Iranian J Publ Health, Vol.38, No.1, 2009, pp.74-84

Original Article

Comparison of Neural Network and Principal Component-Regression Analysis to Predict the Solid Waste Generation in Tehran

**R Noori, MA Abdoli, M Jalili Ghazizade, R Samieifard*

Dept. of Environmental Engineering, Graduate Faculty of Environment, University of Tehran, Iran

(Received 10 Jun 2008; accepted 20 Dec 2008)

Abstract

Background: Municipal solid waste (MSW) is the natural result of human activities. MSW generation modeling is of prime importance in designing and programming municipal solid waste management system. This study tests the short-term prediction of waste generation by artificial neural network (ANN) and principal component-regression analysis.

Methods: Two forecasting techniques are presented in this paper for prediction of waste generation (WG). One of them, multivariate linear regression (MLR), is based on principal component analysis (PCA). The other technique is ANN model. For ANN, a feed-forward multi-layer perceptron was considered the best choice for this study. However, in this research after removing the problem of multicolinearity of independent variables by PCA, an appropriate model (PCA-MLR) was developed for predicting WG.

Results: Correlation coefficient (R) and average absolute relative error (AARE) in ANN model obtained as equal to 0.837 and 4.4% respectively. In comparison whit PCA-MLR model (R= 0.445, MARE= 6.6%), ANN model has a better results. However, threshold statistic error is done for the both models in the testing stage that the maximum absolute relative error (ARE) for 50% of prediction is 3.7% in ANN model but it is 6.2% for PCA-MLR model. Also we can say that the maximum ARE for 90% of prediction in testing step of ANN model is about 8.6% but it is 10.5% for PCA-MLR model.

Conclusion: The ANN model has better results in comparison with the PCA-MLR model therefore this model is selected for prediction of WG in Tehran.

Keywords: *Prediction of waste generation, Artificial neural network, Multivariable linear regression, Principle component analysis*

Introduction

Nowadays, large quantities of municipal solid waste (MSW) are generated. The waste management organizations must collect, transport, process and finally dispose the residues in an economical and environmental efficient way. Therefore, it is important for MSW managers to obtain accurate predictions of solid waste quantities which are generated (1). The short-term prediction of future MSW generation can facilitate better planning with respect to collection, personnel, truck utilization, transportation to the landfill and final disposal (2). Conventional forecasting methods for MSW generation frequently use the demographic and socioeconomic factors in a per-capita basis. The per-capita coefficients may be taken as fixed over time or they may be projected to change with time. In the other hand, generation quantity is affected by many different factors. Such factors include geographical situation, seasons, collection frequency, onsite process, people's food hobbies, economic condition, recovery and reuse boundaries, existence of law and people's cultural conditions (3). Bruvoll and Ibenholt (4) extended such considerations and forecasted the waste generation (WG) on the basis of a macroeconomic model. McBean and Fortin (5) dealt with certain aspects of MSW management by means of correlations among socioeconomic and solid waste composition and quantities. In recent decades, numerous forecasting techniques have been developed to simulate the environmental process such as generation quantitative, and some of the published forecasting tech-

***Corresponding author:** Tel.: +98 917 9331208, E-mail: roohollahnoori@gmail.com

niques were developed with multiple linear and nonlinear regression techniques, which require the user to specify a priori a mathematical model of the empirical correlation. Multiple linear techniques are mainly used to model the linear relationship between a dependent variable and one or more independent variables. However, when the nonlinear phenomenon is significant to some extent within the data series investigated, the multiple linear will fail to develop an appropriate predictive model. Therefore, nonlinear and dynamic modeling techniques such as artificial neural network (ANN) are necessary for building an accurate and reliable predictive model.

Recently, use of ANN in management of MSW like a proposed model based on ANN, predict rate of leachate flow rate in place of disposal solid wastes in Istanbul, Turkey (6), prediction for energy content of Taiwan MSW using multilayer perceptron neural networks (7), HCl emission characteristics and back propagation neural networks prediction in MSW/coal co-fired fluidized beds (8), recycling strategy and a recyclables assessment model based on an ANN (9) and prediction of heat production from urban solid waste by ANN and multivariate linear regression (MLR) in the city of Nanjing, China (10), have been become in current. The results of these researches have shown the high performance of ANN in prediction of various environmental parameters such as generation.

In this paper we used two methods, ANN and MLR models. We have proposed a new application of principal component analysis (PCA) for using in process of feed data in MLR model for weekly WG prediction in Tehran (PCA-MLR). The goals of present research were prediction of WG with use of ANN and PCA-MLR models and at last selecting an appropriate model for WG prediction in Tehran.

Material and Methods

Case study and data

Tehran is the most populated city in Iran. In the latest years, increasing of emigration has been caused in expanding the WG and as a result making a problem for the municipal solid waste management system (MSWMS) in Tehran. According to the recycling and material conversion organization report, with production of 2.75 million tons MSW in 1384, Tehran was biggest center of WG in Iran. So offering a suitable model for WG forecasting is essential for suitable programming and decision making in organization related to WG. Having seasonal patterns of WG have an effective role for estimating the amount of generated waste in one city, so a weekly time series model of WG with 12 lag time (equal a season) has been made for forecasting the WG in Tehran. In this model weight of waste in t+1 week (W_{t+1}), is a function of waste quantity in t (W_t), t-1 (W_{t-1}) … and t-11 (W_{t-11}) weeks. Besides the time series of generation, another input data (the thirteenth data), consist the number of trucks which carry waste in week of t (Tr_{t+1}). The weekly fluctuation of WG in Mashhad has been shown in Fig. 1.

Artificial neural networks

ANNs customary architecture is composed of three layers; input layer (distributes inputs in the network), hidden layer (processes inputs), and output layer (extracts result in return for inputs). Among all the ANNs paradigms available, a feed forward multilayer perceptron was considered to be the best choice for this study. Feed forward multilayer preceptrons have been shown to have a computational superiority in comparison to other paradigms (11). Feed forward multilayer perceptron can have more than one hidden layer; however theoretical works have shown that a single hidden layer is sufficient for ANNs to approximate any complex nonlinear function (12). Therefore, in our experiment, one hidden layer feedforward multilayer perceptron used for the prediction of weekly MSW generation. The activation functions chosen were the sigmoid hyperbolic tangent function in the hidden and output layers. The error correction learning with the Levenberg-Marquardt (L–M) algorithm (13, 14) was chosen for training the networks. The L–M algorithm has been proved to have the fast-

est convergence on networks which contain up to a few hundreds weights (15). To improve the generalization of the models, the stop training algorithm (STA) approach (16, 17) used. The use of STA reduced the training time four times and it provided better and more reliable generalization performance than the use of L–M algorithm alone. To implement STA in practice, the available data are split into three parts:

1) A training set, used to determine the network parameters, weights and biases;

2) A validation set, used to estimate the network performance and decide when the training stopped;

3) A test set, used to verify the effectiveness of the stopping criterion and to estimate the expected performance in the future.

Principal component analysis

PCA is one of the multivariate statistical methods which can use to reduce input variables complexity when we have a huge volume of information and we want to have a better interpretation of variables (18). By using of this method, input variables change into principal components (PCs) that are independent and linear compounds of input variables (19). Instead of direct use of input variables, we change them into PCs and then we use them as input variables. In this method, the information of input variables will present with minimum losses in PCs (20). PCs specified by the equation in below.

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + ... + a_{ip}X_P \quad [1]$$

In equation [1], Z_i represents specific PCs; a_i is related eigen vector and X_i are also input variables. This information achieved by solving equation [2] (21):

$$|R - I\lambda| = 0 \quad [2]$$

Where, I is unit matrix, R is variance-covariance matrix and λ is eigen value. From these eigen values, we can achieve the eigen vectors.

Details for mastering the art of PCA is published elsewhere (22-27).

Multivariable linear regression

Regression model in matrix form can be shown as:

$$Y = X\beta + e \quad [3]$$

In equation [3], β is regression coefficient matrix, e is fitting error matrix and Y is response matrix. By solving equation [3] for β we will have:

$$\beta = (X'X)^{-1}(X'Y) \quad [4]$$

In equation [4], X' is transpose of X. For calculating inverse of $(X'X)$, it is necessary that the independent variables have not high relativity, because in this situation $(X'X)$ matrix can not become inverse and we will have more error. To solve this problem, we should remove the multicolinearity between independent variables with PCA method. The variance inflated factor (VIF) criterion is usually applied to check the results. The ideal value for VIF is one. The higher the VIF values, the more multicolinearity between independent variables exist.

Models evaluation

In this research, we used two common indices for comparing network output result: correlation coefficient (R) and average absolute relative error (AARE) indices. These criteria show the average of error in model and don't give any information about the error distribution, so to test the robustness of the network output result, it is important to test the model using some other performance evaluation criterion such as threshold statistics (TS) (28-31). The TS not only gives the performance index in terms of weekly predicting WG but also the distribution of the prediction errors. The TS for a level of $x\%$ is a measure of the consistency in forecasting errors from a particular model. The TS represented as TS_x and expressed as a percentage. This criterion can be expressed for different levels of absolute relative error (ARE) from the model. It is computed for the $x\%$ level as:

$$TS_x = \frac{Y_x}{n} \times 100 \quad [5]$$

In equation [5] Y_x is the number of predicted WG (out of n total computed) for which ARE is less than $x\%$ from the model.

Results

Sensitivity analysis

In this research, to know the percentage of every in-

put variables effect on W_{t+1} the sensitivity analysis was performed. Its results are showed in Fig. 2. As can be seen from Fig. 2, the W_{t+1} got the most effect from amount of W_t and W_{t-4}. This can be the result of hobbies and economical situation of people.

Principle component analysis

After standardization of input variables For PCA application, variance-covariance symmetrical matrix *R* was formed from order 13 (equivalent to the number of input variables).

The variance-covariance symmetrical matrix *R*:

$$R = \begin{bmatrix} 1.00 & -0.02 & 0.09 & 0.16 & 0.19 & 0.04 & -0.09 & -0.08 & -0.15 & -0.03 & 0.12 & 0.38 & 0.63 \\ -0.02 & 1.00 & 0.58 & 0.27 & 0.05 & 0.02 & -0.02 & -0.08 & -0.12 & -0.11 & -0.04 & 0.00 & 0.01 \\ 0.09 & 0.58 & 1.00 & 0.57 & 0.23 & 0.04 & -0.02 & -0.07 & -0.12 & -0.13 & -0.10 & -0.01 & 0.03 \\ 0.16 & 0.27 & 0.57 & 1.00 & 0.56 & 0.22 & 0.00 & -0.05 & -0.08 & -0.12 & -0.12 & -0.08 & -0.01 \\ 0.19 & 0.05 & 0.23 & 0.56 & 1.00 & 0.56 & 0.22 & 0.01 & -0.06 & -0.05 & -0.09 & -0.07 & -0.04 \\ 0.04 & 0.02 & 0.04 & 0.22 & 0.56 & 1.00 & 0.55 & 0.24 & 0.02 & -0.01 & -0.03 & -0.09 & -0.10 \\ -0.09 & -0.02 & -0.02 & 0.00 & 0.22 & 0.55 & 1.00 & 0.58 & 0.25 & 0.08 & 0.02 & -0.01 & -0.10 \\ -0.08 & -0.08 & -0.07 & -0.05 & 0.01 & 0.24 & 0.58 & 1.00 & 0.60 & 0.30 & 0.12 & 0.05 & 0.01 \\ -0.15 & -0.12 & -0.12 & -0.08 & -0.06 & 0.02 & 0.25 & 0.60 & 1.00 & 0.62 & 0.33 & 0.15 & 0.03 \\ -0.03 & -0.11 & -0.13 & -0.12 & -0.05 & -0.01 & 0.08 & 0.30 & 0.62 & 1.00 & 0.62 & 0.34 & 0.15 \\ 0.12 & -0.04 & -0.10 & -0.12 & -0.09 & -0.03 & 0.02 & 0.12 & 0.33 & 0.62 & 1.00 & 0.61 & 0.32 \\ 0.38 & 0.00 & -0.01 & -0.08 & -0.07 & -0.09 & -0.01 & 0.05 & 0.15 & 0.34 & 0.61 & 1.00 & 0.61 \\ 0.63 & 0.01 & 0.03 & -0.01 & -0.04 & -0.10 & -0.10 & 0.01 & 0.03 & 0.15 & 0.32 & 0.61 & 1.00 \end{bmatrix}$$

After solving equation [2], 13 eigen values and for every eigen value 13 eigen vector were obtained and by using them, 13 PCs were formed from input variables. The characteristics of these PCs are presented in Fig. 3 and Table 1.

PCA-MLR model

After removing the multicolinearity between independent variables, an appropriate PCA-MLR model was developed for prediction of MSW by stepwise algorithm. The results are showed in Table 2.

Finally, PCA-MLR model constructed for predicting quantity of generated waste that its equation is given below:

$$W_{(t)} = 49168332.3 + 1177698.5 \times (PC3) - 904395.4 \times (PC2) - 823336.1 \times (PC9) \quad [6]$$

The results for training and testing the PCA-MLR model are given in Figs. 4 to 7.

ANN model

To achieve the best network structure for WG prediction, various structures of feed-forward neural networks with three layers and different number of neurons in hidden layer was investigated. Finally, with consideration on R and AARE, a structure with three layers that have 13-22-1 neurons respectively was selected for the best architecture of network. The results for network training and testing are given in Figs. 8 to 11. Finally, TS is applied to assess the error distribution in models in testing stage (Fig. 12).

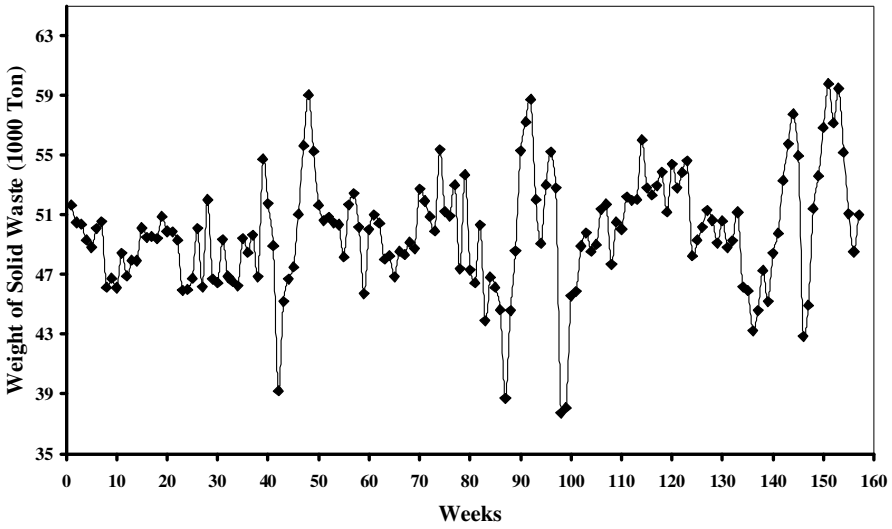

Fig. 1: Variation of waste generation in Tehran

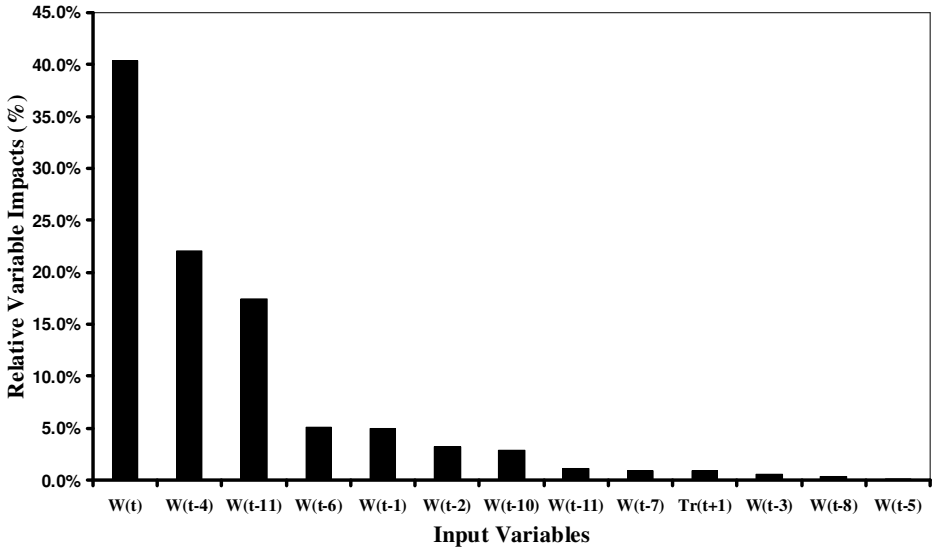

Fig. 2: Percentage of every input variables effect on generated MSW

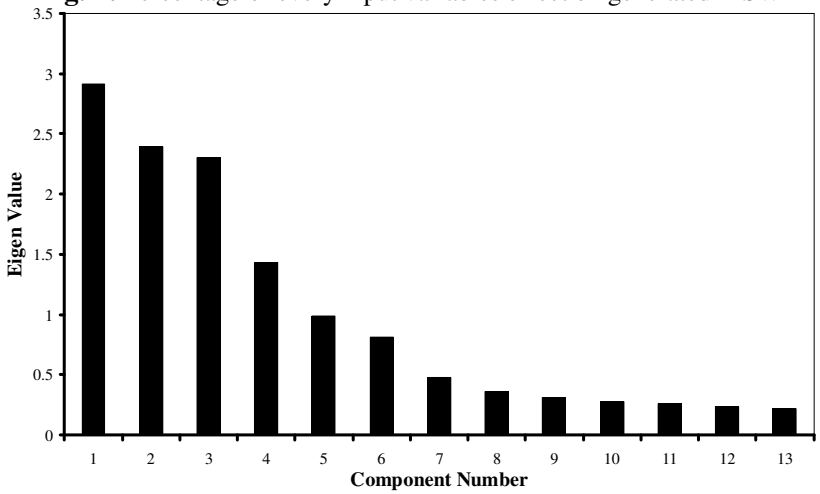

Fig. 3: Specification of each component

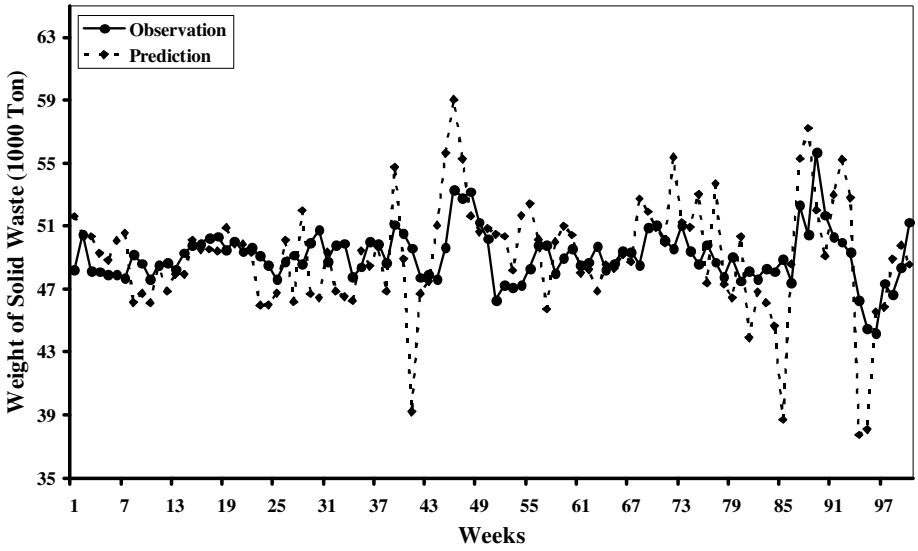

Fig. 4: Observed and predicted solid waste from training of PCA-MLR model

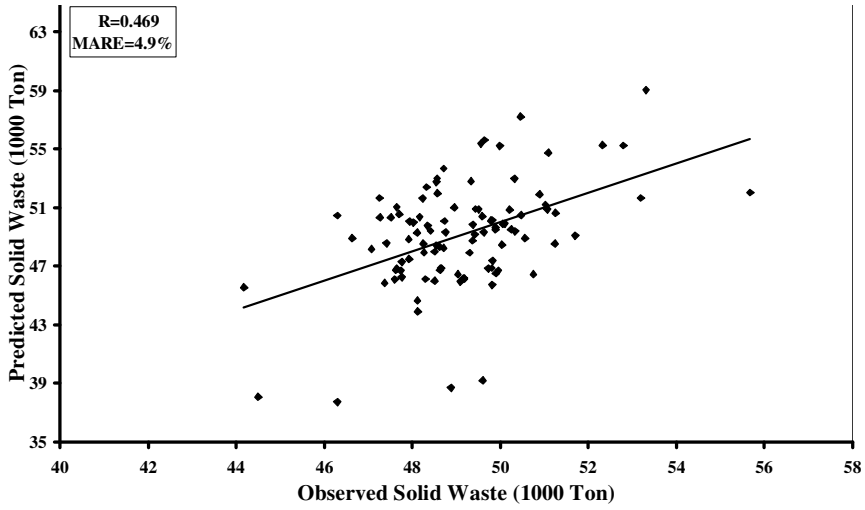

Fig. 5: Scatter plot of observed and predicted solid waste from training of ANN model

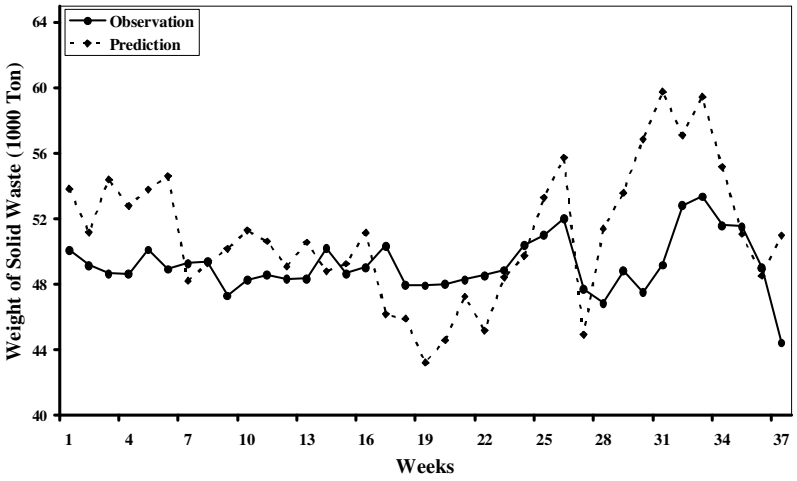

Fig. 6: Observed and predicted solid waste from testing of PCA-MLR model

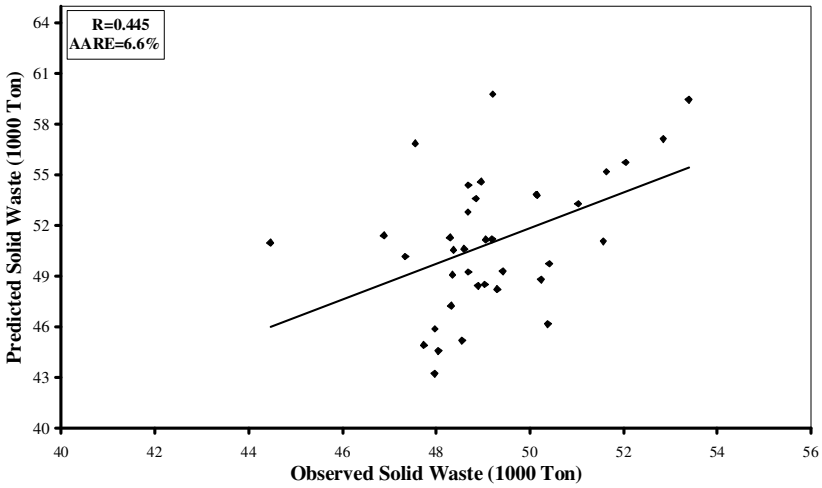

Fig. 7: Scatter plot of observed and predicted solid waste from testing of PCA-MLR model

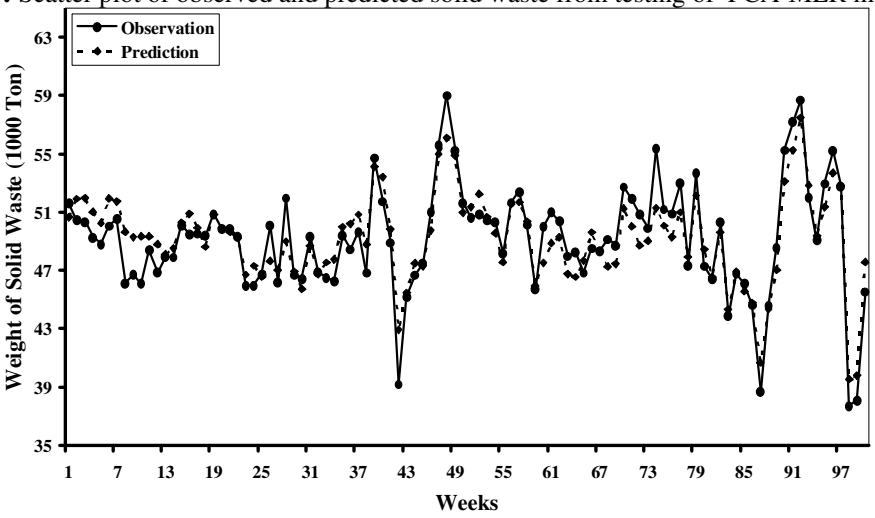

Fig. 8: Observed and predicted solid waste from training of ANN model with structure (13-22-1)

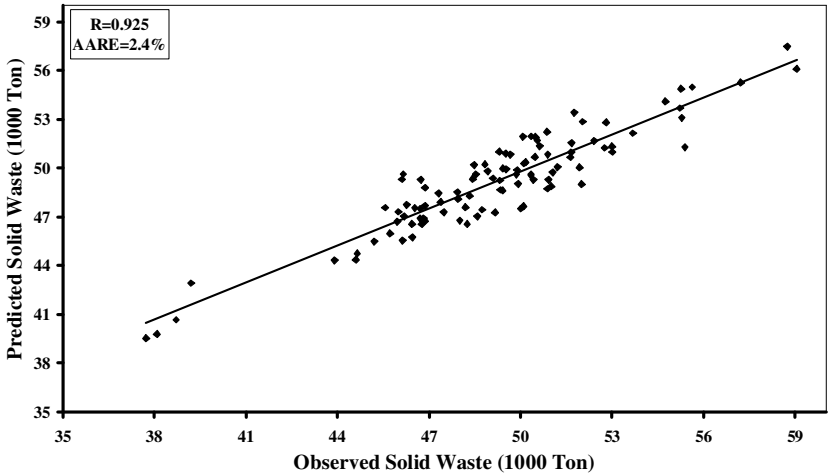

Fig. 9: Scatter plot of observed and predicted solid waste from training of ANN model with structure (13-22-1)

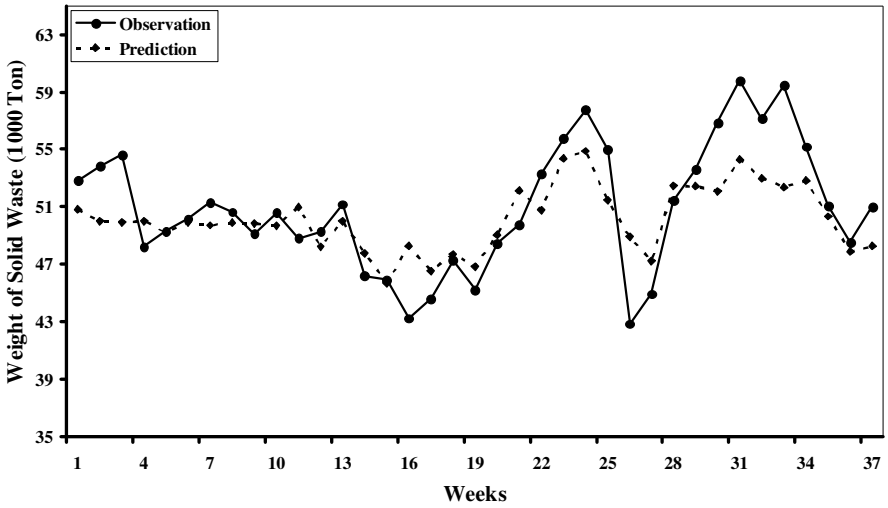

Fig. 10: Observed and predicted solid waste from testing of ANN model with structure (13-22-1)

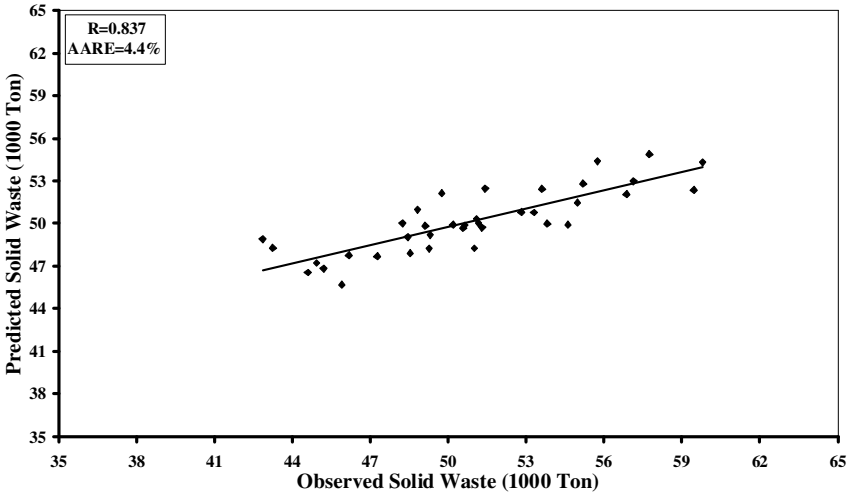

Fig. 11: Scatter plot of observed and predicted solid waste from testing of ANN model with structure (13-22-1)

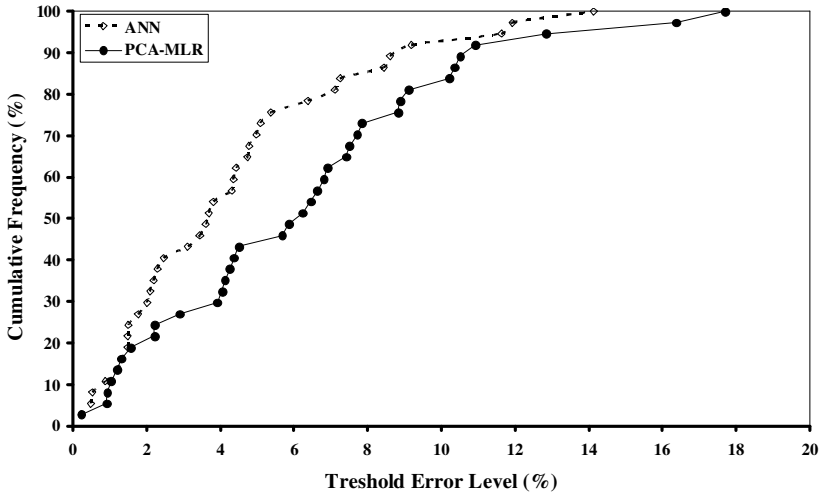

Fig. 12: Distribution of forecast error for PCA-MLR and ANN models in testing step

Table 1: Specification of each PC created

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12	PC13
W_t	0.15	0.06	**0.71**	-0.44	0.15	-0.27	0.21	0.26	0.18	-0.07	-0.16	0.1	-0.01
W_{t-1}	-0.31	0.22	0.35	0.61	0.34	0.3	0.32	-0.15	0.19	-0.01	0.05	0.04	-0.03
W_{t-2}	-0.38	0.38	0.47	0.54	0.13	-0.07	-0.13	0.21	**-0.27**	-0.06	-0.15	-0.12	0.06
W_{t-3}	-0.39	0.55	0.41	0.19	-0.3	-0.27	-0.25	0	0.12	0.17	0.16	0.17	-0.09
W_{t-4}	-0.26	0.69	0.22	-0.28	-0.42	-0.01	0.08	-0.21	0.06	-0.24	0	-0.18	0.09
W_{t-5}	-0.07	**0.75**	-0.11	-0.39	-0.04	0.34	0.18	0.01	-0.17	0.27	-0.13	0.04	-0.08
W_{t-6}	0.19	0.68	-0.34	-0.19	0.39	0.2	-0.16	0.15	0	-0.2	0.18	0.15	0.09
W_{t-7}	0.47	0.52	-0.37	0.08	0.37	-0.3	-0.1	-0.03	0.17	0.1	-0.06	-0.25	-0.13
W_{t-8}	0.66	0.3	-0.31	0.32	-0.08	-0.35	0.14	-0.17	-0.08	0.01	-0.14	0.2	0.19
W_{t-9}	0.76	0.15	-0.04	0.3	-0.36	0.01	0.23	0.19	-0.1	-0.11	0.15	-0.03	-0.22
W_{t-10}	0.74	0.02	0.26	0.17	-0.25	0.36	-0.13	0.15	0.19	0.15	-0.02	-0.09	0.22
W_{t-11}	0.64	-0.04	0.56	-0.04	0.08	0.23	-0.28	-0.24	-0.04	-0.12	-0.16	0.09	-0.18
Tr_{t+1}	0.44	-0.07	0.69	-0.23	0.27	-0.15	0.09	-0.14	-0.2	0.12	0.26	-0.09	0.09

Table 2: result of MLR with application of PCA on primary variables.

Input Component	R	Sig. F Change	VIF
PC3	0.332	0.001	1.066
PC3,PC2	0.412	0.01	1.061
PC3,PC2,PC9	0.469	0.015	1.01

Discussion

As distinguished from Table 2, in PCA-MLR model, by performing PCA from 13 PCs, just 3 PCs were meaningful to enter the model. It estimates the WG with regard to these new input variables. In Table 2, it is obvious that the new obtained PCA-MLR model resulted from PCs have VIF value near one (i.e. an ideal quantity). According to equation [6], PC3 parameter has the most important effective subject on amount of WG of Tehran. The W_t has the most effect On PC3. The other effective parameters on quantity of W_{t+1} are PC2 and PC9 that W_{t-5} and W_{t-2} have the most effect on them, respectively.

We can conclude that ANN model showed better results than PCA-MLR model, in a way that R and AARE have more desirable values in this model. In addition, TS is applied to assess the error distribution in models (Fig. 12). For example, after considering this graph, we can say that the maximum ARE for 50% of prediction is 3.7% in ANN model but it is 6.2% for PCA-MLR model. Also we can say the maximum ARE for 90% of prediction in testing step of ANN model is about 8.6% but it is 10.5% for PCA-MLR model. So the ANN model has better results in comparison with the PCA-MLR model.

Acknowledgements

The authors wish to thank the Civil, Environmental, Laboratory and Consulting Engineering (CELCO) Company and its director, MSc Amir Khakpour, for the financial support of this study.

The authors declare that they have no conflict of interests.

References

1. Chang NB, Lin YT (1997). An analysis of recycling impacts on solid waste generation by time series intervention modeling. *Resour Conserv Recycl*, 19:165-86.
2. Matsuto T, Tanaka N (1993). Data analysis of daily collection tonnage of residential solid waste in Japan. *Waste Manage Res*, 11: 333-43.
3. Tchobanoglous G, Eliaseen R, Theisen H (1977). *Solid Waste: Engineering Principles and Management*. 1st ed. McGraw Hill. Tokyo.

4. Bruvoll A, Ibenholt K (1997). Future waste generation forecast on the basis of a macroeconomic model. *Resour Conserv Recyc*, 19: 137-49.
5. McBean EA, Fortin MHP (1993). A forecast model of refuse tonnage with recapture and uncertainty bounds. *Waste Manage Res*, 11: 373-85.
6. Karaca F, Özkaya B (2006). *NN-LEAP:* A neural network-based model for controlling leachate flow-rate in a municipal solid waste landfill site. *Env Model Softw*, 21: 1190-97.
7. Shu HY, Lu HC, Fan HJ, Chang MC, Chen JC (2006). Prediction for energy content of Taiwan municipal solid waste using multilayer perceptron neural networks. *J Air Waste Mana Asso*, 56: 852-58.
8. Chi Y, Wen JM, Zhang DP, Yan JH, Ni MJ, Cen KF (2005). HCl emission characteristics and BP neural networks prediction in MSW/coal co-fired fluidized beds. *J Env Sci*, 17: 699-704.
9. Liu ZF, Liu XP, Wang SW, Liu GF (2002). Recycling strategy and a recyclability assessment model based on an artificial neural network. *J Mate Proces Tech*, 129: 500-506.
10. Dong C, Jin B, Li D (2003). Predicting the heating value of MSW with a feed forward neural network. *Waste Manag*, 23: 103-106.
11. Hornik K, Stinchcombe M, White H (1989). Multilayer feed forward networks are universal approximators. *Neu Net*, 2: 359-66.
12. Cybenko G (1989). Approximation by superposition of a sigmoidal function. *Math. Control Signals Syst*, 2: 303-14.
13. Marquardt DW (1963). An algorithm for least-squares estimation of nonlinear parameters. *J Soci Indus Appl Math*, 11: 431-41.
14. Levenberg K (1944). A method for the solution of certain non-linear problems in least squares. *Quart J Appl Math*, 2: 164-68.
15. Demuth H, Beale M (1998). *Neural Network Toolbox for use with Matlab*. 2nd ed. The Mathworks. Natick.
16. Sarle WS (1995). Stopped training and other remedies for over fitting. *Proceedings of the 27th Symposium on the Interface of Computing Science and Statistics*, 352-60.
17. Bishop CM (1995). *Neural Network for Pattern Recognition*. 2nd ed. Oxford University Press. New York.
18. Camdevyren H, Demyr N, Kanik A, Keskyn S (2005). Use of principal component scores in multiple linear regression models for prediction of Chlorophyll-a in reservoirs. *Ecol Model*, 181: 581-89.
19. Lu WZ, Wang WJ, Wang XK, Xu ZB, Leung AYT (2003). Using improved neural network to analyze RSP, NOX and NO2 levels in urban air in Mong Kok, Hong Kong. *Env Monit Asses*, 87: 235-54.
20. Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, Fernandez L (2000). Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. *Wat Res*, 34: 807-16.
21. Johnson RA, Wichern DW (1982). *Applied Multivariate Statistical Analysis*. 2nd ed. Prentice-Hall Inc. Englewood Cliffs.
22. Davis JC (1986). *Statistical and Data Analysis in Geology*. 2nd ed. John Wiley & Sons. New York.
23. Manly BFJ (1986), *Multivariate Statistical Methods: A Primer*. 2nd ed. Chapman & Hall. London.
24. Noori R, Ashrafi K, Ajdarpour A (2008). Comparison of ANN and PCA based multivariate linear regression applied to predict the daily average concentration of CO: A case study of Tehran. *J Phy Eart Spac*, 34: 135-52 (Language: Persian).
25. Noori R, Kerachian R, Khodadadi A, Shakibayinia A (2007). Assessment of importance of water quality monitoring stations using principal component and factor analyses: a case study of the Karoon

River. *J Water & Wastewater*, 63: 60-9 (Language: Persian).
26. Tabachnick BG, Fidell LS (2001). *Using Multivariate Statistics*. 3rd ed. Allyn and Bacon. Boston. London.
27. Wackernagel H (1995). *Multivariate Geostatistics*. *An Introduction with Applications*. 2nd ed. Springer. New York and London.
28. Jain A, Indurthy SKVP (2003). Comparative analysis of event based rainfall-runoff modeling techniques-deterministic, statistical, and artificial neural networks. *J Hydro Eng (ASCE)*, 8: 93-98.
29. Jalili M, Noori R (2008). Prediction of municipal solid waste generation by use of artificial neural network: a case study of Mashhad. *Inter J Env Rese*, 2: 22-33.
30. Noori R, Abdoli MA, Ameri A, Jalili-Ghazizade M (2008). Prediction of municipal solid waste generation with combination of support vector machine and principal component analysis: a case study of Mashhad. *Environmental Progress*, DOI 10.1002/ep.
31. Noori R, Abdoli MA, Farokhnia A, Abbasi M (2009). Results uncertainty of solid waste generation forecasting by hybrid of wavelet transform-ANFIS and wavelet transform- neural network. *Expert Systems with Applications*. DOI 10.1016/j. eswa. 2008.12.035.