Int. J. Advance Soft Compu. Appl, Vol. 9, No. 2, July 2017
ISSN 2074-8523

Forecasting of Monthly Marine Fish Landings using Artificial Neural Network

Najah Nasir [1], Ruhaidah Samsudin [1], and Ani Shabri [2]

[1] Fakulti Komputeran,
Universiti Teknologi Malaysia, 81310, Johor, Malaysia
[2] Fakulti Sains, Universiti Teknologi Malaysia, 81310, Johor, Malaysia
e-mail: najah6@live.utm.my, ruhaidah@utm.my,
ani@utm.my

Abstract

Management of marine resources have gradually become more important during these past years because of the increased awareness of these resources becoming limited. Forecasting of fish landings is one of the many ways that can contribute to a better decision making for fisheries management. Being a renowned forecasting model, artificial neural network with back propagation was selected for this research with enhancement made by pre-processing the data using empirical mode decomposition. The monthly marine landings data of East Johor and Pahang which has 144 observations each, was gathered from the Department of Fisheries Malaysia website. A ratio of 92:8 was used to divide the data into training and testing sets. Data pre-processing was done in R software whereas the forecasting models were developed in MATLAB software. Results from the proposed model are then compared to a conventional artificial neural network using the root-mean-square error and mean absolute error values, wherein it was shown that the proposed model could outperform the conventional model.

Keywords: *Artificial neural network, Empirical mode decomposition, Fish landing, Forecast.*

1 Introduction

Fish landing is the quantity of fish brought to shore after being collected from the sea. Different from landings, catch includes discards but most often, the landings record is also the total catch. Resources of the sea seemed unlimited in past years when the sea was used as a means of transport but nowadays, discussions on the sustainability of marine reserves can turn into arguments about privileges, responsibilities, and necessities of the public and private communities. Due to the rising demand for food, energy, and manufactured goods by the growing human population, natural resources are constantly under pressure. The limitations and potential problems when dealing with land are normally easily seen but since the ocean reserves cannot be clearly observed or calculated, it became more intricate and provocative.

Previous researchers have implemented different models in the area of fish landing forecasting as listed in the following Table 1. Particularly in Malaysia, the study on fish landing forecasting appears to be an uncommon practice as only the research by [1], including [2] is found. Note that this field is also infrequently researched by scientists around the globe as there are only a handful of studies conducted in the past decade. Despite this situation, it is important for more research related to forecasting of fish landing be performed as the results could help in better management of marine resources.

Table 1: Models employed for fish landing forecasting

Year	Author	Model
2015	Shabri and Samsudin	Discrete wavelet transform-ARIMA
2014	Bako	Box-Jenkins, Error Trend and Seasonal state space exponential
2006	Koutroumanidis et al.	Regression, exponential smoothing, auto-regressive integrated moving average (ARIMA), harmonic regression, dynamic regression, vector auto-regression
2003	Kaunda-Arara et al.	ARIMA
2001	Craine et al.	Nonlinear regression, ARIMA transfer function
1998	Park	ARIMA
1998	Venugopalan and Srinath	Regression, exponential smoothing, ARIMA, harmonic regression
1993	Houde and Rutherford	Nixon’s equation [15]

Research in artificial intelligence, especially in the fields of computational learning theory and pattern recognition, has allowed for the development of machine learning as a subtopic of computer science. Machine learning can be explained as a study discipline to provide computers with a general automation of data handling and learning ability [9]. Instead of exactly adhering to fixed program commands, machine learning algorithms make data-driven decisions by the construction of a model from sample inputs. Approaches of machine learning include among others, similarity and metric learning, deep learning, clustering, Bayesian networks, artificial neural networks (ANN) and sparse dictionary learning.

Inspired by biological neural networks, ANN is a collection of models capable of approximating functions, even for those with inputs that have large numbers.

ANN is input adaptive and possesses learning abilities because the system of interrelated neurons with which they are commonly portrayed, exchange messages with each other and have adjustable numeric weights. A neural network computational model was created by [10] based on algorithms and mathematics named as the threshold logic. Resultant to this, two different fields for ANN research was undertaken, in which one focuses on the brain's biological processes while the other is on ANN utilisation in artificial intelligence.

Applicable to machine learning and data mining projects, the phrase "garbage in, garbage out", which is synonymous with data pre-processing indicates that it is imperative for such projects. Because control of data-gathering methods is unrestricted, values could be out-of-range, are missing or have impossible combinations. Results would be deceiving if data was not cautiously assessed for such problems, therefore data representation and quality are predominant before any analysis [11]. Knowledge discovery during the training phase would be extra challenging if irrelevant, redundant, noisy or unreliable data are present, besides requiring significant processing time since data needs to be prepared and filtered. Methods for data pre-processing include cleaning, feature extraction, transformation, normalisation and feature selection, where the end product is the final training set. An algorithm for every data pre-processing activity is presented by [12].

2 Methodology

Trend analysis and generating predictions of the future based on past and present data are parts of the forecasting process, also generally known as prediction. The process might be indicated by judgmental or statistical methods wherein time series, cross-sectional or longitudinal data is utilised. Various categories of forecasting processes exist including time series and artificial intelligence methods. Methods useful for time series data are moving average, Kalman filtering and Box-Jenkins methods, to name a few. Artificial neural networks, support vector machines and group method of data handling are some examples of artificial intelligence methods.

2.1 Artificial neural network

Before a paper by [14], the importance of backpropagation algorithm was not fully appreciated wherein it was originally introduced in the 1970s. In the paper, several neural networks with backpropagation were described to work faster than other learning approaches, rendering it viable for utilisation on previously insoluble problems. Today, the algorithm has become the learning backbone of neural networks. The backpropagation algorithm is expressed by partial derivatives of the cost function with respect to weights or biases in the network. Backpropagation explains the cost function's rate of change when the weights and biases were adjusted, resulting in different behaviours of the network. Propagation and weight update are main stages of the backpropagation learning algorithm, as detailed in the following;

Stage 1 *Propagation*

1. Forwardly propagate inputs through the network to produce the output.
2. Backwardly propagate the output through the network to generate the delta, i.e. difference between the target and actual output.

Stage 2 *Weight update*

1. Multiply the delta and input to get the gradient of weight.
2. Subtract a percentage of the gradient from the weight.

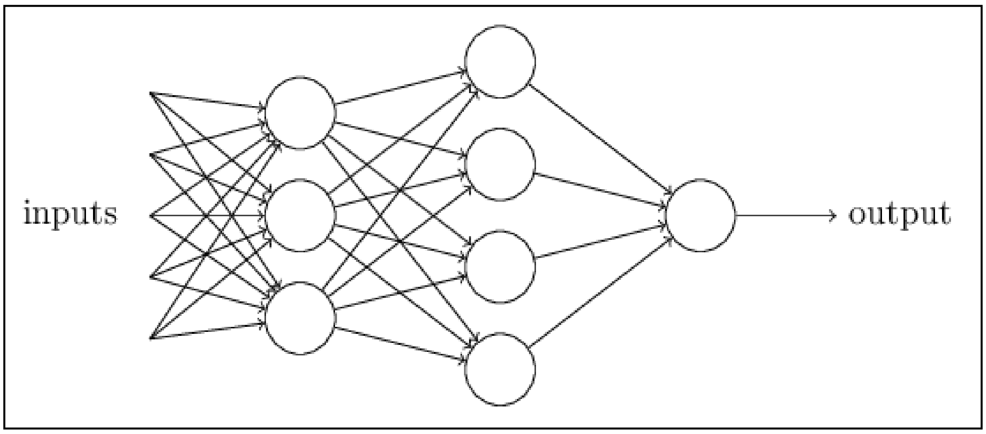


Fig. 1: Example of a network with three layers

2.2 Empirical mode decomposition

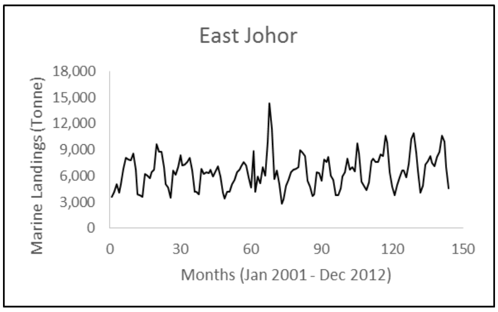
The empirical mode decomposition (EMD), proposed by [13] is a signal decomposition method. Several intrinsic mode functions (IMF) and a trend are products of the decomposition process where they can be seen as expansions of the data [13]. This technique is suitable for use on nonstationary and nonlinear time series data. EMD resembles an algorithm or empirical method instead of a theoretical tool that is applicable to a dataset, different from other transforms such as Fourier's. [13] reasoned that IMF is named so because there is only a single oscillation mode in every cycle of zero crossings. The definition given for IMF is a function satisfying conditions of having a zero average value for the envelope of local maxima and minima, including having an equal or a difference of at most one for the number of extrema and zero crossings [13]. [13] devised a structured procedure for extracting the IMF from data, named as the sifting process, which is explained in the following Fig. 2.

1. Identify extrema.
2. Join local maxima via cubic spline line for the upper envelope and local minima for the lower envelope.
3. Calculate component h_1, which is the difference between data and mean of envelopes m_1, that is,
$$h_1 = X(t) - m_1 \qquad (1)$$
4. Repeat Step 1 and 2 until h_k is an IMF, i.e.
$$h_2 = h_1 - m_2,\ \dots, h_k = h_{k-1} - m_k \qquad (2)$$
where h_{k-1} replaces $X(t)$ of equation (1) for subsequent siftings.
5. Assign $h_k = c_k$ as k-th IMF data component.
6. Compute residue by
$$r_i = X(t) - c_i \qquad (3)$$
7. Repeat Step 6 for all r_j, as in
$$r_2 = r_1 - c_2,\ \dots,\ r_n = r_{n-1} - c_n \qquad (4)$$
with $X(t)$ of equation (3) substituted for r_{i-1}.
8. Check stopping criteria, whether
   a. c_n or r_n is less than the substantial consequence value
   b. no IMF can be extracted due to r_n becoming a monotonic function.
9. Obtain n-empirical modes and a residue by adding equations (3) and (4) to get
$$X(t) = \sum_{i=1}^{n} c_i + r_n \qquad (5)$$

Fig. 2: Sifting process of EMD

3 Data

Data used for this research were taken from the List of Annual Fisheries Statistics which is available for download at the Department of Fisheries Malaysia official portal (http://www.dof.gov.my/en/fishery-statistics). The statistics relevant to this research are Landings of Marine Fish by Month and State from the year 2001 to 2012, particularly those of Pahang and East Johor, which are plotted in Fig. 3. There are 144 observations in the set for each state, where it is decided that a ratio of 92:8 would be used for division into training and testing sets. Such decision was for the reason that data from 2001 to 2011 with 132 observations (11-year monthly data) constitutes around 91.67 percent, while the remaining 12 observations (one-year monthly data) from 2012 is about 8.33 percent. Based on the following Fig. 3, it can be seen that the data pattern is quite irregular but there seemed to be a similar general trend in the data per each year cycle. Apparently, the trend is upwards in the early months until around midyear, then goes downward towards the end of the year.



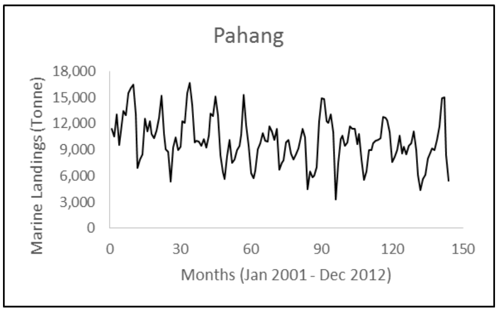


Fig. 3: Marine landings per month for East Johor and Pahang

4 Results

Experiments were performed by combining ANN with EMD and the results obtained were compared to a general ANN model. The ANN model was developed using built-in tools of the MATLAB software. Although the function to produce an ANN network is embedded with data normalisation code, explicit normalisation of data was done before feeding it into the function. Iterations were implemented to get average values of the results where the ANN model is built, trained and simulated at each loop. For this research, the model performance is measured using root-mean-square error (RMSE) and mean absolute error (MAE).

Theorem 4.1 *The RMSE and MAE are given by the following equation (1) and (2) respectively, where* n *is the number of data,* a *is the actual output, while* p *is the predicted output.*

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - p_i)^2} \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i| \tag{2}$$

The RMSE is a regularly utilised measure representing the standard deviation of differences between the values predicted by a model and values observed. The purpose of RMSE is to sum the prediction errors at various time points into a single measure. Although a reliable test for accuracy, RMSE can only compare the forecasting error of numerous models for an individual variable as it is scale-dependent [16]. As its name suggests, the mean absolute error is an average of absolute errors, where it is used in statistics to calculate the proximity of forecasts or predictions to eventual outcomes. Even though a regular measure of prediction error in time series analysis [16], MAE is sometimes alternatively used with the term "mean absolute deviation", perhaps in confusion with the standard definition of both terms. [17] identified MAE as a scale-dependent accuracy measure and cannot be utilised for comparison of series with different scales because it is on the same scale as data concerned.

The monthly marine landings data for East Johor and Pahang is lagged for 12 columns (x_{t-1}, x_{t-2}, …, x_{t-12}) which resulted in only 131 observations per column. Following the ratio of 92:8, the first 120 data would be used for training while the remaining 11 is for testing. Of all the 12 columns created by lagging the original data, only those deemed important enough for the use of forecasting are kept. This was achieved by utilising the partial auto-correlation function (PACF) package for R software environment. The original data for each state is simply sent to PACF as parameters which allow the function to generate plots that show which data lags are relevant as inputs to the forecasting model. The plots produced by PACF are displayed by Fig. 4. From the plots, it can be seen that only some of the lags exceed the threshold value, indicated by the dotted line, meaning that these lags are important input values. Due to having only 12 columns of lagged data, lags above x_{t-12} are ignored. Significant lags for East Johor are x_{t-1}, x_{t-2}, x_{t-3}, x_{t-5}, x_{t-8}, x_{t-10}, x_{t-12} while for Pahang are x_{t-1}, x_{t-2}, x_{t-10}, x_{t-11}, x_{t-12}.

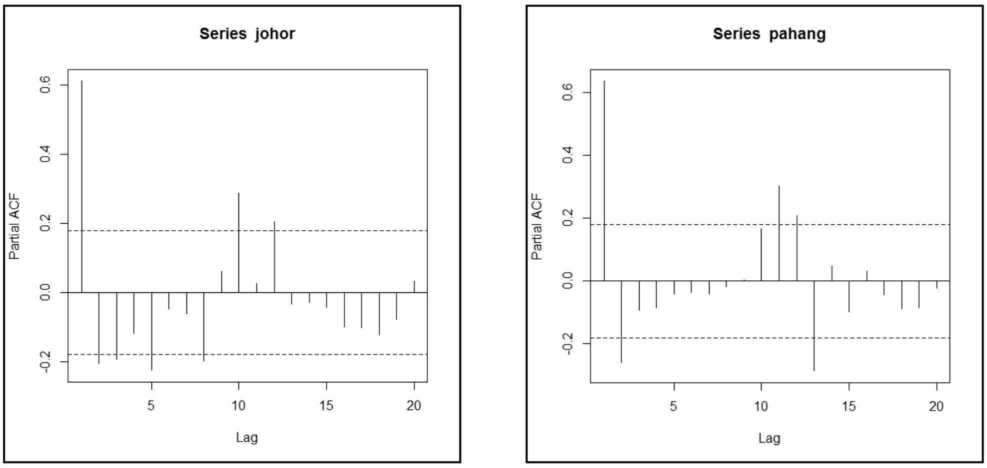


Fig. 4: PACF plot for East Johor and Pahang showing significant lags

The selected lag columns are rearranged side by side with the unwanted columns deleted. This dataset is then supplied to the ANN model programmed in MATLAB. For simplicity purposes, the ANN code is executed only once but with the application of basic ensemble method. Default values are used for all functions of creating the ANN model, where the number of hidden layer is 13, transfer function per layer is hyperbolic tangent sigmoid and linear, whereas the training function is Levenberg-Marquardt backpropagation. Performance calculation of the ANN model is by comparing the predicted output to actual output.

The EMD-ANN model is the joining of EMD to ANN such that EMD is for processing the data further before it is used by ANN for forecasting. Workflow of the implementation of this model is the same as ANN like what was explained previously but with an additional step of applying EMD. EMD is executed using the ‘EMD’ package of the R software environment. The EMD package requires a data array as its parameter so the original data with 144 observations for each state is sent to it. Fig. 5 shows the decomposition of data by the package into several IMF including the residue or trend for East Johor and Pahang. Steps detailed formerly are then repeated for each of the IMF and also the residue, i.e. each IMF is lagged, PACF executed on the lagged IMF, the IMF rearranged for important lags and finally forecasted using an ANN model. Results returned by ANN per each IMF and the residue are then summed before being utilised for the performance evaluation of ANN.

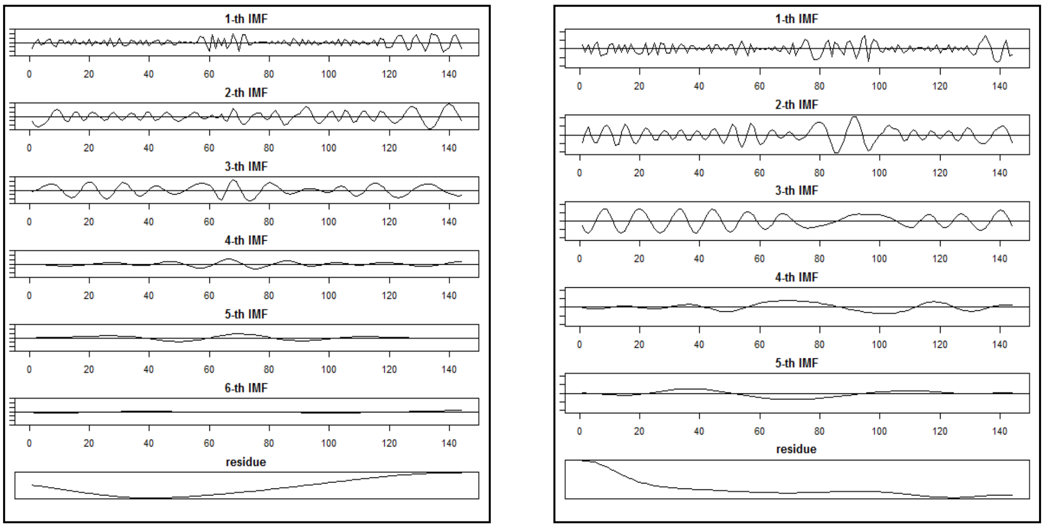


Fig. 5: IMF and residue for East Johor (left) and Pahang (right)

The performance of ANN and EMD-ANN models are measured for comparison to each other. The RMSE for ANN is 1246.542 while EMD-ANN is 1074.213 with the usage of East Johor data. MAE computed for models forecasting East Johor data are 1057.025 and 905.9176 for ANN and EMD-ANN respectively. Applying Pahang data, ANN gives an RMSE value of 1551.45 whereas EMD-ANN yielded 1421.657. Values of 1247.108 and 1180.667 were found for the computation of MAE on ANN and EMD-ANN models after forecasting for Pahang. The values calculated for each model can be viewed in Table 2.

Table 2: Performance measure of each model

	East Johor		Pahang	
	RMSE	MAE	RMSE	MAE
ANN	1246.542	1057.025	1551.450	1247.108
EMD-ANN	1074.213	905.918	1421.657	1180.667

5 Discussion

Based on Table 2, the lowest values for RMSE is by the EMD-ANN model for both East Johor and Pahang data. The situation is also the same for MAE where EMD-ANN produces lower values compared to ANN for each dataset. In general, the ANN model resulted in higher RMSE and MAE contrasting to models wherein data is pre-processed. Since data pre-processing is considered to be capable of assisting a forecasting model in generating better results, the trend in differences between the RMSE and MAE values for both models is as anticipated. More accountable or greater quality predictions are attainable by sufficiently pre-processing the data [18]. Although identification of a suitable data pre-processing method is mostly through trial and error, [19] and [20] have demonstrated that applying de-seasoning and de-trending contributed to improvements, apart from [21] recommending against using complicated techniques.

The random behaviour appearing in complex processes which are usually visible at various time intervals with different frequencies may be overlooked by Fourier or wavelet decompositions but can be easily distinguished by an EMD [22]. [23] stated that EMD is adaptive compared to Fourier analysis because its basis functions are obtained entirely from the signal and not by summation of fixed sinusoids. Nevertheless, EMD has limitations with the cubic spline as its basis function considering that the smoothness in phase function derivatives could not be ensured except for the continuity in data [24]. The sinusoidal form of IMF components can be employed by forecasting models, thus the research with EMD as a data pre-processor is widespread for example by [25], [26], [27], [28], [29], and [30], to list a few.

Notice from the results that the value of RMSE and MAE tend to be lower for East Johor than for Pahang, irrespective of the forecasting models. For instance, the ANN model’s performance for East Johor is better when measured against its corresponding performance for Pahang. This condition possibly resulted from the different time series patterns of each state [18]. Referring to Fig. 3, it is evident that the yearly trend for Pahang is slightly irregular as opposed to the somewhat regular trend for East Johor. A very sharp peak occurred in 2006 for East Johor specifically around June to August, before the rapid plunge to the lowest value for the 2001-2012 record about the period of September 2006 until January 2007. [31] reported that near the end of 2006 up to early 2007, southern Malaysia was struck by a sequence of floods caused by the rainfall exceeding average quantities, credited to Typhoon Utor. Following that, officials were compelled to evacuate more than 50,000 residents in Johor [32], making it likely that the drop in fish landings during these times resulted from the decreasing amount of fishery activities.

The dataset used in this research looks quite simple if compared to datasets with a large volume or has many variables but the need for data lagging during pre-processing could make the final input set have thousands of data points. In an article by [34], complex datasets are fundamentally “big” and come from distinct sources. The difficulty to separate between signal and noise, as well as the computational resources required are problems arising from processing large quantities of data [34]. According to [33], being derived from various processes and having numerous purposes are attributes of complex datasets. Concerning the flooding incident and research data likely having a relation, it could be perceived that the dataset values obtained were produced from an assortment of processes rather than a single discrete action [33].

The correlation between fish landings and months were not investigated in this research. It was noted however that certain relationships which could influence the amount of fish caught might exist between the landings and other variables, such in the case of this research, the 2006-2007 flooding. [35] and [36] claimed that meteorological elements can impact the variety and quantity of fishes as implied by the correlation between fish landings and values of temperature, wind, rainfall, including the Southern Oscillation Index (SOI). They noticed in their study that fish landings dwindled throughout the North-East monsoon in Sabah and low index values with regards to the El Nino Southern Oscillation suggested elevated fish landing amounts [35, 36]. Motivated by these findings, a possible relationship between landings and months was spotted, probably triggered by the parameters stated before. It was discovered that for the 2001-2012 landings record of East Johor and Pahang, its quantities per four-month cycle in each year would culminate in the same months.

Table 3: Highest landings in every one-third of a year for East Johor

	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12
T_1												
Jan	3.6	3.7	4.7	4.2	3.4	8.8	2.8	4.8	3.8	4.4	3.8	4.9
Feb	4.2	3.6	3.5	3.9	4.2	4.2	3.3	3.7	4.6	5.3	5.0	7.3
Mar	5.0	6.2	6.7	6.8	4.2	5.9	4.8	3.9	5.9	7.7	5.9	7.7
Apr	4.1	6.0	6.2	6.2	5.0	5.1	5.5	6.5	6.4	7.9	6.6	8.3
T_2												
May	5.1	5.7	7.0	6.5	5.5	7.0	6.5	6.3	8.0	7.6	6.6	7.5
Jun	6.8	6.4	8.4	6.3	6.4	6.0	6.7	5.5	6.7	7.6	5.8	7.1
Jul	8.1	6.7	7.2	6.7	6.7	10.0	6.9	7.9	7.0	8.5	7.4	8.2
Aug	7.8	9.6	7.3	6.0	7.1	14.3	7.0	7.6	6.5	8.3	10.2	8.7
T_3												
Sep	7.8	8.7	7.6	6.5	7.6	11.4	8.9	8.2	9.7	10.6	10.9	10.6
Oct	8.6	8.8	8.1	7.1	7.2	5.6	8.7	6.0	8.5	9.8	8.7	9.9
Nov	6.7	7.0	6.6	6.0	5.9	6.6	8.2	5.5	5.3	6.4	6.6	7.0
Dec	3.9	5.1	4.2	4.1	4.6	5.0	5.4	3.8	4.8	4.8	4.1	4.6

The original data was converted to decimal values by division over a thousand and rounded off to the tenths for simpler tabulation of data. The months in a year were split into cycles of four-months interval each, denoted as T_1 for the January-April section, T_2 for May-August and T_3 for September-December. The maximum value for each T per year in the 2001-2012 landings record is then highlighted as illustrated in Table 3 and 4, correspondingly for East Johor and Pahang, after which the frequency of a highlighted month is counted. The highlighted months were detected having a tendency to be similar for both states across the years such that the first and second most highlighted month in T_1 is March and April, in T_2 is July and August while in T_3 is September and October, as identified by the months with a dark background in Table 3 and Table 4. This may be an indication that some external factors could influence the amount of fish caught as validated in the study by some researchers [35, 36, 37, 38, 39, 40] but the exact factors in the case of this research are unknown.

Table 4: Highest landings in every one-third of a year for Pahang

	'01	'02	'03	'04	'05	'06	'07	'08	'09	'10	'11	'12
T_1												
Jan	11.4	7.9	8.8	10.0	5.6	5.7	7.5	6.5	7.4	6.5	8.2	5.6
Feb	10.5	8.5	5.3	9.9	8.3	6.6	7.8	5.9	10.3	8.9	9.0	6.1
Mar	13.1	12.6	9.2	9.4	10.1	9.1	9.8	6.0	10.6	8.9	10.6	8.0
Apr	9.5	11.1	10.4	10.2	7.5	9.7	10.1	7.0	9.5	9.8	8.6	8.7
T_2												
May	11.1	12.3	8.9	9.2	7.8	10.9	8.6	12.3	9.8	10.0	9.4	9.2
Jun	13.4	10.8	9.3	10.6	9.0	10.1	7.8	14.9	11.7	10.2	8.4	8.9
Jul	13.0	10.3	12.3	13.2	9.4	9.9	8.5	14.9	11.4	10.3	9.5	10.2
Aug	15.5	11.2	12.1	12.9	10.7	11.6	9.1	12.2	11.4	12.7	9.8	11.7
T_3												
Sep	16.0	12.7	15.5	15.1	15.3	11.1	10.0	12.1	9.6	12.6	11.1	14.9
Oct	16.5	15.2	16.6	12.9	11.8	10.1	11.4	13.1	10.8	12.4	8.9	15.0
Nov	13.0	10.8	14.1	8.3	9.4	11.4	10.4	11.0	7.9	11.0	6.1	8.4
Dec	6.9	9.1	9.8	6.4	6.3	6.7	4.4	3.3	5.5	7.6	4.4	5.5

6 Conclusions

Strategists, policy makers, business executives and many others resort to forecasting for help in strategic planning, policy planning, investment and so on. With the aim of avoiding great cost or loss, forecasting has become increasingly beneficial and crucial. Therefore, this research could contribute in the form of proposing a methodology to forecast landings of fish in Malaysia, which can then help managers of fish production and fishing management.

It was discovered that models which combine machine learning with data pre-processing techniques especially those applied on fish landing data are hard to find in literature. By conducting this research, it was revealed that the proposed idea of performing pre-processing before supplying data into a forecasting model is capable of providing better performance compared to conventional models. Furthermore, a framework of the model development has been provided thus enabling future researchers to understand the methodologies employed in this research and allow them to plan enhancements on the models. More research could be done on other data pre-processing methods coupled with other forecasting models.

ACKNOWLEDGEMENTS
This research is supported by Universiti Teknologi Malaysia under the project with vote number R.J130000.7826.4F681.

References

[1] Shabri, A., & Samsudin, R. 2015. Fishery Landing Forecasting Using Wavelet-Based Autoregressive Integrated Moving Average Models. *Mathematical Problems in Engineering*, *2015*.

[2] Bako, H. Y. 2014. *Forecasting pelagic fish in Malaysia using ETS State Space Approach* (Doctoral dissertation, Universiti Tun Hussein Onn Malaysia).

[3] Koutroumanidis, T., Iliadis, L., & Sylaios, G. K. 2006. Time-series modelling of fishery landings using ARIMA models and Fuzzy Expected Intervals software. *Environmental Modelling & Software*, *21*(12), 1711-1721.

[4] Kaunda-Arara, B., Rose, G. A., Muchiri, M. S., & Kaka, R. 2004. Long-term trends in coral reef fish yields and exploitation rates of commercial species from coastal Kenya. *Western Indian Ocean Journal of Marine Science*, *2*(2), 105-116.

[5] Craine, M., Chengb, Y. W., Caputi, N., & Chubb, C. 2001. Improving Catch Predictions for Management of the Western Rock Lobster (Panulirus cygnus) Fishery Using Time Series Analysis. In *MODSIM 2001 International Congress on Modelling and Simulation*.

[6] Park, H. H. 1998. Analysis and prediction of walleye pollock (Theragra chalcogramma) landings in Korea by time series analysis. *Fisheries Research*, *38*(1), 1-7.

[7] Venugopalan, R., & Srinath, M. 1998. Modelling and forecasting fish catches: Comparison of regression, univariate and multivariate time series methods. *Indian Journal of Fisheries*, *45*(3), 227-237.

[8] Houde, E. D., & Rutherford, E. S. 1993. Recent trends in estuarine fisheries: predictions of fish production and yield. *Estuaries*, *16*(2), 161-176.

[9] Samuel, A. L. 1959. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, *3*(3), 210-229.

[10]McCulloch, W. S., & Pitts, W. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

[11]Pyle, D. 1999. *Data preparation for data mining* (Vol. 1). Morgan Kaufmann.

[12]Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. 2006. Data Preprocessing for Supervised Leaning. *International Journal of Computer Science, 1*(2), 111-117.

[13]Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., ... & Liu, H. H. 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* (Vol. 454, No. 1971, pp. 903-995). The Royal Society.

[14]Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1988. Learning representations by back-propagating errors. *Cognitive modeling*, *5*(3), 1.

[15]Nixon, S. W. 1988. Physical energy inputs and the comparative ecology of lake and marine ecosystems. *Limnology and Oceanography*, *33*(4), 1005-1025.

[16]Hyndman, R. J., & Koehler, A. B. 2006. Another look at measures of forecast accuracy. *International journal of forecasting*, *22*(4), 679-688.

[17]Hyndman, R. J., & Athanasopoulos, G. 2014. *Forecasting: principles and practice*. OTexts.

[18]Janeski, M., & Kalajdziski, S. 2011. Neural network model for forecasting Balkan stock exchanges. In *International Conference on Intelligent Computing* (pp. 17-24). Springer Berlin Heidelberg.

[19]Zhang, G. P., & Qi, M. 2005. Neural network forecasting for seasonal and trend time series. *European journal of operational research*, *160*(2), 501-514.

[20]Zhang, G. P., & Kline, D. M. 2007. Quarterly time-series forecasting with neural networks. *IEEE transactions on neural networks*, *18*(6), 1800-1814.

[21]Crone, S. F., Lessmann, S., & Stahlbock, R. 2006. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *European Journal of Operational Research*, *173*(3), 781-800.

[22]Ngo, L. B. 2011. Application of the empirical mode decomposition on the characterization and forecasting of the arrival data of an enterprise cluster.

[23]Liang, H., Lin, Q. H., & Chen, J. D. Z. 2005. Application of the empirical mode decomposition to the analysis of esophageal manometric data in gastroesophageal reflux disease. *IEEE Transactions on biomedical engineering*, *52*(10), 1692-1701.

[24]Pao, S. H., Young, C. N., Tseng, C. L., & Huang, N. E. 2010. Smoothing empirical mode decomposition: A patch to improve the decomposed accuracy. *Advances in Adaptive Data Analysis*, *2*(04), 521-543.

[25]Drisya, G. V., & Kumar, K. S. 2015. Empirical mode decomposition and chaos based prediction model for wind speed oscillations. In *2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS)* (pp. 306-311). IEEE.

[26]Zhang, H., Singh, V. P., Wang, B., & Yu, Y. 2016. CEREF: A hybrid data-driven model for forecasting annual streamflow from a socio-hydrological system. *Journal of Hydrology*, *540*, 246-256.

[27]Lahmiri, S. 2016. Interest rate next-day variation prediction based on hybrid feedforward neural network, particle swarm optimization, and multiresolution techniques. *Physica A: Statistical Mechanics and its Applications*, *444*, 388-396. 388-396.

[28]Zhang, C., Wei, H., Zhao, J., Liu, T., Zhu, T., & Zhang, K. 2016. Short-term wind speed forecasting using empirical mode decomposition and feature selection. *Renewable Energy*, *96*, 727-737.

[29]Yang, D., & Yang, K. 2016. Multi-step prediction of strong earthquake ground motions and seismic responses of SDOF systems based on EMD-ELM method. *Soil Dynamics and Earthquake Engineering*, *85*, 117-129.

[30]Liu, H., Chen, Q., Li, B., Mao, X., Mao, K., & Peng, F. 2011. On-line chatter detection using servo motor current signal in turning. *Science China Technological Sciences*, *54*(12), 3119-3129.

[31]Typhoon Utor to blame. 2006. *The Star Online*. Retrieved from http://www. thestar.com.my/.

[32]2 killed, 60,000 displaced in worst flooding in a century in southern Malaysia. 2006. *USA Today*. Retrieved from http://www.usatoday.com/.

[33]Skillicorn, D. 2007. *Understanding complex datasets: data mining with matrix decompositions*. CRC press.

[34]Castle, E. 2015. *7 Signs You're Dealing with Complex Data*. Retrieved from https://www.sisense.com/blog/7-signs-youre-dealing-with-complex-data/.

[35]Ho, D. J., Siti Maryam, D., Jafar-Sidik, M., & Aung, T. 2013. Influence of weather condition on pelagic fish landings in Kota Kinabalu, Sabah, Malaysia. *J Tropic Biol Conserv*, *10*, 11-21.

[36]Jafar-Sidik, M., Aung, T. H., & Singh, A. M. 2010. Sensitivity of fish landings to some meteorological parameters: a case study. *American Journal of Environmental Sciences*, *6*(2), 177-183.

[37]Moses, B. S. 1987. The influence of flood regime on fish catch and fish communities of the Cross River floodplain ecosystem, Nigeria. *Environmental Biology of Fishes*, *18*(1), 51-65.

[38]Strong, J. W., & Criddle, K. R. 2014. A Market Model of Eastern Bering Sea Alaska Pollock: Sensitivity to Fluctuations in Catch and Some Consequences of the American Fisheries Act. *North American Journal of Fisheries Management*, *34*(6), 1078-1094.

[39]Breitburg, D. L., Craig, J. K., Fulford, R. S., Rose, K. A., Boynton, W. R., Brady, D. C., ... & Hart, D. R. 2009. Nutrient enrichment and fisheries exploitation: interactive effects on estuarine living resources and their management. *Hydrobiologia*, *629*(1), 31-47.

[40]Keppeler, F. W., Hallwass, G., & Silvano, R. A. M. 2016. Influence of protected areas on fish assemblages and fisheries in a large tropical river. *Oryx*, 1-12.