT	RF
2016	Selangor	93,460	128,676	123,024	100,175
2016	Terengganu	44,703	120,131	68,313	89,028
2016	Pahang	163,878	99,206	157,101	105,018
2016	Kedah	120,418	186,165	113,674	112,234
2016	Perak	367,347	267,373	346,853	327,642
2017	Selangor	125,517	− 56,797	105,560	98,035
2017	Terengganu	47,269	− 86,850	54,985	76,642
2017	Pahang	142,550	165,023	113,674	112,033
2017	Kedah	118,066	172,372	112,086	116,345
2017	Perak	266,556	299,393	346,853	310,549
Error matrices		MAE	34,346	28,054	18,163
		R^2	0.640	0.890	0.860
		NSE	0.082	0.840	0.860

particular predicted value. Therefore, the RF model was used to average multiple DTs to improve the accuracy and reduce data overfitting. The R^2 and NSE values of the RF model were 0.86 and 0.86, respectively, which were better than those of the other ML models with the testing dataset. Thus, according to this research, the RF regression model is suitable for predicting marine fish landings (tonnes) in the abovementioned Malaysian states.

Here, the NSE value for the RF model was 0.86, indicating a good fit [18]. The dataset contained all five major states in both the validation and testing phases. Thus, this research successfully predicted marine fish landings in five central states of Malaysia. Decision-makers in the fishery industry typically plan based on the fishing market's resource requirements, which are highly dependent on accurate 1- to 2-year forecasts of fish landings [19]. Therefore, this predictive model can be a valuable component included in the construction of decision support systems for Malaysia's fisheries sector.

Acknowledgements The authors would like to express sincere gratitude to the Universiti Tenaga Nasional and Institute for Environment and Development, Universiti Kebangsaan Malaysia for all the support to conduct this research work.

Funding Funding was provided by universiti tenaga nasional (BOLDRefresh2025) and Kementerian Pendidikan (TRGS/1/2015/UKM/02/5/2).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Sumaila UR (2019) Comparative valuation of fisheries in Asian Large Marine Ecosystems with emphasis on the East China Sea and South China Sea LMEs. Deep Sea Res Part II Top Stud Ocean 163:96–101
2. Kathijotes N, Alam L, Kontou A (2015) Aquaculture, coastal pollution and the environment, *Aquacu Ecosyst Adapt Sustain* pp 139–163
3. Ho DJ, Maryam DS, Jafar-Sidik M, Aung T (2013) Influence of weather condition on pelagic fish landings in Kota Kinabalu, Sabah, Malaysia. J of Trop Bio Con (JTBC) 10:11–21
4. Subarna D (2018) The effect of monsoon variability on fish landing in the Sadeng Fishing Port of Yogyakarta, Indonesia. IOP Conf Ser Earth Environ Sci 139(1):012027
5. Atindana SA, Ofori-Danson PK, Brucet S (2019) Modelling the effects of climate change on shellfish production in marine artisanal fisheries of Ghana. AAS Open Res 2:16
6. Yadav VK, Jahageerdar S, Adinarayana J (2020) Modelling framework to study the influence of environmental variables for forecasting the quarterly landing of total fish catch and catch of small major pelagic fish of north-west Maharashtra Coast of India, *Natl Acad Sci Lett* pp 1–4
7. Anuja A, Yadav VK, Bharti VS, Kumar NR (2017) Trends in marine fish production in Tamil Nadu using regression and autoregressive integrated moving average (ARIMA) model. J Appl Nat Sci 9(2):653–657
8. Majid R, Mir SA (2018) Advances in statistical forecasting methods: An overview. Eco Aff 63(4):815–831
9. Rahman LF, Marufuzzaman M, Alam L, Bari MA, Sumaila UR, Sidek LM (2021) Developing an ensembled machine learning prediction model for marine fish and aquaculture production. Sustainability 13(16):9124
10. Marufuzzaman M, Bin Ibne Reaz M, Rahman LF, Farayez A (2017) A location based sequence prediction algorithm for determining next activity in smart home. *J Eng Sci Technol Rev* 10(2)
11. Knudby A, LeDrew E, Brenning A (2010) Predictive mapping of reef fish species richness, diversity and biomass in Zanzibar using IKONOS imagery and machine-learning techniques. Remote Sens Environ 114(6):1230–1241

12. Alam L, Mokhtar M, Ta GC, Halim SA, Ahmed MF (2017) Review on regional impact of climate change on fisheries sector. Nov J 4(1):1–5
13. Tehrany MS, Pradhan B, Jebur MN (2013) Spatial prediction of flood susceptible areas using rule-based decision tree (D.T.) and a novel ensemble bivariate and multivariate statistical models in GIS. J Hydrol 504:69–79
14. Pal M, Mather PM (2003) An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sens Environ 86(4):554–565
15. Brame M (2007) Avoiding overfitting of decision trees, Principles of data mining, pp 119–134
16. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Duchesnay E (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
17. Ahmed M, Sultan M, Elbayoumi T, Tissot P (2019) Forecasting GRACE data over the African watersheds using artificial neural networks. Remote Sens 11(15):1769
18. Moriasi DN, Arnold JG, VanLiew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans ASABE 50(3):885–900
19. Felthoven RG, Paul CJM (2004) Directions for productivity measurement in fisheries. Mar Policy 28(2):161–169