NIPES Journal of Science and Technology Research 3(4) 2021 pp.233 – 244 pISSN-2682-5821, eISSN-2682-5821

Journal of Science and Technology Research

Journal homepage: www.nipesjournals.org.ng

Performance of Multiple Linear Regression and Artificial Neural Network in Predicting Risk Index

[1*]Ekiugbo A., [2]Amiolemhen P. and [3]Ariavie. G.O.

[1,2]Department of Production Engineering, Faculty of Engineering, University of Benin, Benin City, Edo State, Nigeria.
[3]Department of Production Engineering, Faculty of Engineering, University of Benin, Benin City, Edo State, Nigeria.
*afokeekiugbo@gmail.com, patrick.amiolemhen@uniben.edu, ariaviefe@uniben.edu

Article Info

Keywords:
Risk index, Vandalism, Reliability analysis, Multiple linear regression, Artificial neural network

.
Received 26 October 2021
Revised 19 November 2021
Accepted 16 November 2021
Available online 12 December 2021

https://doi.org/10.37933/nipes/3.4.2021.24

https://nipesjournals.org.ng

Abstract

The act of deliberate damage to oil and gas equipment and facilities has been a common phenomenon in Nigeria and has posed huge risk of economic, social and political effects on the people, companies and the environment. The target of this study is to assess the performance of multiple linear regression (MLR) and artificial neural network (ANN) for the prediction of risk index. Twenty-six (26) years secondary data were obtained from the archive of the Nigerian National Petroleum Corporation (NNPC), covering the period of 1989 to 2016 and capturing information on risk index, vandalism, rupture, spillage and volume. To assess the quality of the data, preliminary analysis involving; reliability analysis using one-way analysis of variance and test of homogeneity was done. To examine the correlation and determine the exact relationship between the risk index and other independent variables, selected linear and non-linear models were employed. For the linear model, multiple linear regression (MLR) was used and for the non-linear model, artificial neural network (ANN) was employed. The calculated p-value of 0.000 based on the reliability test shows that the data are reliable and adequate. On the overriding influence of the selected independent variables, vandalism was observed to be positively correlated with risk index thus making this variable the most significant variable that influence risk index compares to other independent variables such as rupture, spills and volumes. On the best fit model for risk index prediction, result of regression plot of output revealed that; artificial neural network with R^2 value of 0.9973 was acclaimed the best model ahead of multiple linear regression with R^2 value of 0.5640.

1. Introduction

Pipelines play an important role in modern societies and are crucial for providing fuels for power generating transportation. In the light of the hazardous properties of the product been transported through these pipelines a ruptured pipeline has the potential to cause serious environmental damage and this problem is further compounded by the fact that many developing countries have

not established proper guidelines that will help to minimize exposure to such risk [1]
The act of deliberate damage to oil and gas facilities has been a common problem in Nigeria and has posed huge risk of economic and social effects on the people, companies and the environment [2, 3]. [4] reported that it is an undisputable fact that risk control measures increase the importance of firms and may reduce business distress. Interruptions in oil production caused by fires and accidents easily lead to significant economic losses, and potential hazards to humans and the environment. [5] defined the concept of risk as the likelihood of specific consequences happening, they considered the theory of error and concluded that such error is valuable input to any probabilistic risk analysis. [6] did a study on pipeline risk and discovered that it is different from other plant risk because the risk is associated with a line source rather than a series of point sources of risk. [7], in his comparative studies of the Nigerian oil industry with its US counterpart, maintained that the Nigeria oil companies need more hazard analysis and risk assessment because most parts of Nigeria meet the criteria defined in the U.S as high consequence areas for oil spills (populated area, drinking water area, or productive ecosystem). Oil companies in Nigeria are implicitly required by Nigerian law to comply with international standards like the API standards for high consequence areas and, therefore, require more hazard analysis and risk assessment. Intentional Damage (Sabotage), is a deliberate act of people causing damage to assess, properties and the environment. It happens due to several reasons, causing disruption in the production and distribution of petroleum products. A major cause of sabotage includes economic backwardness of the region, where the pipeline is located, leading to agitation and violence. In particular, [8] attributed lack of employment and environmental degradation as major factors causing pipeline sabotage in the Niger-Delta region of Nigeria, where incessant vandalism of oil and gas production facilities and installations are rife. The impact of sabotage can be very disastrous leading in some cases to loss of lives, extensive environmental pollution, degradation, and huge economic losses [9, 10]

2. Methodology

2.1 Description of study area

The map of oil pipeline network in Nigeria is presented in Figure 1. Distinction of the different Pipelines is based on the colour codes with green colour representing pipeline carrying oil red for gas and blue for products, such as gasoline, propane and ethylene.

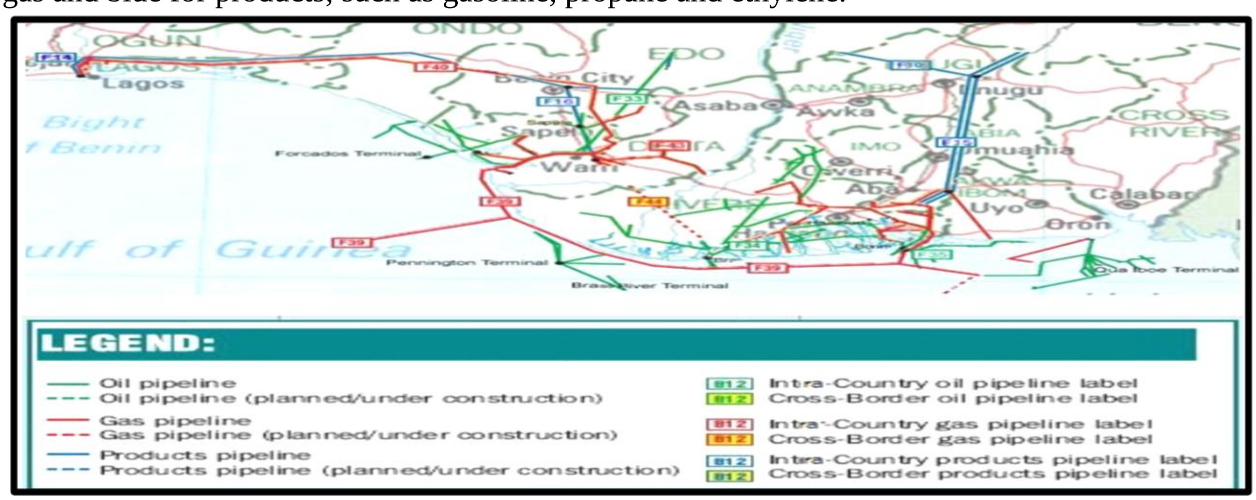

Figure 1: Nigeria pipelines map (Source: theodora.com/pipelines/)

2.2 Data Collection

The data used are pipeline data. Twenty-three (23) years' secondary data covering from 1990 to 2016 on risk index, pipe vandalism, pipe rupture, oil spill and oil volume obtained from the archive of the Nigerian national petroleum corporation annual report bulletin of operations was employed.

2.3 Preliminary Analysis of Data

2.3.2 Test of Homogeneity

Frequency analysis of data requires that the data be homogeneous and independent. Homogeneity test was carried out to establish the fact that the data used for the analysis are from the same population distribution. Homogeneity test is based on the cumulative deviation from the mean as expressed using the mathematical equation presented by [11].

$$S_k = \sum_{i=1}^{k} \left(X_i - \bar{X} \right) \quad k = 1, \qquad (1)$$

where

X_i = The record for the series X_1 X_2 ----------- X_n

$\bar{X}$ = The mean

S_{ks} = the residual mass curve

For a homogeneous record, one may expect that the S_{ks} fluctuate around the zero-center line in the residual mass curve. To perform the homogeneity test, a software package (Rainbow) for analyzing time series data was employed.

2.3.3 Reliability Analysis of the Data:

Reliability analysis of the data was done to ascertain the fitness of the data for the selected analysis. Descriptive analysis of the reliability test was based on the data scale (measured in terms of weight and order of distribution). The summary statistics was done to compute the data means, variance, covariance and correlations using the intra class correlation coefficient. The null and alternate hypothesis of reliability was formulated as follows;

H_0: for p-value < 0.05, data is reliable

H_1: for p-value > 0.05, data is not reliable

For computed (p-value < 0.05), the null hypothesis was accepted and it was concluded that the data are reliable

2.4 Risk Index Prediction Models

2.4.1 Multiple Linear Regression Model

To apply multiple linear regression models, the independent variables that influences risk index were thoroughly filtered and analyzed. The selected independent variables include;

i. No. of vandalization (X_1)
ii. No. of Rupture (X_2)
iii. No. of Spills (X_3)
iv. Volume of oil (X_4)

The dependent variable is risk index. To ascertain the dependence of the selected independent variables on the dependent variable, multiple linear regression models was applied to generate a regression equation of the form:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \text{---------} \; b_5 x_5 + e \qquad (2)$$

where;

X_1, X_2 ---------- X_n = the selected independent variables

Y = the dependent variable (Rate of accident),

β_0 , β_1 are the regression constant;

Ƹ is the deviation.

To execute the multiple linear regression modelling and generate the regression equation, statistical software (EViews 9.0) was employed.

2.4.2 Artificial Neural Network

To apply neural network, 60% of the data was employed to train a network, 25% of the data was used to validate the network while the remaining 15% was used to test the performance of the network. The neural network modelling and prediction was done with the aid of a neural network modelling software (MATLAB 10.1).

To avoid the problem of weight variation which can subsequently affect the efficiency of the training process, the input and output data were first normalized to obtain a value of between 0.1 and 1.0 using the normalization equation proposed by [11] as follows;

$$x_i = \frac{x - x_{\min}}{x_{\max} - x_{\min}} + 0.1 \qquad (3)$$

Where;

x_i is the normalized value of the input and output data

x_{min} and x_{max} is the minimum and maximum value of the input and output data

Input and output data training resulting in the design of network architecture is of paramount importance in the application of neural network to data modelling and prediction. To obtain the optimal network architecture that possesses the most accurate understanding of the input and output data, two major factors were considered;

i. First is the selection of the most accurate training algorithm and secondly,
ii. The number of hidden neurons.

Based on this consideration, different training algorithm and hidden neurons were tested to determine the best training algorithm and the optimum number of hidden neurons that will produce the most accurate network architecture. Selectivity of the training algorithm and the optimum number of hidden neurons was based on the coefficient of determination (R^2) and the mean square error value (MSE) [13].

To train the network, 3 runs of 1000 epochs, each with a precision rate of 0.00001 and a learning rate of 0.05 was used. In addition, cross validation data representing about 25% of the total input data was introduced to monitor the progress of training and prevent the network from memorizing the input data instead of leaning which is a common problem associated with overtraining [11]. The progress of the training was monitored using the mean square error (MSE) graph for training and cross validation.

3.0 Results and Discussion

Descriptive statistics of the data employed for the study is presented in Table 1.

Table 1: Descriptive statistics of data used

Descriptive Statistics	N	Range	Minimum	Maximum	Mean		Std. Deviation	Variance	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Risk index	23	.1	.2	.3	.255	.0091	.0436	.002	.139	.481	-1.625	.935
No. of vandalization	23	155	350	505	436.83	11.089	53.182	2.828E3	-.175	.481	-1.522	.935
No. of rupture	23	35	13	48	29.57	1.523	7.304	53.348	.150	.481	1.302	.935
No. of spills (tons)	23	514	1	515	232.30	28.840	138.313	1.913E4	.427	.481	-.573	.935
Volume (barrels)	23	106790	150	106940	3.69E4	6486.680	31109.022	9.678E8	1.274	.481	.876	.935
Valid N (listwise)	23											

From the results of Table 1, it was observed that the mean ± deviation for risk index is 0.255±0.0436, for number of vandalization it is 436.83±53.182, for number of rupture it is 29.57±7.304, for number of spills it is 232.30±28.840 while for volume it is 36900±31109.022. In order to ascertain that the data used are from the same population distribution, test of homogeneity was done and results obtained is presented in Figures 1, 2, 3 and 4.

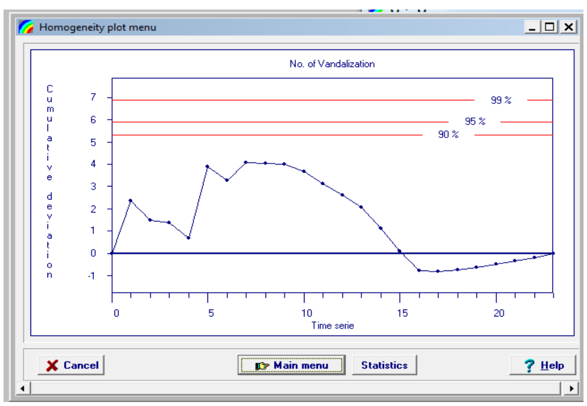

Figure 1: Homogeneity test of risk index of number of vandalization

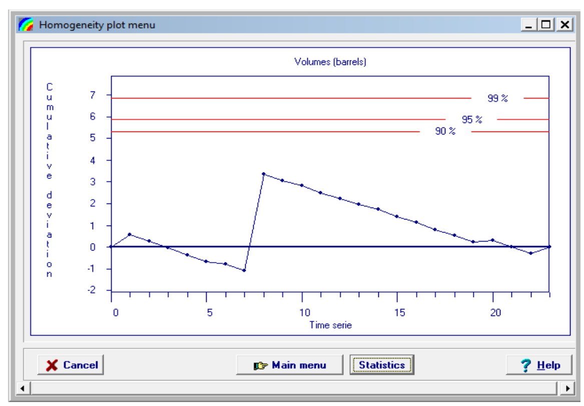

Figure 3: Homogeneity test for volumes (barrels)

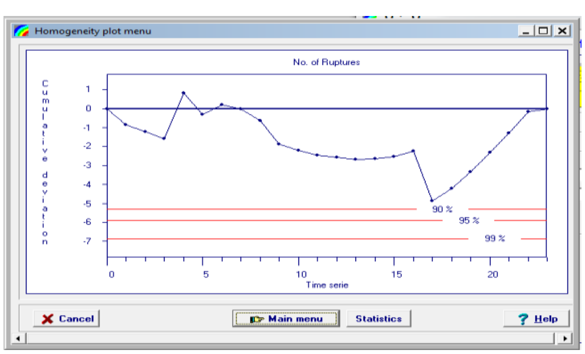

Figure 2 Homogeneity test of number of rupture

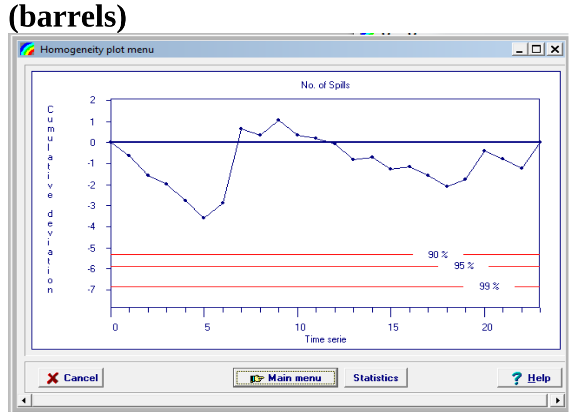

Figure 4: Homogeneity test of number of spills

From the result of Figures 1 to 4, it was observed that the data fluctuates around the zero-center line of the residual mass curve an indication that the data are statistically homogeneous. Before model development, it was partinent to test the reliability of the data in order to ascertain the data fitness. The calculated inter-item correlation ad covariance statistics of reliability is presented in Table 2.

Table 2: Inter-item correlation and covariance statistics

Inter-Item Correlation Matrix

	Risk Index	No. of Vandalization	No. of Rupture	N0. of Spills	Volume (barrels)
Risk Index	1.000	.500	-.369	-.165	-.336
No. of Vandalization	.500	1.000	.149	.587	-.020
No. of Rupture	-.369	.149	1.000	.542	.331
N0. of Spills	-.165	.587	.542	1.000	.308
Volume (barrels)	-.336	-.020	.331	.308	1.000

Inter-Item Covariance Matrix

	Risk Index	No. of Vandalization	No. of Rupture	N0. of Spills	Volume (barrels)
Risk Index	.002	1.158	-.117	-.992	-455.024
No. of Vandalization	1.158	2828.332	57.694	4319.510	-32618.121
No. of Rupture	-.117	57.694	53.348	547.638	75191.903
N0. of Spills	-.992	4319.510	547.638	19130.585	1323335.745
Volume (barrels)	-455.024	-32618.121	75191.903	1.323E6	9.678E8

From the result of Table 2, it was observed that vandalization is the only independent variable that is positively correlated with risk index thus making this variable the most significant variable that influence risk index compares to other independent variables such as rupture, spills and volumes. The computed coefficient of correlations; -0.369 for No. of rupture, -0.165 for No. of spills, and -0.336 for volume were observed to be relatively weak and indicative of the absence of collinearity problem in the regression variables. The highest coefficient of (-0.165) which is between No. of spill and risk index did not pose any challenge of multicollinearity since the value is still far less than 10. Hence, we can conclude that there is no issue of multicollinearity and that the regression variables are clearly correlated with the dependent variable. This is evident in the intra-class correlation coefficient presented in Table 3.

Table 3: Computed intra-class correlation coefficients

Intraclass Correlation Coefficient

	Intraclass Correlation[a]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.001[b]	-.100	.182	1.004	22	88	.469
Average Measures	.004[c]	-.827	.526	1.004	22	88	.469

Again, we observed from the result of Table 3 that the single and average measure intra-class correlation coefficients are relatively weak (0.001 and 0.004) and indicative of the absence of multicollinearity. To ascertain the reliability of the data, one-way analysis of variance (ANOVA) was generated and presented in Table 4.

Table 4: Analysis of variance table

ANOVA

		Sum of Squares	df	Mean Square	F	Sig
Between People		4.270E9	22	1.941E8		
Within People	Between Items	2.478E10	4	6.195E9	32.028	.000
	Residual	1.702E10	88	1.934E8		
	Total	4.180E10	92	4.544E8		
Total		4.607E10	114	4.041E8		

At 0.05 df, with a computed p-value of 0.000 as observed in Table 4, the null hypothesis of reliability was accepted and it was concluded that the data are good and can be employed for further analysis. To develop the model for risk index prediction, multiple linear regression and artificial neural network were employed. The coded least square regression equation developed for this study is presented in Equation 1 and the regression output is presented in Table 5.

Table 5: Output of Regression Analysis

Equation: EQ1 Workfile: RISK DATA::Untitled\

View | Proc | Object | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: RISK_INDEX
Method: Least Squares
Date: 03/16/20 Time: 16:31
Sample: 1 23
Included observations: 23

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.030079	0.072885	0.412696	0.6847
N0__OF_SPILLS	-0.000167	7.29E-05	-2.297272	0.0338
NO__OF_RUPTURE	-0.001026	0.001103	-0.930180	0.3646
NO__OF_VANDALIZATION	0.000685	0.000159	4.293926	0.0004
VOLUME__BARRELS_	-1.38E-07	2.29E-07	-0.603111	0.5540

R-squared	0.601641	Mean dependent var	0.254783
Adjusted R-squared	0.513117	S.D. dependent var	0.043575
S.E. of regression	0.030406	Akaike info criterion	-3.958720
Sum squared resid	0.016641	Schwarz criterion	-3.711873
Log likelihood	50.52528	Hannan-Quinn criter.	-3.896638
F-statistic	6.796346	Durbin-Watson stat	1.434503
Prob(F-statistic)	0.001621		

From the result of Table5, the following observations were made

i. With a regression (p-value) of 0.6847, it was concluded that the regression analysis was not significant at 0.05 degree of freedom. Hence, it was declared that the relationship between risk index and selected sets of independent variables such as number of vandalization, number of spills, number of rupture and volume is not linear.

ii. Independent variables, namely; number of spills and number of vandalization were observed to have a very strong influence on the risk index with a probability (p-value) of 0.0338 and 0.0004 respectively.
iii. The poor regression terms such as coefficient of determination and adjusted coefficient of determination was apparently due to the fact that the variables are not from a normal probability distribution.

Using the result of Table 5, the overall regression equation was thereafter generated and presented in Equation 2

$$(\text{Risk Index}) = 0.030079 - 0.000167(\text{Spills}) - 0.001026(\text{Rupture}) + 0.000685(\text{Vandalization}) + 1.38\text{E-}07(\text{Volume}) \quad (2)$$

To assess the strength of the regression model, a regression plot of output between the observed risk index and regression predicted risk index was obtained and presented in Figure 5.

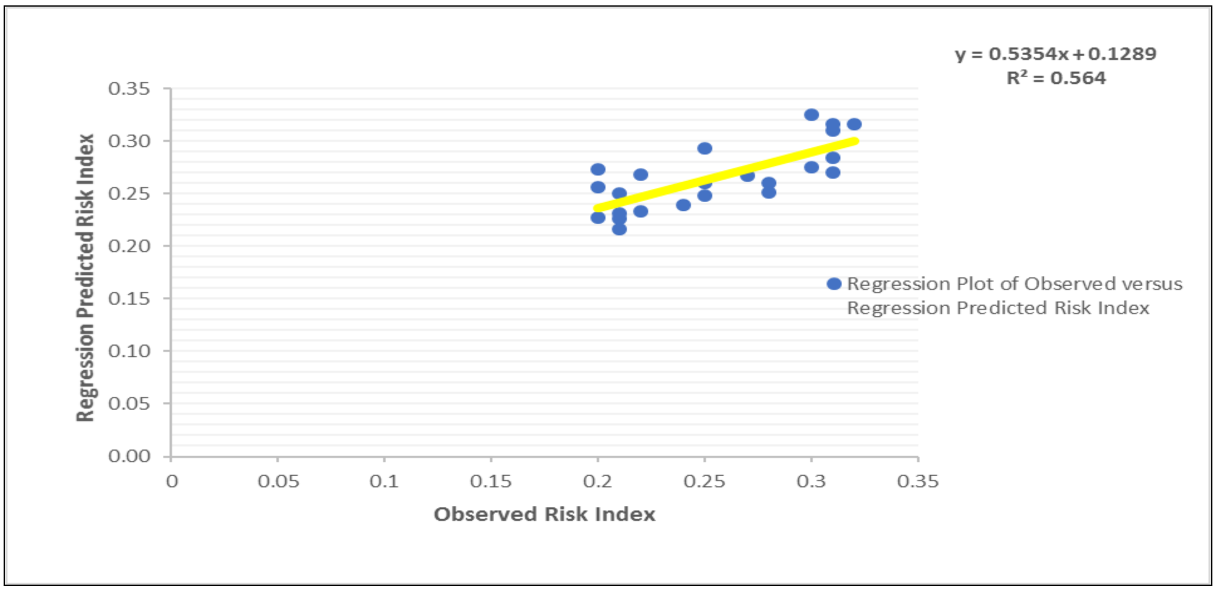

Figure 5: Observed versus predicted regression predicted risk index

Coefficient of determination (R^2) value of 0.564 shows the low strength of regression model in predicting the risk index. On account of the poor strength of regression model, artificial neural network (ANN) was employed to predict the risk index.
To obtain the optimal network architecture that possesses the most accurate understanding of the input and output data, two factors were considered. First, is the selection of the most accurate training algorithm and secondly, the number of hidden neurons. Based on this consideration, different training algorithm and hidden neurons were selected and tested to determine the best training algorithm and accurate number of hidden neurons that will produce the most accurate network architecture. Selection of optimum training algorithm was based on the computed coefficient of determination (r^2) and the mean square error (MSE). Table 6 shows the performance of the different training algorithm tested.

Table 6 Selection of optimum training algorithm for ANN modelling and prediction

S/No	Training Algorithm (Learning Rule)	Training MSE	Cross Validation MSE	R-Square (r^2)
1	Gradient information (Step)	0.0455	0.0434	0.63
2	Gradient and weight change (Momentum)	0.0403	0.0721	0.84
3	Gradient and rate of change of gradient (Quick prop)	0.0665	0.0388	0.72
4	Adaptive step sizes for gradient plus momentum (Delta Bar Delta)	0.0504	0.0407	0.61
5	Second order method for gradient (Conjugate gradient)	0.0511	0.0700	0.64
6	Improved second order method for gradient (Levenberg Marquardt)	0.0018*	0.0014*	0.97*

Result of Table 6 revealed that improved second order method of gradient also known as Levenberg Marquardt Back Propagation training algorithm was the best learning rule and was therefore adopted in designing the neural network. To determine the exact numbers of hidden neuron, different numbers of hidden neurons were selected to create a trained network using Levenberg Marquardt Back Propagation training algorithm. Performance of the trained network was assessed using mean square error (MSE) and coefficient of determination r^2. The number of hidden neurons corresponding to the lowest MSE and the highest r^2 as presented in Table 7 was selected to design the neural network.

Table 7 Selection of optimum number of hidden neurons for ANN

S/No	Number of Hidden Neurons	Training MSE	Cross Validation MSE	R-Square (r^2)
1	3	0.0229	0.0633	0.81
2	4	0.0401	0.0321	0.75
3	6	0.0599	0.0205	0.72
4	8	0.0310	0.0401	0.87
5	10	0.0035	0.0019	0.98

Based on the results of Table 6 and 7, Levenberg Marquardt Back Propagation training algorithm having 10 hidden neurons in the input layer and output layer was used to train a network of 4 input processing element (PE) and 1 output processing element. The input layer of the network uses the hyparbolic targent (tan-sigmoid) transfer function to calculate the layer output from the network input while the output layer uses the linear (purelin) transfer function. The number of hidden neuron was set at 10 neurons per layer and the network performance was monitored using the mean square error of regression (MSEREG). In addition, a learning rate of 0.01, momentum coefficient of 0.1, target error of 0.01, analysis update interval of 500 and a maximum training cycle of 1000 epochs was used. The network generation process divides the input data into training data sets, validation and testing. For this study, 60% of the data was employed to perform the network training, 25% for validating the network while the remaining 15% was used to test the performance of the network. A performance evaluation plot which shows the progress of training, validation and testing is presented in Figure 6.

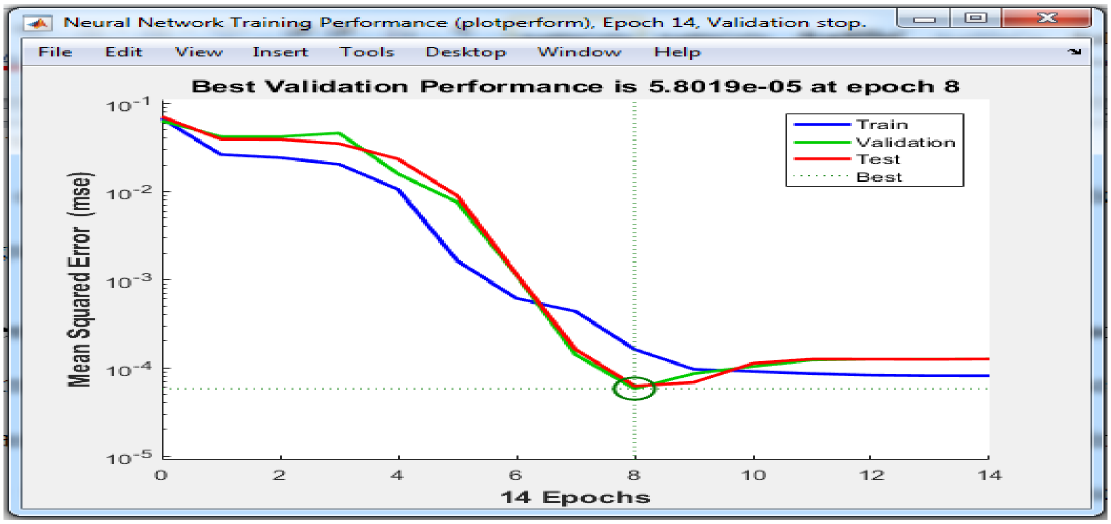

Figure 6: Performance curve of trained network for predicting risk index

From the performance plot of Figure 6, no evidence of over fitting was observed. In addition, similar trend was observed in the behaviour of the training, validation and testing curve which is expected since the raw data were normalized before use. Lower mean square error is a fundamental criterion used to determine the training accuracy of a network. An error value of 5.8019e-05 at epoch 8 is an evidence of a network with strong capacity to predict risk index. The training state, which shows the gradient function, the training gain (Mu) and the validation check, is presented in Figure 7.

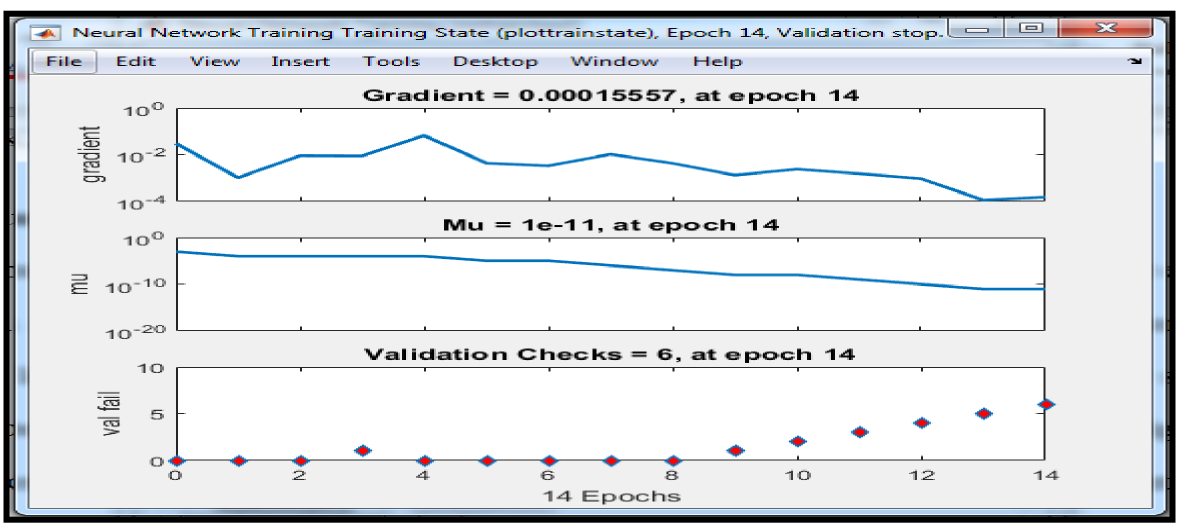

Figure 7: Neural network training state for predicting risk index

Back propagation is a method used in artificial neural networks to calculate the error contribution of each neuron after a batch of data training. Technically, the neural network calculates the gradient of the loss function to explain the error contributions of each of the selected neurons. Lower error is better. Computed gradient value of 0.00015557 as observed in Figure 8 indicates that the error contributions of each selected neuron are very minimal. Momentum gain (Mu) is the control parameter for the algorithm used to train the neural network. It is the training gains and its value must be less than one. Momentum gain of 1.0e-11 shows a network with high capacity to predict risk index. The regression plot which shows the correlation between the input variables (number of vandalization, number of spills, number of rupture and volume) and the target variable (risk index) coupled with the progress of training, validation and testing is presented in Figure 8.

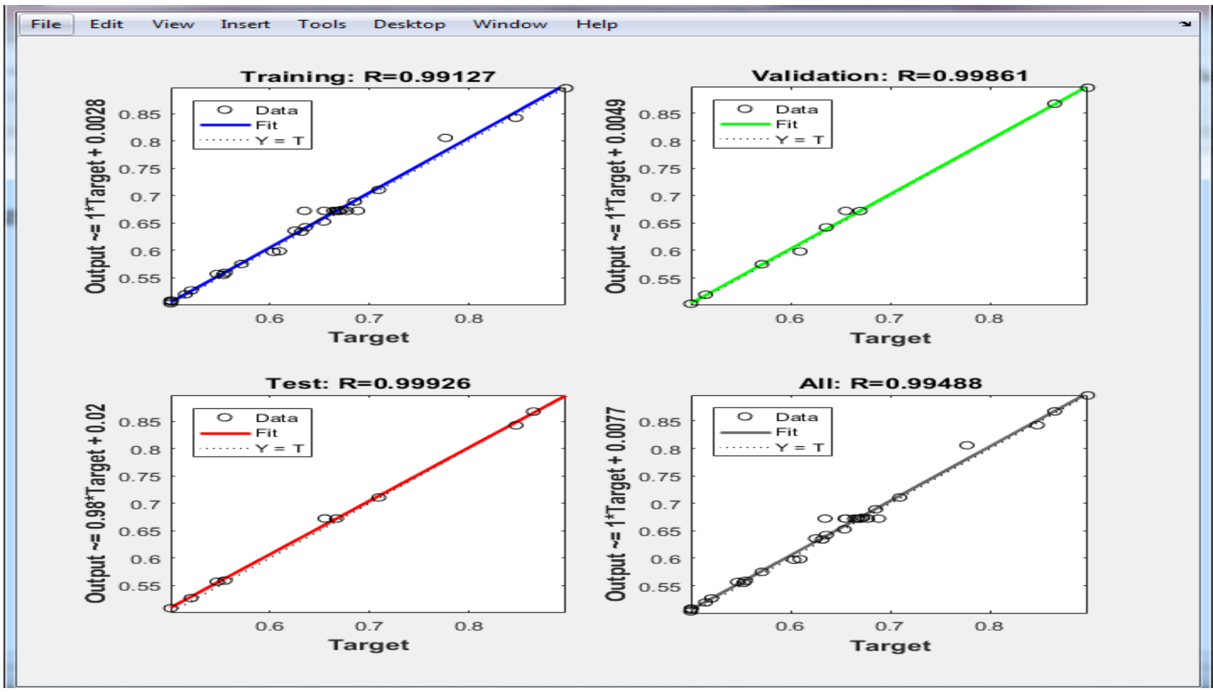

Figure 8: Regression plot showing the progress of training, validation and testing

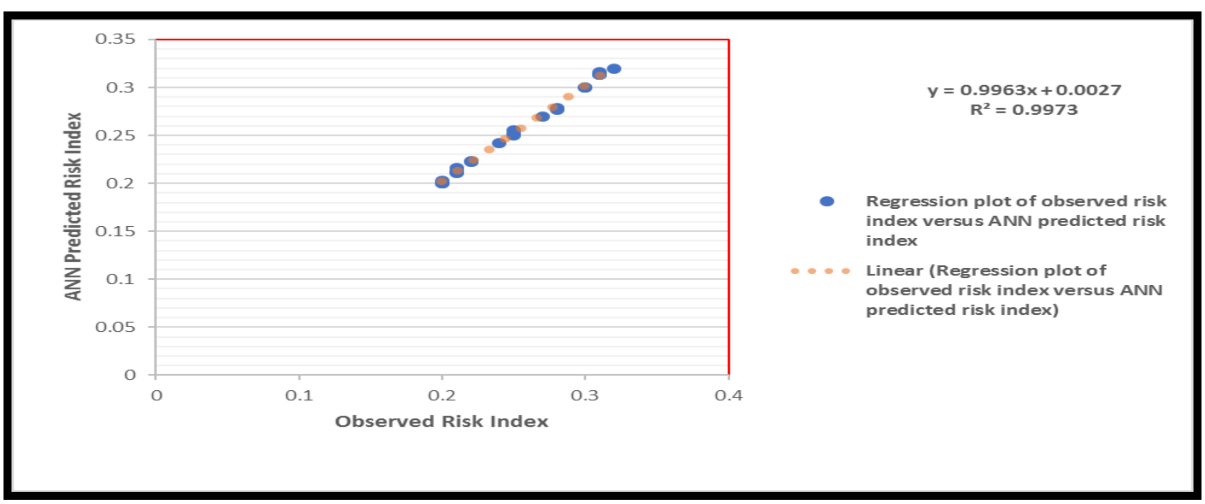

Figure 9: Regression plot of observed versus ANN predicted risk index

Based on the computed values of the correlation coefficient (R) as observed in Figure 8, it was concluded thet the network has been accurately trained and can be employed to predict the risk index. To assess the strength of ANN model, a regression plot of output between the observed risk index and ANN predicted risk index was also obtained and presented in Figure 9. Coefficient of determination (r^2) values of 0.9973 as observed in Figures 9 was employed to draw a conclusion that the trained network can be used to predict risk index using different sets of input parameters.

4.0 Conclusion

This study attempt to predict risk index using selected sets of independent variables, namely; vandalism, spill, pipe rupture and oil volume on risk index. The study employed two forms of supervised learning techniques, namely; multiple linear regression technique and artificial neural network. Twenty-six (26) years secondary data obtained from the archive of the Nigerian National Petroleum Corporation (NNPC), covering from 1991 to 2016 was used. The quality of the data was examined by means of selected statistical techniques, namely reliability analysis using one-way analysis of variance and test of homogeneity. The outcome revealed that the data are adequate and homogeneous.
On the performances of the different prediction models employed in this study, it was observed that; though the selected model shows potential capability to predict risk index with artificial neural network been a better model compared to regression method. The direct implication is that; the exact relationship between risk index and the selected independent variables cannot be determined using linear modelling techniques. Although the content of this study is not completely exhaustive on the subject matter, it has provided additional information to the existing literatures on risk index assessment, modelling and prediction.

References

[1] Abubakar, U. (2015) An overview of the occupational safety and health systems of Nigeria, UK, USA, Australia and China: Nigeria being the reference case study. American Journal of Educational Research, 3(11), pp. 1350-1358.

[2] Adebiyi K.A and Owaba-Charles, O. E (2009). Towards setting a sustainable manufacturing safety program in Nigeria Disaster Prevention and Management. 18(4): 388-396.

[3] Adeogun, B. K., and Okafor, C. C. (2013) Occupational health, safety and environment (HSE) trend in Nigeria. International Journal of Environmental Science, Management and Engineering Research, 2 (1), pp. 24-29.

[4] Aishatu, S., Chukwudi, O. J., and Hauwa'u, I. (2016). The Environmental Impact of Pipeline Vandalism. A Challenge to Biodiversity in Port Harcourt Area of Rivers State, Nigeria.

[5] Ekpu, A. and Ehighelua, I. (2004) Sabotage in the Oil Industry-The role of the Law, Environmental Policy and Law, Vol. 34, Issue 6, pp. 274 – 280.

[6] Ekuerhare, B.U. and Orubu, C. O., (1996); Economics of the Niger Delta, Technical Report, Niger Delta Environmental Survey, Port-Harcourt, pp:56-67

[7] El-Choueiry, J., Feather, A., Jeffrey, J. and Shaw, S. 2013. Federal Republic of Nigeria: Risk Assessment Report. Norman Paterson School of International Affairs, Carleton University. [Online]. Available: http://www,carleton.ca.cifp/app/1413, pp: 33-67

[8] Jo, Y. D and Ahn, B. J. (2005) A method of quantitative risk assessment for transmission pipeline carrying natural gas', Journal of Hazardous Materials, A123, pp. 1–12.

[9] Steiner, R. (2010), Double standard: Shell practices in Nigeria compared with International standards to prevent and control pipeline oil spills and the deep water horizon oil spill. Amsterdam, The Netherlands: Milieudefensie, pp:11-35.

[10] Taylor, J. R. (1994), Risk analysis for process plant, pipelines and transport, Taylor and Francis. The Nigeria Petroleum Industry Bill (2002). http://www.nigeria- law.org/Legislation/LFN/2012, pp: 23-45

[11] Yanting, Z. and Liyun, X. (2011) 'Research on Risk Management of Petroleum Operations', Energy Procedia, 5(0), pp. 2330-2334