Journal of Theoretical and Applied Information Technology
20th July 2015. Vol.77. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A CONCEPTUAL FRAMEWORK FOR APPROACHING PREDICTIVE MODELING USING MULTIVARIATE REGRESSION ANALYSIS VS ARTIFICIAL NEURAL NETWORK

[1]VIDYULLATHA PELLAKURI, [2] D RAJESWARA RAO, [3]*P* LAKSHMI PRASANNA, [4] M.V.B.T SANTHI

[1]Research Scholar, Department of CSE, KL University, Guntur (DT), Andhra Pradesh.

[2]Department of Computer Science & Engineering, Guntur (DT), Andhra Pradesh.

[3,4]Assistant Professor, Department of CSE, KL University, Guntur (DT), Andhra Pradesh.

E-mail: [1]lekana04cuty@gmail.com, [3]pprasanna@kluniversity.in

ABSTRACT

The use of artificial neural networks is a promising approach for prediction of fine particles concentrations under variable meteorological conditions. This paper analyzes the statistical analysis of Multivariate Regression Analysis (MVRA) versus Artificial Neural Networks (ANN) and investigations were performed on real statistical data set obtained from measurements of the process parameters of recent six months data under industrial conditions. Most influential statistical parameters such as R, R-square, Adjusted R-square, MAE, RMSE are evaluated for choosing right modeling tool in this investigation.

Keywords: *Artificial Neural Network, Back Propagation, Levenberg-Marquardt Algorithm, Meteorological Parameters, Multivariate Regression Analysis, Nntool.*

1. INTRODUCTION

Regression analysis is necessary in modern science problems for almost all data reduction. This is one type of machine learning method using mathematical equations. In a regression analysis, the relationship is studied between the dependent and independent variables and if one input is chosen it is linear regression model and two or more inputs are selected it is termed as multiple linear model. There are three types of regression models which are linear regression, Non-Linear regression and Polynomial regression. In this paper, ambient pollutant respirable particulate matter is assumed as dependent variable and two weather parameters wind speed, temperature and two ambient pollutants - oxides of nitrogen, particulate matter are considered as independent variables. Numerous studies have been made for prediction of air pollutants which are detailed in this literature survey. In 2015, Andrew C. Comrie [1] examines the multiple regression models versus neural networks on nonlinear relationships between the pollutants and weather parameters for ozone predictions and proved that neural network techniques are somewhat better than regression models for daily ozone prediction. In 2014, Elmira Shamshiry etal [2] studied the forecasting of solid waste generation in Langkawi Island using ANN model and Multiple Regression Analysis (MRA). In his study, comparison is made on statistical parameters such as R2, MAE, MARE and RMSE and proved that ANN model is better than MRA. Prakash G L etal (2014) [3] developed a multi layer neural network model using neural network toolbox in Matlab for optimal controlling of incinerator in sulphur recovery block of refineries. Mahad S baawian etal (2014) [4] proved that the value of coefficient of determination R^2 exceeds 0.70 for prediction of ground level air pollutants by varying of temperature using the multilayer back propagation algorithm. Lubna B Mohammed etal (2013), [5]used a Marquardt-Levenberg learning algorithm using Matlab and compare the parameters such as root mean square error and maximum coefficient of determination R for prediction of solar radiation in hourly basis. S Sapna etal (2012), [6] applied Levenberg-Marquardt algorithm which gives best performance

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

in medical field for the prediction of diabetic status. Jyoti Kumar Arora and Ponnala Vimal Mosahari (2012) [7] used single layer Artificial Neural Network (ANN) modeling to predict the noise due to road traffic in Agra-Firozabad highway using Levenberg-Marquardt algorithm (LMA) and was found the best performance at1500 Epoch with 0.70 momentum. Mohammad Zaefizadeh etal [8] suggested that yielding more barley is predicted by considering correlation, multiple regressions and multi-layer neural network (ANN) model. Olaniyi S Maliki etal [9] found that two techniques such as Neural Network and Statistical Technique for modeling and forecasting the electrical power generated of Nigeria are used to access power generation up to 2036. Isık Yilmaz a etal [10] used MLP, RBF and ANFIS methods for prediction of swell percent of soil and compared the results with the traditional statistical model of MR (multiple regression) and proved that RBF exhibited a high performance. Tharwat E. Alhanafy etal (2010) [11] applied the techniques of fuzzy logic and neural networks using matlab for forecasting nitrous oxides and sulphur dioxides emissions from industries. Rakesh kundal etal (2006) [12] demonstrated that support vector machine method is better than conventional regression approaches for blast disease prediction in rice which is helpful for plant science community and farmers. The main objective of current research is to predict the ambient pollutant respirable matter using artificial neural network model and compare the results with statistical method MVRA.

2. METHODOLOGY

This research compares two different models such as multivariate regression analysis and back propagation neural networks for prediction of respirable particulate matter which is released from industries and causing severe effects to humans. It is observed that for complex variables, the back propagation Neural Network is better for prediction when compared with multivariate regression analysis. The data on ambient air pollutants and meteorological parameters are collected per hour for six months (180 instances) of winter season from August-14 to January-15 from KTPP. After collection of data the statistical method is evaluated. Before proceeding to regression model, correlation coefficients are calculated for all the parameters and understood that all are correlated with dependent variable. After correlation, the hypothesis test is conducted at 95% confidence level. From this test, probability is greater than 0.05 so that, the relation is linear. The equation y= a + b1 x1 + b2 x2 + b3 x3 + b4x4 is chosen. For selecting the best subset of independent variables, stepwise selection method is adopted. In this method, the most correlated variable is first selected then R-square, Mallows coefficient (Cp) and Adjusted R-square are calculated and subset of variables is selected based on high adjusted R-square and lowest cp value. After applying multivariate linear regression equation $y = 11.23 + 0.297(x1) + 0.131(x2) – 0.972(x3)$. Now, this is the best fit model as far as multivariate regression analysis [13] is concerned. Then the research work is focused on developing multilayer back propagation neural network model using the neural net toolbox in Matlab R20077b [14]. The back propagation neural network consists of nodes that are organized in three tiers of input, hidden and output. The nodes in every layer compute activation functions and process the data from one layer to another layer by random generating of weights. In the output layer, the output node estimated the error and then propagated backward by adjusting the connected weights repeatedly and the network learns the relations between the input variables and predicted variable. In this work, the recorded hourly data for five months was used as training data and the one month was used as validation data. Many training algorithms are there for training neural networks. In this paper, Levenberg-Marquardt (LM) back propagation algorithm [15] is used for training the network because the error function is minimized with the combination of Gauss Newton and gradient descent method. For the use of gradient descent method, the sum of the squared errors are being reduces and the parameters are updated. The back propagation model mainly consists of network architecture and set of performance parameters. In this paper, the network model was run in a supervised manner based on trial and error technique and the architecture of 4-1-1 was developed and parameters such as correlation coefficient (R), coefficient of determination (R2), mean absolute error (MAE), root mean square error (RMSE) are calculated and shown in tables 1& 2.

3. RESULTS AND DISCUSSIONS

In this research work, the statistical method- multivariate regression and artificial neural network model-back propagation are used to investigate their ability to predict respirable particulate matter in KTPP. The data collected on winter season between August-2014 to January-2015 was used for training and testing of neural network. Initially, MVRA is used to predict

Journal of Theoretical and Applied Information Technology
20th July 2015. Vol.77. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

particulate matter with the help of four input variables. Only three variables are selected based on the decisive indicators Adj.R^2 and Cp as shown in table1. The input variable 'temperature' was deleted from the input data due to the fact that the Data was collected during winter season and hence there was not much change in temperature and it did not affect the output variable. At last, the most influence statistical analysis parameters like R, R^2, MAE, RMAE are clearly indicated as shown in the table2. For back propagation neural network, Marquardt-Levenberg training algorithm is considered because it yields the least root mean squared error (RMSE) and higher coefficient of determination (R) for the entire data in training and testing. The Marquardt-Levenberg learning algorithm training automatically stopped at epoch 4. It was found that multilayer perceptron back propagation artificial neural network model yields best results when compared to multivariate regression analysis. The network is trained with 800 epochs with four input nodes, three hidden nodes and one output node with different parameters such as epoch, time to train in seconds, performance goal, validation failures and gradient etc. By Marquardt-Levenberg algorithm, the R values for training and testing are 1.0000 and 0.9996 respectively are shown in fig1, fig2 and the graph between actual and predicted particulate matter values are shown in graphical representation in fig3.

Table1: Selection Of Best Fit Of Subset.

S.no	Subset of variables	R^2	Adj.R^2	C_p= (SSE/Var)+ 2k-n
1.	x1	0.539	0.534	27.626
2.	x1,x2	0.606	0.599	8.392
3.	x1,x2,x3	0.613	0.603	8.149
4.	x1,x2,x3,x4	0.614	0.601	9.999

Table2: Comparison Of MVRA And BPNN Model

S.no	Performance parameters	Multi Variant regression method	Back Propagation neural network model
1	R	0.783	0.9998
2	R^2	0.613	0.9988
3	Adj.R^2	0.603	0.9978
4	MAE	6.283	0.0001
5	RMSE	7.963	0.0001

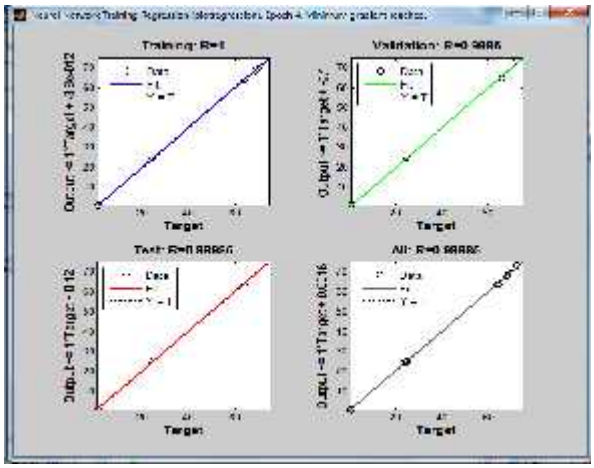

Fig1: BPNN Model Indicates Predicted Vs Observed Values

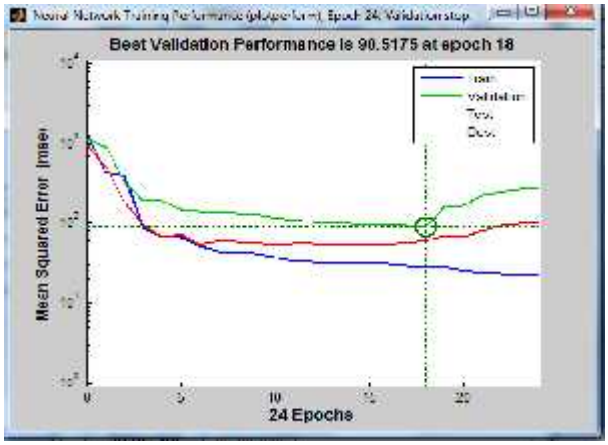

Fig 2: Plotting Of Neural Network Training Performance Parameter MSE

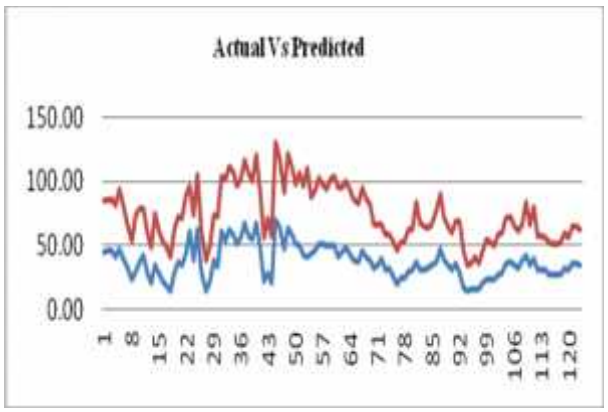

Fig3: Graph Showing Actual Values Vs Predicted Values

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. CONCLUSIONS

This paper presents a detailed analysis and comparison for prediction of particulate matter using air pollutants and meteorological variables. MVRA and multilayer perceptron neural network models have been used as the relationship between input and outputs are non-linear and complex. The data collected from power plant on an hourly basis has been averaged into daily basis which consisted of 184 instances, out of that 70% of data used as training and 30% of data used for testing. ANN model with structure 4-1-1 has been chosen as appropriate model according to three statistical indexes MAE, RMSE and R^2. In multi variant analysis, the performance parameters are 70% accurate but in artificial neural network using back propagation yields 99% accuracy for prediction so that the final results of ANN model are displayed better than that of multivariate regression models. The scope of this paper is limited to data collected during winter season only. The further work is going on to collect three years of data and is predicted by using different soft computing techniques.

5. RECOMMENDATIONS

It is observed that some of the data collected for particulate matter exceeded the limit value of national standards. As the site is situated very near to state highway road and due to heavy vehicle traffic, the release of air pollutants is more. It is recommended that particulate matter can be reduced by firing good coal, by maintaining proper combustion of coal in the boiler and supplying sufficient air for complete combustion of coal. However, the controlling steps like spraying of water on dust, less utilization of petrol and diesel vehicles and planting of trees can be taken up to reduce pollution levels at site.

REFRENCES

[1] Andrew C. Comrie, "Comparing Neural Networks and Regression Models for Ozone Forecasting", *Journal of the Air & Waste Management Association*, ISSN 1047-3289, 25 March 2015, pg: 653-663.

[2] Elmira Shamshiry, Mazlin Bin Mokhtar, and Abdul-Mumin Abdulai, "Comparison of Artificial Neural Network (ANN) and Multiple Regression Analysis for Predicting the Amount of Solid Waste Generation in a Tourist and Tropical Area—Langkawi Island", *International Conference on Biological, Civil and Environmental Engineering* (BCEE-2014) March 17-18, 2014 Dubai (UAE)

[3] Prakash G L, Samson saju, Snehil Mitra and Vedant Sharma," optimal decision support system using Multilayer neural networks for Incinerator control", *international journal of computer science and Information technologies*, ISSN: 0975-9646, vol 5 (6), 2014, 6920-6925.

[4] Mahad S baawian and aisha S AI-Serihi," systematic approach for the prediction of ground level air pollution using artificial neural network", *Aerosol and air quality research*, 14: 124-134, 2014, ISSN: 1680-8584.

[5] Lubna B Mohammed, mohammad A Hamdan, Eman A Abdelhafez and Walid Shaheen," hourly solar radiation prediction based on non-linear autoregressive exogenous neural networks", *Jordan journal of mechanical and industrial engineering*, vol 7, number 1, December 2013, pages 11-18.

[6] S Sapna, Dr. A. Tamilarasi and m Pravin Kumar," Back propagation Learning Algorithm based on Levenberg-Marquardt Algorithm", *Computer Science & Information Technology (CS & IT)*, 2012, pp: 393-398.

[7] Jyoti Kumar Arora and Ponnala Vimal Mosahari," Artificial Neural Network Modelling of Traffic Noise in Agra-Firozabad Highway", *International Journal of Computer Applications* (0975 – 8887), Volume 56– No.2, October 2012.

[8] Mohammad Zaefizadeh, Majid Khayatnezhad and Roza gholamin , "Comparison of Multiple Linear Regressions (MLR) and Artificial Neural Network (ANN) in Predicting the Yield Using its Components in the Hulless Barley", *American-Eurasian J. Agric. & Environ. Sci.*, 10 (1): 60-64, 2011

[9] Olaniyi S Maliki1 , Anthony O Agbo1 , Adeola O Maliki1 , Lawrence M Ibeh2 , Chukwuemeka O Agwu3 , "Comparison of Regression Model and Artificial Neural Network Model for the prediction of Electrical Power generated in Nigeria", *Advances in Applied Science Research*, 2011, 2 (5):329-339

[10] Isık Yilmaz a, Oguz Kaynar b, "Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils", *Expert Systems with Applications* 38 (2011) 5958–5966].

[11] Tharwat E. Alhanafy, Fareed zaghlool and Abdou Saad EI Din Moustafa, "Neuro fuzzy modeling scheme for the prediction of air pollution", *journal of American science*, 2010; 6(12), pg: 605 – 616

[12] Rakesh kundal, amar s kapoor and gajendra ps raghava," machine learning techniques in disease forecasting: a case study on rice blast prediction", *BMC informatics*, November 2006, pg:1-16.

[13] http://en.wikipedia.org/wiki/Linear_regression

[14] http://in.mathworks.com/products/neural-network/

[15] http://www.mathworks.in/help/pdf_doc/nnet/nnet_ug.pdf