Transactions on Machine
Learning and Data Mining
Vol. 13, No. 1 (2020) 3 - 25

P-ISSN: 1865-6781, E-ISSN 2509-9337
ISBN: 978-3-942952-73-6

ibai Publishing
www.ibai-publishing.org

Comparison of Multiple Linear Regression and Artificial Neural Network Models for the Prediction of Solid Waste Generation in Sri Lanka

C. L. Perera[1], MGNAS Fernando[2]

[1]University of Colombo School of Computing, Colombo 07, Sri Lanka
chathralochanie@gmail.com
[2]University of Colombo School of Computing, Colombo 07, Sri Lanka
nas@ucsc.cmb.ac.lk

Abstract. In order to plan Solid Waste Management (SWM) in a sustainable way, accurate forecasting of solid waste generation and composition plays an important role. Predicting the amount of generated waste is difficult task because it is affected by various influencing factors. Recognizing these factors is essential for implementing waste management policies to reduce the amount of waste generation. In addition to population growth and migration, underlying economic development, household size, and employment changes would influence the solid waste generation interactively. The main objectives of this study are to identify significant factors influencing solid waste generation and to develop a model to predict solid waste generation in Sri Lanka. In this study, two predictive models, Multiple Linear Regression (MLR) Models and Artificial Neural Networks (ANN) were used. The MLR model which is a conventional method showed R^2 values of *0.750*, *0.544* and *0.769* for Biodegradable, Non-Biodegradable and Total waste, respectively. On the other hand, ANN model, a non-linear model showed R^2 values of *0.846*, *0.855* and *0.902* for Non-Biodegradable, Biodegradable and Total waste, respectively, which indicated higher predictive accuracy than MLR model. Therefore, in order to develop a prediction model with a higher predictive accuracy, ANN model is recommended.

Keywords: Solid Waste Management, Solid Waste Generation, Forecasting, Influencing Factors, Multiple Linear Regression, Artificial Neural Networks

1 Introduction

Solid waste generation and management is a burning issue globally and it is extremely difficult for the planners and policy formulators to handle this issue. Improper management of solid waste worsens the air quality which causes adverse effects to human

health. The Smart City concept especially has taken the initiative in reducing greenhouse gases, using sustainable resources and managing energy resources efficiently. Developed countries have already adopted a holistic approach to waste management using an integrated solution for diversifying waste and retaining its sustainability. As a developing country, waste generation in Sri Lanka has rapidly increased over the past decade [1] due to growth of population, rapid urbanization, improved standards of living and global technological developments [2]. Therefore, it is very important for Sri Lanka too, to take immediate measures for the effective management of waste generation.

In the past, when open dumping was practiced, the main costs associated with waste management were the collection and transportation of waste. Open dumping at that time was not much costly, as the required land was freely available. However, it is now difficult to find sufficiently land for waste disposal because of its scarcity [1]. The practices commonly followed by some local authorities in Sri Lanka are open burning, land filling and open dumping of wastes even though these practices are not environmentally friendly. The conventional approach adopted by local authorities towards Solid Waste Management (SWM) is focused more on collection and disposal of waste, and no efforts on its reduction and reuse. However, their last option would be to use the waste for landfilling. The general public with their attitude 'we dump- they collect' consider SWM to be the sole responsibility of the local authorities. Open dumps, in general, are low lying degraded lands which are used only for flood retention. The majority of these open dumps are left open, while few of them have a thin layer of soil applied on top for protection [1]. Another issue is that the nature and characteristics of the solid waste generated pose challenges to local authorities and collecting, sorting and disposing and require them to have additional resources and technological support to help them to manage waste.

To provide reliable references for planning the future solid waste management and analyzing the potential of waste to energy utilization in Sri Lanka, the prediction of future waste generation quantity in the whole country is necessary and urgent to be well predicted. Therefore, the main objectives of this study is to identify significant factors influencing solid waste generation, to study the contribution of identified factors to waste generation and to identify methodologies used in other related studies, which can be used to predict solid waste generation in Sri Lanka.

Section 2 gives a literature overview about multilinear regression and artificial neural networks applied to forecast of waste generation. Section 3 describes theoretical concepts used in the study. Section 4 includes the methodology and Section 5 describes the results and description. Finally, we give conclusions in Section 6 about the prediction models devised to forecast of waste generation.

2 Literature Review

Successful planning of a solid management system mainly depends on the prediction accuracy of solid waste generation [3]. Knowing the nature of solid waste generation, such as its quantity and composition will vastly contribute for planning, operation and

optimization of Solid Waste Management System (SWMS) [4]. Prediction of waste generation is a very complex process, since it depends on many attributes both in quantitative and qualitative terms [5]. It has been found that the physical components of waste depend on the number of variables such as number of residents, household size, age groups, income, consumption pattern etc. [4]. Due to uncertainties and unavailability of historical records regarding waste generation with relevant local authorities, data driven modeling methods are needed for its prediction [6]. Modeling can be used as a tool for the accurate planning and management of waste [3].

2.1 Multiple Linear Regression Models

There are many researches who have carried out regression analysis and time series to forecast solid waste generation. Yuwanwei et al, predicted waste generation in China, using a Multiple Linear Regression (MLR) model [7] using urban population, GDP and consumption level of residents as input variables. Otoniel et al forecasted residential and non-residential solid waste using MLR [8]. Zaini et al, predicted waste generation by studying the influences of different types of houses [9]. Sara, et al, predicted the amount of residential solid waste (RSW) by considering the influences of education, income per household, and number of residents [10]. Mohammad et al, predicted solid waste generation by considering number of employees, population, household income, and temperature [11]. Hoang, et al showed that per capita urban household waste generation is *70–80%* higher compared to a rural household [12].

These models however require large number of historical data for prediction, even for a short period. In addition, the dynamic properties of waste generation cannot be fully characterized in these model formulations. To effectively handle these problems, a new analytical dynamic approach of addressing predictions of waste with reasonable accuracy, needs to be undertaken.

2.2 Artificial Neural Networks

Artificial Neural Networks (ANNs) are simplified computational models of the brain. They ANNs are capable of classifying patterns, clustering, approximating functions, forecasting and optimizing results [3].

Eduardo, et al. forecasted waste generation using an ANN using population, percentage of urban population, Years of Education, Number of Libraries and Indigent Population as influencing factors [13]. Nayseang et al. predicted the future solid waste generation in Bangkok by employing a regression model and an ANN [3]. Results revealed that ANN model had better results having R^2 value of *0.96* in comparison with the multiple regression model having R^2 value of *0.86*. Patel et al. forecasted municipal solid waste generation considering population of the town during current year, total received as tax, longitude, latitude as influencing factors [14]. Kumar et al. predicted waste generation using an ANN which uses Radial Basis Functions (RBFs) as activation functions [15]. Kannangara et al. applied decision trees and ANN to build a model for accurate prediction waste generation. Results indicated that ANN model had the best performance [16]. Kontokosta et al. presented a new analytical approach which combines

gradient boosting regression trees and ANN models to estimate daily and weekly waste generation at the building scales [17]. Elmira et al. predicted the weekly waste generation using an ANN by considering number of trucks, personnel and fuel cost as influencing factors [18]. Siti et al. predicted SWG using an ANN based on population growth factor [19]. David et al. developed Autoregressive Moving Average and the ANN models in forecasting of MSW generation [20]. Mohammad et al. evaluated the accuracy of the prediction of SWG by comparing between the results of the multivariate regression model and ANN is performed [21]. Jingwei et al. compared the performance of prediction of SWG using an ANN and partial least square–support vector machine (PLS-SVM) [22]. Elmira et al. compared the performance of prediction of SWG using an ANN and Multiple Regression Analysis (MRA) considering types of trucks and their trips, number of personnel in per trips as influencing factors [23].

3 Theoretical Concepts

3.1 Principal Component Analysis

Principal Component Analysis (PCA) is a method of data reduction, which aims to identify a small number of derived variables from a larger number of original variables in order to simplify the subsequent analysis of the data [24]. The sequence of steps needed to be followed in PCA are stated below:

1. Selection of influencing factors which affect the waste generation.
2. Assessment of the suitability of data for the PCA using the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity [24]. It is a measure of how suited your data is for Factor Analysis. This test measures sampling adequacy for each variable in the model. KMO values between *0.5* and *1* indicate the sampling is adequate and values less than *0.5* indicate the sampling is not adequate [24]. Bartlett's test of sphericity checks for the hypothesis that the correlation matrix is an identify matrix, which means that all of the variables are uncorrelated. The score from Bartlett's test of sphericity with significance at *95%* ($p < 0.05$) is considered appropriate for the PCA.
3. Kaiser's criterion or the eigenvalues rule, i.e., only components with eigenvalues of *1.0* or more are retained for further investigation was used to determine Principal Components *PCs*. PCA with Varimax rotation was used to facilitate interpretation of factor loadings L_{ik}. Coefficients C_{ik}, were used to obtain factor scores for selected factors. Factors with Eigen values greater than *1* were used to employ Multiple Regression Model.

3.2 Multiple Linear Regression (MLR) Model

Factor Scores obtained from PCA, were used as independent variables for predicting waste generation [25].

The regression equation is presented as:

$$Y = a + b_1 * FS_1 + b_2 * FS_2 + b_3 * FS_3 + e \quad (1)$$

where Y is the dependent variable (waste generation), a is regression constant (Intercept), b_1, b_2 and b_3 are regression coefficients of Factor Scores FS, e is the error term of the regression model.

3.3 Artificial Neural Network

A neural network is a mathematical representation inspired by the human brain and its ability to adapt on the basis of the inflow of new information [25]. They have the ability to approximate any nonlinear mathematical function, which is useful especially when the relationship between the variables is not known or is complex [26]. The most common type of ANN was tested in this research - the multilayer perceptron (MLP), a feed forward network that can use various algorithms to minimize the objective function [27, 28]. A simplified architecture of a MLP ANN is presented in Fig. 1. The input layer of an ANN consists of n input units with values x_i, $i = 1, 2, ..., n$, and randomly determined initial weights w_i usually from the interval $[-1, 1]$. Each unit in the hidden (middle) layer receives the weighted sum of all x_i values as the input. The output of the hidden layer denoted as Y_c is computed by summing the inputs multiplied with their weights [29], according to Equation [2]:

$$Y_c = f \sum_{i=1}^{n} w_i x_i \quad (2)$$

where f is the activation function selected by the user (sigmoid, tangent hyperbolic, exponential, linear, step or other).

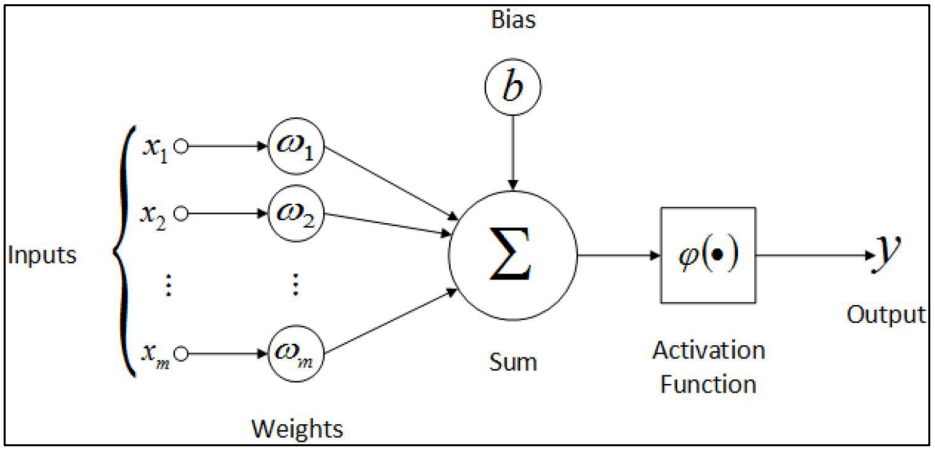

Fig. 1. Structure of Neural Network

3.3.1 Activation Function.

In ANN, the activation function of a node defines the output of that node given an input or set of inputs. There are some common activation functions; Linear Activation Function: Identity Function Non-Linear Activation Functions: Tangent hyperbolic function, Sigmoidal function

3.3.2 Learning by Gradient Descent Optimization Algorithm Error Minimization.

The learning rule of a perceptron in an ANN is that adjusts the network weights w_{mn} in order to minimize the difference between the actual outputs y_{ki} and the target outputs t_{ki}. This difference can be quantified by defining the sum squared error function, summed overall output units i and all training patterns m:

$$\mathrm{E}\,(w_{mn}\,) = \frac{1}{2}\sum_{k=1}^{m}\sum_{i=1}^{n}(t_{ki} - y_{ki})^2 \qquad (3).$$

3.3.3 Hidden Nodes Selection in Hidden Layer.

Deciding the optimum number of neurons in the hidden layer is an important process of building the neural network architecture, since hidden layers influence significantly on the final output. There are some various approaches to find out number of hidden nodes in hidden layer [30].

Try and Error method is characterized by repeated, varied attempts which are continued until success or until the agent stops trying. According to Rule of thumb method, the number of hidden neurons should be in the range between the size of the input layer and the size of the output layer [30, 31].

3.3.4 Early Stopping Approach.

This approach is implemented to avoid the network from overfitting with effective manner [32]. The first subset is implemented as the training set, which is used to initialize weights and biases to the network [33]. The second subset, validation set is used to monitor the error occurring though training procedure [32, 33]. The third subset, test set is not used during training, but used to assess performance [32].

Sum of squared error (*SSE*) and relative error (*RE*) are calculated for both Training and Test sets. *SSE* provides an indication of the root mean square error (*RMSE*) which is a reliable method to measure performance of a neural network [34]. Rule of thumb is to increase the number of hidden neurons if the training error is more. If the training error is satisfactory, but test error is more, it is presumed that the training has led to over-fitting [34]. During training, the validation error generally decreases at the early phase, but when the network stars overfitting, the validation error increases [33].

MLR and ANN models have demonstrated their success in previous literature. Therefore, they are used in this study to develop predictive models and evaluate their performance. However, few researches have addressed waste management and, also in forecasting waste generation. Further, socioeconomic variables have been evaluated so far, but the impact of climatic factors is not assessed in previous studies. Thus, in this study, impact of climatic factors on waste generation is assessed along with socioeconomic, demographic and geographic variables.

4 Methodology

4.1 Data Collection

In this study, daily waste generation data will be collected from fifteen local authorities, that are located at districts of Colombo and Gampaha. Data pertaining to the quantity and composition (sorted degradable waste and sorted non-degradable waste) of daily waste from 2012-2018 will be collected for the study.

Socio-economic and demographic attributes were collected from the Census and Statistics Department. Weather and climatic attributes such as rainfall, humidity and temperature were collected from the Department of Meteorology. Table 1 indicates the selected variables to develop the models.

Table 1. Selected Variables for the Models

Type of indicator	Variables
Waste-related indicator	Total solid waste generated (Biodegradable & Non-Biodegradable waste)
Population indicators	Male and Female population, Total population aged 0-19 years, Total population aged 20 & above
Educational Attainment	Primary**,** Secondary, GCE O/L, GCE A/L, Degree and above, No schooling
Economic indicators	Mean household income, Food expenditure, Nonfood expenditure
Employment status	Unemployed , Employed, Economically not active population
Weather indicators	Rainfall, Temperature, Humidity

4.2 Principal Component Analysis

Initially, there were nineteen input variables. PCA was used as the preliminary step in the development of prediction model. The suitability of data for the PCA was assessed using the Kaiser–Meyer–Olkin (KMO) measure of sampling adequacy and Bartlett's test of sphericity.

Components with eigenvalues of *1.0* or more are retained for further investigation. PCA with Varimax rotation was used to facilitate interpretation of factor loadings L_{ik}. Coefficients C_{ki} were used to obtain factor scores for selected factors. Factors with Eigen values greater than *1* were used to employ Multiple Regression Model.

4.3 Multiple Regression Model

A step wise multiple regression model is developed using the factor scores obtained from PCA.

4.4 Artificial Neural Network

ANN model is constructed with the multilayer perceptron algorithm. Predictor variables consist of Covariates which are scale dependent variables (nineteen numeric variables). Normalized method was chosen for the rescaling of the scale dependent variables to improve the network training. Further, a portion of *60:20:20* of the data is allocated for training, test and validation sets, respectively. Moreover, Gradient Descent Optimization algorithm is used to estimate the synaptic weights. To decide the optimum number of hidden neurons, forward approach and Rule of thumb method is used. Then an optimum number of hidden neurons are selected when *SSE* and *RE* are minimized.

5 Results and Discussion

5.1 Results of PCA and MLR Model

Table 2 shows the results of the KMO and Bartlett's sphericity test. Overall Kaiser's measure of sampling adequacy is equal to *0.780*, indicating that the sample size is adequate to apply the PCA. The significance value of Bartlett's sphericity test is less than 0.05 and it also implied that PCA is applicable to our data set with $P<0.05$.

Table 2. KMO and Bartlett's Test

KMO of Sampling Adequacy.		0.780
Bartlett's Test of Sphericity	Sig.	0.000

According to the results of PCA as shown in Table 3, out of seventeen principal components only four principal components PC_1-PC_4, explaining *84.498%* of variance, were retained.

Table 3. Extraction of PCs

Component	Extraction Sums of Squared Loadings		
	Total	**% of Variance**	**Cumulative %**
PC_1	9.380	52.108	52.108
PC_2	3.181	17.674	69.782
PC_3	1.541	8.563	78.345
PC_4	1.107	6.153	84.498

The principal component scores of selected PC_1-PC_4 shown in Table 4 are used as predictor variables for MLR. According to Table 4, variables Male, Female, Age_0-19, Age_20_and_above, Education_Primary, Education_Secondary, Education_GCE_O/L, Education_GCE_A/L, Education_Degree_and_above, No_schooling and employed population are belong to PC_1. Mean_household_income, Food_expenditure and Non_food_expenditure belong to PC_2, Weather attributes (Rainfall,

Temp_max and Relative_Humidity) belong to PC_3. PC_4 has only one variable, unemployed population.

Table 5 and 6 shows the coefficients and model summary, respectively for the three dependent variables (Biodegradable, Non-Biodegradable and Total waste).

Table 4. Component Score Coefficient Matrix

Dependent Variable	Component			
	1	2	3	4
Male	0.111	-0.014	0.000	-0.014
Female	0.112	-0.015	0.000	-0.017
Age_0-19	0.104	0.003	-0.001	-0.011
Age_20_and_above	0.110	-0.015	0.000	0.004
Education_Primary	0.117	-0.022	0.003	-0.115
Education_Secondary	0.113	-0.025	0.003	-0.117
Education_GCE_O/L	0.100	-0.030	0.003	-0.125
Education_GCE_A/L	0.095	0.002	0.003	0.130
Education_Degree_and_above	0.083	0.017	0.004	0.194
No_schooling	0.118	-0.081	0.006	-0.309
Employed	0.113	-0.016	-0.011	-0.062
Economically_not_active	0.110	-0.017	-0.001	0.003
Mean_household_income	-0.044	0.329	-0.035	0.356
Food_expenditure	-0.029	0.305	0.018	0.027
Non_food_expenditure	-0.022	0.305	-0.008	-0.026
Rainfall	-0.003	0.014	0.489	0.070
Temp_max	0.030	-0.094	-0.440	-0.292
Relative_Humidity	0.031	-0.103	0.465	-0.261
Unemployed	-0.023	0.021	0.083	0.527

Table 5. Coefficients for the MLR Models

Dependent Variable		Unstand-ardized Co-efficients	Sig.
Biodegradable	(Constant)	27815	0.00
	PC1	46214	0.00
	PC2	3876	0.00
	PC3	2545	0.00
	PC4	11508	0.00
Non-Biodegradable	(Constant)	31181	0.00
	PC1	29619	0.00
	PC2	-21390	0.00
	PC3	-1512	0.00
	PC4	9679	0.00
Total	(Constant)	58996	0.00
	PC1	75834	0.00
	PC2	-17514	0.00
	PC3	1033	0.00
	PC4	21188	0.00

According to results of Table 6, it can be seen, that all variables are significant at the *95%* confidence interval. Equations for the dependent variables can be deduced according to Section 2.1 equation [1].

Table 6. Model Summary

Dependent Variable	R	R Square (R^2)	Adjusted R Square
Biodegradable	0.866	0.750	0.750
Non-Biodegradable	0.737	0.544	0.544
Total	0.877	0.769	0.769

$$Y_{Biodegr} = 46214 * PC_1 + 3876 * PC_2 + 2545 * PC_3 + 11508PC_4 + 27815 \quad (4)$$

$$Y_{Non-Biodegr} = 29619 * PC_1 - 21390PC_2 + 1512 * PC_3 + 9676PC_4 + 31181 \quad (5)$$

$$Y_{total} = 75834 * PC_1 - 17514 * PC_2 + 1033 * PC_3 + 21188 * PC_4 + 58996 \quad (6)$$

Fig. 2, 3 and 4 shows the histograms indicating the distribution of residuals for the three dependent variables.

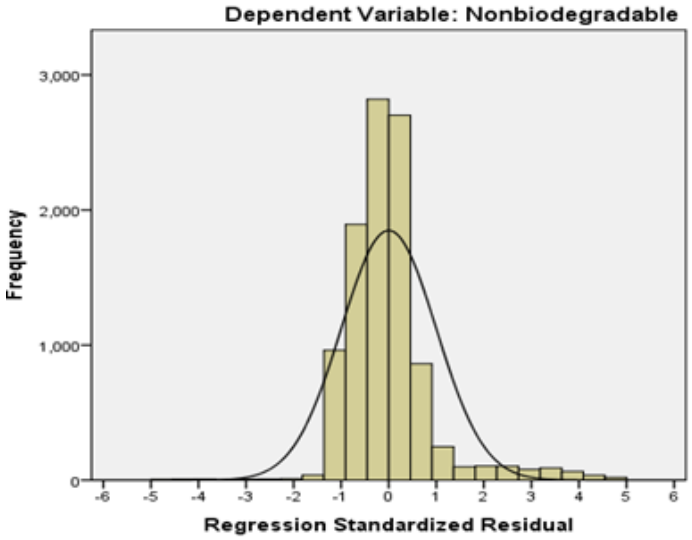

Fig. 2. Histogram of Frequency versus Regression Standardized Residual for Non-Biodegradable Waste

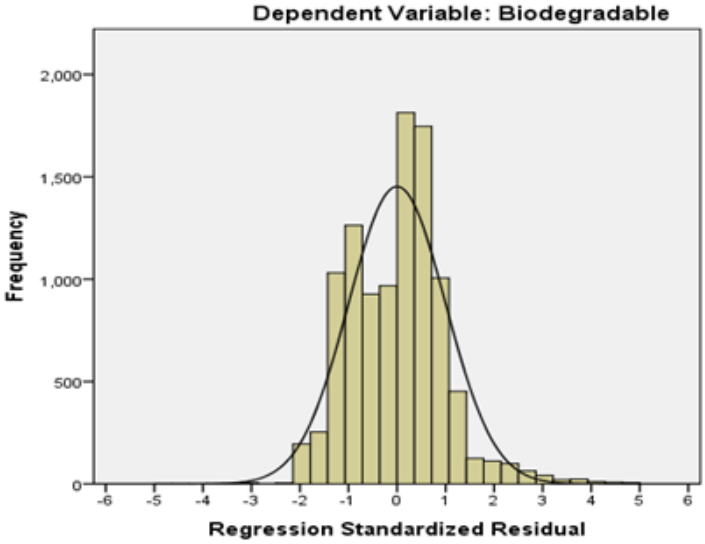

Fig. 3. Histogram of Frequency versus Regression Standardized Residual for Biodegradable Waste

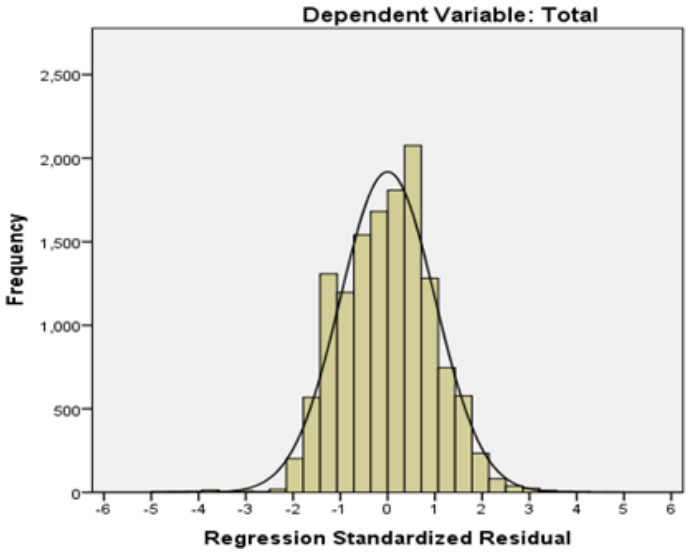

Fig. 4. Histogram of Frequency versus Regression Standardized Residual for Total Waste

5.2 Results for the ANN Model

For all modeling cases, all possible combinations of activation functions between the hidden layer and output layer have been tested and the results are tabulated in Table 7. According to Table 7, the hyperbolic tangent function for the hidden layer and sigmoid function for the output layer shows the least errors.

Table 7. Sum of Squared Error *SSE* for different Combinations of Activation Functions

Activation Functions		**Sum of Squared error**	
Hidden Layer	**Output Layer**	**Training Error**	**Test Error**
Sigmoid	Identity	2354.090	752.044
Hyperbolic tangent	Identity	1825.334	634.947
Sigmoid	Hyperbolic tangent	127.861	47.332
Hyperbolic tangent	Hyperbolic tangent	55.504	16.908
Sigmoid	Sigmoid	39.147	16.233
Hyperbolic tangent	**Sigmoid**	**15.527**	**4.595**

5.2.1 Optimum number of neurons in the Hidden Layer.

After choosing the activation functions, the optimum number of neurons in the hidden layer is found by running the ANN by varying the number of hidden neurons from one to nineteen. Table 8 presents the sum of squared error values for training and test sets and average overall relative error values for training, test and validation sets, while varying the number of hidden neurons.

Fig. 5 and 6 indicates a graphical representation of the same. According to the similarity of the above figures, it is proved that when there are eight hidden neurons, SSE´s and Average Overall Relative Error for training, test and validation sets are minimized. Further, test SSE, starts to go up after eight neurons, possibly indicating overfitting. Therefore, optimum number of hidden neurons for the ANN structure is eight.

Table 8. Sum of Squared Error and Average Overall Relative Error Values for Training and Test Sets

No. of hidden neurons	Sum of Squared Error		Average Overall Relative Error		
	Training Error	Test Error	Training Error	Test Error	Validation Error
1	27.296	9.673	0.259	0.275	0.241
2	17.602	5.528	0.167	0.18	0.171
3	16.202	4.930	0.155	0.147	0.166
4	16.299	4.934	0.164	0.194	0.172
5	25.35	8.406	0.153	0.142	0.154
6	25.093	9.349	0.136	0.145	0.14
7	39.68	16.469	0.15	0.171	0.138
8	**14.369**	**4.300**	**0.142**	**0.147**	**0.132**
9	23.488	7.816	0.142	0.151	0.135
10	21.467	7.719	0.134	0.119	0.129
11	23.86	8.496	0.141	0.141	0.164
12	33.164	13.777	0.151	0.159	0.143
13	16.921	5.223	0.143	0.142	0.152
14	22.834	7.880	0.139	0.144	0.133
15	21.302	7.228	0.143	0.176	0.155
16	15.621	4.794	0.144	0.126	0.136
17	21.62	7.288	0.18	0.17	0.179
18	22.587	8.247	0.135	0.161	0.155
19	30.84	10.692	0.142	0.126	0.128

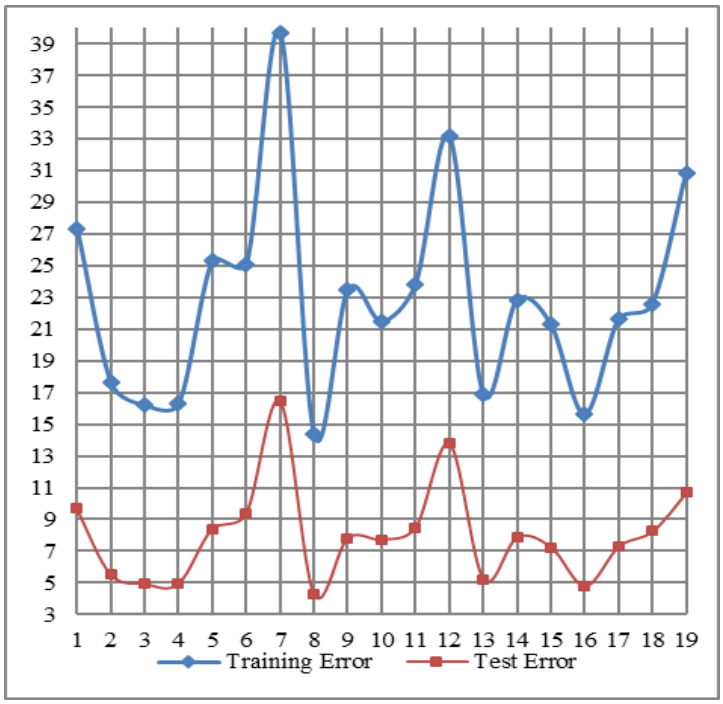

Fig. 5. Sum of Squared Training Error *SSE* versus Number of Neurons in the Hidden Layer

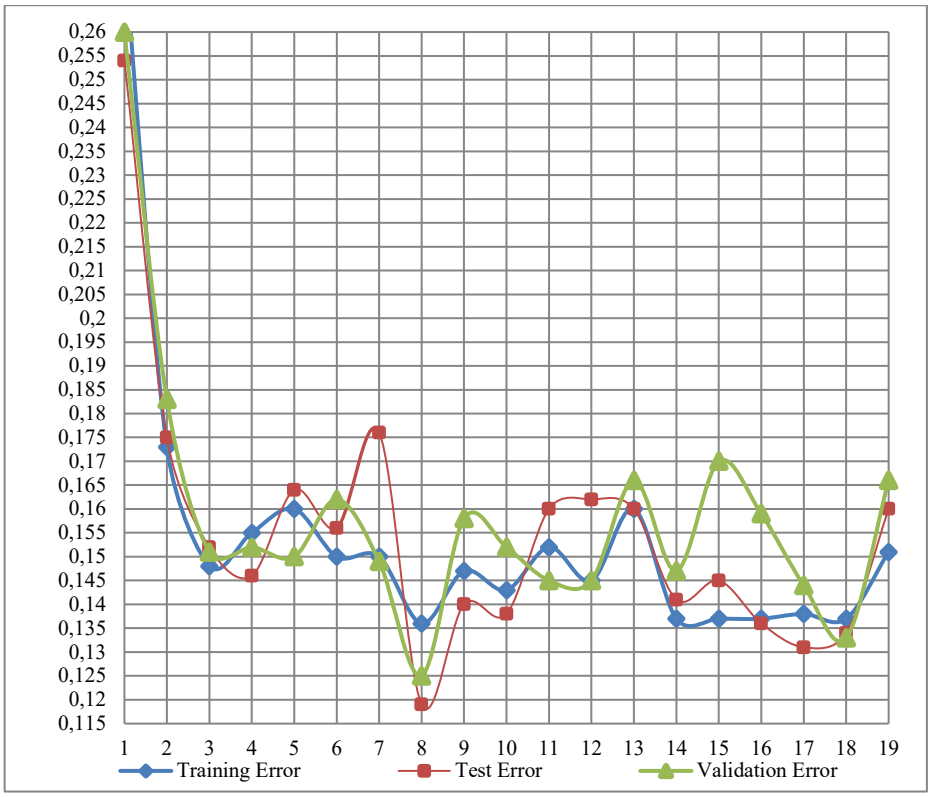

Fig. 6. Average Overall Relative Error of Training, Test and Validation Sets versus Number of Neurons in the Hidden Layer

5.2.1 Features of the selected ANN Structure.

Table 9 presents the network information of the chosen network structure. In the architectural point of view, it is a *19-8-3* neural network (nineteen independent variables, eight neurons in the hidden layer and three dependent variables).

Table 9. Network Information

Input Layer	Covariates	1	Male
		2	Female
		3	Age_0-19
		4	Age_20_and_above
		5	Education_Primary
		6	Education_Secondary
		7	Education_GCE_O/L
		8	Education_GCE_A/L
		9	Education_Degree_and_above
		10	No_schooling
		11	Unemployed
		12	Employed
		13	Economically_not_active
		14	Mean_household_income
		15	Food_expenditure
		16	Non_food_expenditure
		17	Rainfall
		18	Temp_max
		19	Relative_Humidity
	Number of Units[a]		19
	Rescaling Method for Covariates		Normalized
Hidden Layer(s)	Number of Hidden Layers		1
	Number of Units in Hidden Layer 1[a]		8
	Activation Function		Hyperbolic tangent
Output Layer	Dependent Variables	1	Nonbiodegradable
		2	Biodegradable
		3	Total
	Number of Units		3
	Rescaling Method for Scale Dependents		Normalized
	Activation Function		Sigmoid
	Error Function		Sum of Squares
a. Excluding the bias unit			

5.2.3 Summary of the ANN Model.

The model summary in Table 10 shows information about the results of the neural network training. The sum of squared error is shows because the hidden and output layers use the hyperbolic tangent and sigmoid activation functions, respectively. This is the error function that the network tries to minimize during training.

Table 10. Model Summary

Training	Sum of Squares Error		14.369
	Average Overall Relative Error		0.136
	Relative Error for Scale Dependents	Nonbiodegradable	0.160
		Biodegradable	0.149
		Total	0.102
	Stopping Rule Used	1 consecutive step(s) with no decrease in error	
	Training Time		0:00:01.97
Testing	Sum of Squares Error		4.300
	Average Overall Relative Error		0.119
	Relative Error for Scale Dependents	Nonbiodegradable	0.138
		Biodegradable	0.135
		Total	0.087
Holdout	Average Overall Relative Error		0.125
	Relative Error for Scale Dependents	Nonbiodegradable	0.146
		Biodegradable	0.134
		Total	0.097

The scatter plot of the observed against predicted waste for the test data set are given in Fig. 7, 8 and 9 for the Non-Biodegradable, Biodegradable and Total Waste respectively, for the ANN model. The R^2 values obtained are *0.846*, *0.855* and *0.902* for Non-Biodegradable, Biodegradable and Total waste, respectively.

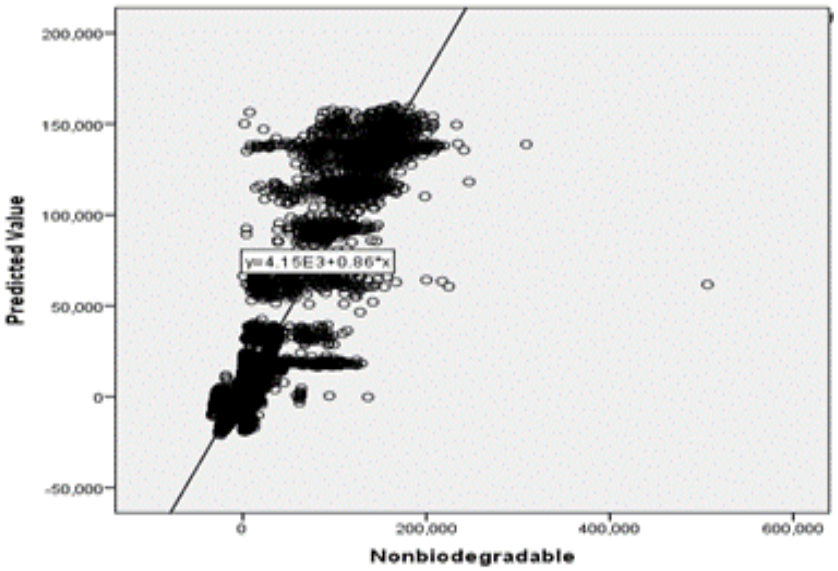

Fig. 7. Graph of Predicted versus Observed Waste for Non-Biodegradable Waste

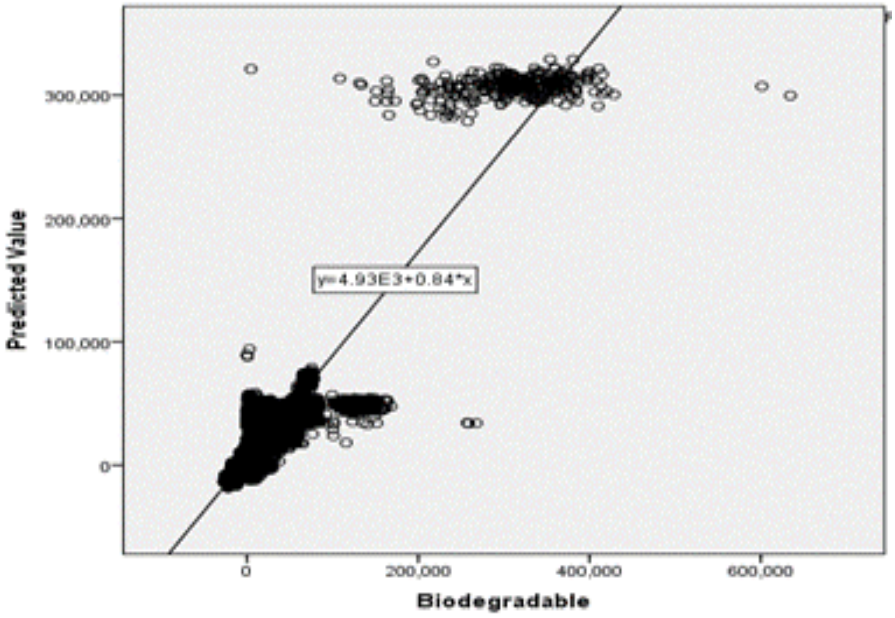

Fig. 8. Graph of Predicted versus Observed Waste for Biodegradable Waste

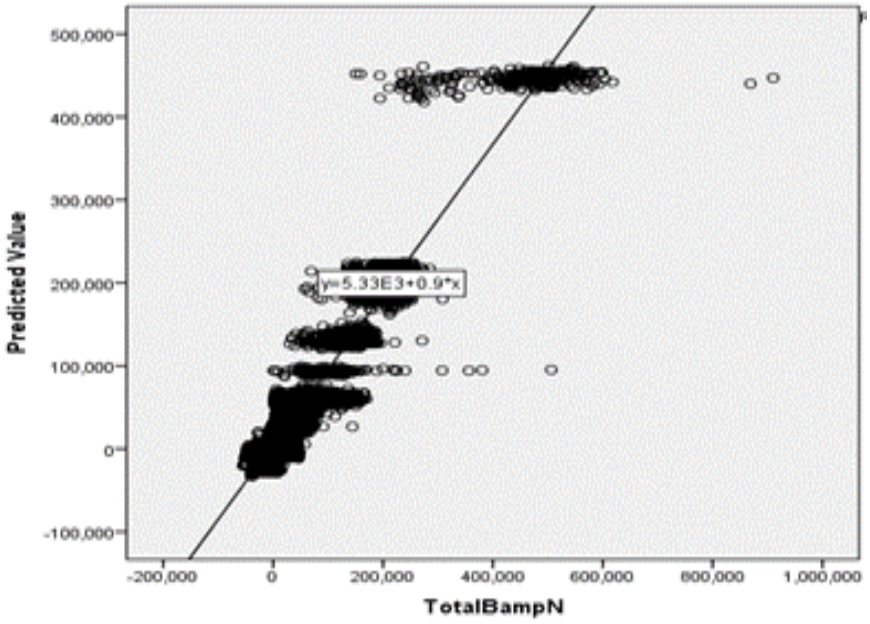

Fig. 9. Graph of Predicted versus Observed Waste for Total Waste

5.2.4 Importance of Independent Variables

Table 11 shows the results of the Variable Important Analysis, which computes the importance and the normalized importance of each variable in determining the neural network. Fig. 10 shows a graphical representation of the normalized importance of predictor variables. The analysis is based on the training and testing samples. The importance of an independent variable is a measure of how much the network's model-predicted value changes for different values of the independent variable. Moreover, the normalized importance is simply the importance values divided by the largest importance values and expressed as percentages. From the following table, it is evident that GCE_A/L population contributes most in the neural network model construction, followed by economically not active and Degree and above population.

Table 11. Independent Variable Importance

Variables	Importance	Normalized Importance
Male	0.059	49.7%
Female	0.044	37.3%
Age_0-19	0.063	53.6%
Age_20_and_above	0.041	34.2%
Education_Primary	0.021	17.6%
Education_Secondary	0.024	20.2%
Education_GCE_O/L	0.070	59.0%
Education_GCE_A/L	0.119	100.0%
Education_Degree_and_above_population	0.113	95.5%
No_schooling	0.028	23.4%
Unemployed	0.031	26.3%
Employed	0.083	70.4%
Economically_not_active	0.114	96.5%
Mean_household_income	0.026	22.1%
Food_expenditure	0.064	54.3%
Non_food_expenditure	0.023	19.5%
Rainfall	0.006	4.8%
Temp_max	0.040	33.9%
Relative_Humidity	0.030	25.4%

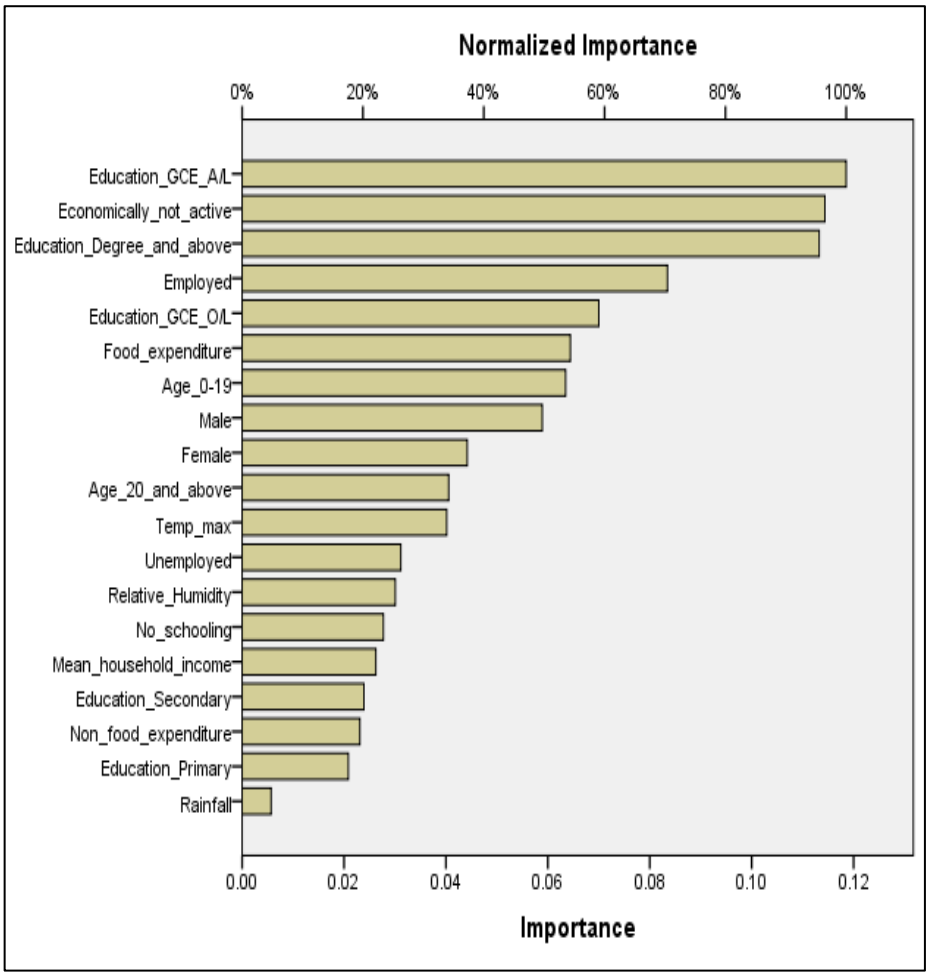

Fig. 10. Normalized Importance

6 Conclusion

This study presents a systematic process to identify the significance of factors affecting waste generation and a methodology for developing solid waste generation models with various socioeconomic, demographic and geographic variables and climatic factors.

PCA was carried out to investigate the influencing factors to waste generation and to avoid the effects of multicollinearity among them. Additionally, the regression relationships for estimating waste generation, based on the selected key factors from the PCA, are developed. The PCA shows four components of key factors that can explain at least up to *84.498%* of the variation of all variables. Then an MLR analysis carried out with the factor scores obtained from PCA which showed R^2 values of *0.750*, *0.544* and *0.769* for Biodegradable, Non-Biodegradable and Total waste, respectively. Neural

Network model is best fitted with R^2 values of *0.846*, *0.855* and *0.902* and lower *RE* values of *0.138*, *0.135* and *0.087* for Non-Biodegradable, Biodegradable and Total waste, respectively. Therefore, ANN model which showed higher predictive accuracy, is concluded as the appropriate model. Further, it is concluded that Education_GCE_A/L population contributes most in the ANN model construction, followed by Economically_not_active and Education_Degree_and_above_population.

Data availability is a limitation of this study, however, despite the limited data, the proposed model reached satisfactory R^2 and lower error values and learnt to model the desired output with a good accuracy. This study provides a reliable method for estimating solid waste generation, providing decision makers, useful information for waste management policy development.

References

1. Bandara, N. J. G. J.: Municipal Solid Waste Management – The Sri Lankan Case. Developments in Forestry and Environment Management in Sri Lanka. (2004)
2. Wijerathna, D.M.C.B., Lee, K., Koide, T., Jinadasa, K.B.S.N., Kawamoto, K., Iijima, S., Herath, G.B.B., Kalpage, C.S., Mangalika, L.: Solid Waste Generation, Characteristics And Management Within The Households In Sri Lankan Urban Areas. (2013)
3. Nayseang, S., Supachart, C.: Development of an appropriate model for forecasting municipal solid waste generation in Bangkok. Elsevier (2017)
4. Hoang M., Fujiwara, T., Pham, S., Nguyen, K.: Predicting waste generation using Bayesian model averaging. Global J. Environment Science Management (2017).
5. Kontokosta, E., Hong, B., Johnson, N., Starobin, D.: Verifying the performance of Artificial Neural Network and Multiple Linear Regression in predicting the seasonal municipal solid waste generation rate: A case study of Fars Province, Iran. Waste Management (2015).
6. Rotchana, I., Salam, P., Kumar, S., Untong, A.: Forecasting of municipal waste quantity in a developed country using multivariate grey models. Waste Management (2015).
7. Yuanwei, W., Yali, X., Jiongyu, Y., Weidou, N.: Prediction of Municipal Solid Waste Generation in China by Multiple Linear Regression Method. International Journal of Computers and Application (2013).
8. Otoniel, B., Gerardo, B., Javier, V.: Forecasting Generation of Urban Solid Waste in Developing Countries-A Case Study in Mexico. Journal of the Air & Waste Management Association (2011).
9. Zaini, S., Simon, G.: The Development of Predictive Model for Waste Generation Rates in Malaysia. Research Journal of Applied Sciences, Engineering and Technology (2013).
10. Sara, O., Gabriela, L., Carolina, A.: Mathematical modeling to predict residential solid waste generation. Elsevier (2008).
11. Mohammad, A., Maliheh, F., Behboudian, S.: Multivariate Econometric Approach for Solid Waste Generation Modeling: A Case Study of Mashhad, Iran. Environmental engineering science (2011).
12. Hoang, M., Fujiwara, T., Pham, S., Nguyen, K.: Predicting waste generation using Bayesian model averaging. Global J. Environ. Sci. Manage (2017).
13. Eduardo, O., Samarasinghe, S., T. Lynn,: A Model for Assessing Waste Generation Factors and Forecasting Waste Generation using Artificial Neural Networks: A Case Study of Chile. Waste and Recycle (2004).

14. Patel, V., Meka, S.: Forecasting of Municipal Solid Waste Generation for Medium Scale Towns Located in the State of Gujarat, India. International Journal of Innovative Research in Science, Engineering and Technology (2013).
15. Kumar, J., Venkata, K., Rao, P.: Prediction of Municipal Solid Waste with RBF Net Work-A Case Study of Eluru, A.P, India. International Journal of Innovation, Management and Technology (2011).
16. Kannangara, M., Dua, R., Ahmadi, L., Bensebba, F.: Modeling and Prediction of regional municipal solid waste generation and diversion in Canada using machine learning approaches. Elsevier (2018).
17. Kontokostaa, C., Honga, B., Nicholas, Johnson E., Starobin, D.: Using machine learning and small area estimation to predict building-level municipal solid waste generation in cities. Elsevier (2018).
18. Elmira, S., Nadi, B., Bin, M., Komoo, I., Hashim, H., YAhya, N.: Forecasting Generation Waste Using Artificial Neural Networks. Waste Management and Pollution Control (2012).
19. Siti, H., Nur, U., Hasmah, M, Shahida, N., Azra, S.: Neural Network Prediction for Efficient Waste Management in Malaysia. Indonesian Journal of Electrical Engineering and Computer Science (2018).
20. David, L., Wani, S.: Forecasting solid waste generation in Juba Town, South Sudan using Artificial Neural Networks (ANNs) and Autoregressive Moving Averages (ARMA). Journal of Environment and Waste Management (2017).
21. Mohammad, A., Falah, M., Salehi, R., Behboudian, S.: Long term Forecasting of Solid Waste Generation by the Artificial Neural Networks. Environmental Progress & Sustainable Energy (2012).
22. Jingwei, S., He, J.: A Multistep Chaotic Model for Municipal Solid Waste Generation Prediction. Environmental Engineering Science (2014).
23. Elmira, S., Bin, M., Abdulai, A.: 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 (2014).
24. Piyawat, W., Mukand, S.: Principal Component and Multiple Regression Analyses for the Estimation of Suspended Sediment Yield in Ungauged Basins of Northern Thailand. Water (2014).
25. Alabi, M., Issa, S., Afolayan, R.: An Application of Artificial Intelligent Neural Network and Discriminant Analyses On Credit Scoring. Mathematical Theory and Modeling (2013).
26. Sarwat, E., Ghada, I.: Estimation of Air Quality Index by Merging Neural Network with Principal Component Analysis. International Journal of Computer Application (2018).
27. Keshavarzi, A., Sarmadian, F.: Comparison of Artificial Neural Network and Multivariate Regression Methods in Prediction of Soil Cation Exchange Capacity. International Journal of Geological and Environmental Engineering (2010).
28. Waghmare, S., Sakhale, C.: Formulation of Experimental Data Based model using SPSS (Linear Regression) for Stirrup Making Operation by Human Powered Flywheel Motor. International Research Journal of Engineering and Technology (2015).
29. Maliki, S., Agbo, A., Maliki, A., Chukwuemeka, L.: Comparison of Regression Model and Artificial Neural Network Model for the prediction of Electrical Power generated in Nigeria. Advances in Applied Science Research (2011).
30. Panchal, F., Panchal, M.: Review on Methods of Selecting Number of Hidden Nodes in Artificial Neural Network. International Journal of Computer Science and Mobile Computing (2014).

31. Thomas, J., Petridis, M., Walters, S.: On predicting the Optimal Number of Hidden Nodes. International Conference on Computational Science and Computational Intelligence (2015).
32. Chandanashive, V., Kambejar, A.: Estimation of Building Construction Cost using Artificial Neural Networks. Journal of Soft Computing in Civil Engineering (2019).
33. Zhang, Z.: Neural Networks: further insights into error function, generalized weights and others (2016).
34. Sonali, A., Akhtar, J.: Neural Nets for Stock Indices: Investigating Effect of Change in Hyper-parameters. Theoretical Economic Letters (2019).