Application of Artificial Neural Network to Trip Attraction of Condominiums in Metro Manila

Nicanor R. ROXAS, Jr.
Manufacturing Engineering and Management Department
De La Salle University
2401 Taft Avenue, Manila
E-mail: nicanor.roxas@dlsu.edu.ph

Alexis M. FILLONE
Professor
Civil Engineering Department
De La Salle University
2401 Taft Avenue, Manila
E-mail: alexis.fillone@dlsu.edu.ph

Abstract: The Philippines has experienced a significant growth in economy. In light of this, a lot of buildings and condominiums have been constructed all over Metro Manila. Due to the abrupt increase in the number of buildings, the number of jobs, and the population of Metro Manila has increased as well, resulting in the greater number of trips. The original research of Tecson, Sia, and Viray estimated the number of people attracted to the condominiums and identified the important predictor variables that influence the number of people attracted to condominiums by using regression models. In this paper, the ANN is applied to the same dataset and its application is explored. The modelling results are then compared. Lastly, the application of ANN to such problems is verified, and the limitations of the proposed ANN model are determined.

Key words: Trip Attraction, Artificial Neural Network, Linear Regression

1. INTRODUCTION

The Philippines, in the recent years, has experienced a significant growth in economy. In light of this, a lot of buildings and condominiums have been constructed all over Metro Manila, and more are currently underway. These buildings are used for a number of purposes such as residential, commercial, and office locations. Due to the abrupt increase in the number of buildings, the number of jobs, and the population of Metro Manila has increased as well. Along with the increase in population comes the increase in the number of vehicles which is accompanied by the problem of parking spaces. This problem arises from the overflowing demand from tenants, residents, and customers which cannot be served by their parking facilities.

The original research of Tecson, Sia, and Viray aims to estimate the number of people attracted to the condominiums and to identify important predictor variables that influence the number of people attracted to condominiums. All of these were achieved by using the least squares method, which is the common method used in transportation engineering to forecast trip generation (Mathew and Rao; Institute of Transportation Engineers, 1992). However, for the purpose of this research, the application of the artificial neural network (ANN) to predict the number of people attracted to condominiums using the same dataset is explored. The estimated ANN models will use the same dataset as the original researchers in order to facilitate comparison between the results of multiple linear regression and ANN. Several factors that possibly affect the number of people attracted to condominiums all over the Metro are investigated. Properly understanding these factors may help alleviate the congestion through proper traffic management, efficient parking facility design, and allotting enough spaces to cater the growing demand.

2. LITERATURE REVIEW

There have been a significant number of studies which relate trips and vehicular attractions to land uses in the region and most, if not all, of them use regression as the tool for predicting demands. ANN has been used extensively as a predicting tool in civil engineering (Shahin, Maier, and Jaksa; Karlaftis and Vlahogianni, 2011). It is similar to traditional statistical models in that the model parameters are adjusted in order to minimize the error between predicted and observed values engineering (Shahin, Maier, and Jaksa). ANNs have also been shown to outclass traditional statistical methods in engineering (Shahin, Maier, and Jaksa). It has also been discussed in literature that when modelling datasets with complex relationships among the variables, nonlinearities, or even missing data, neural networks are said to be more flexible than traditional statistical tools (Kalyoncuoglu and Tigdemir, 2004; Karlaftis and Vlahogianni, 2011). However, ANN should always be applied with caution since the predicting power is sometimes preferred over the explanatory capabilities of a model (Karlaftis and Vlahogianni, 2011); i.e. causality. In a paper by Kalyoncuoglu and Tigdemir (2004), they used ANN in predicting the amount of risk a driver in Turkey would face given his/her characteristics. Also, in a paper by Cai, Yin, and Xie (2009), they used ANN to hourly pollutant concentrations in China based on traffic-related, background concentration, meteorological, and geographical variables. These are just some of the applications of ANN in transportation engineering, and there are other applications of ANN in travel behavior, traffic flow, and traffic management (Kalyoncuoglu and Tigdemir,2004; Cai, Yin, and Xie, 2009) . In this paper, the ANN is applied to the same dataset of Tecson, Sia, and Viray and the ANN modeling results are compared with the results of the least squares regression. In so doing, the application of ANN to such problems can be verified, and if so, determine the limitations of the proposed model.

3. METHODOLOGY

The dataset from Tecson, Sia, and Viray was derived from counts in randomly selected condominiums in Metro Manila. A total of thirty sample counts of various condominiums were included in the dataset resulting from field data collection and survey. The building administrators and staff were also interviewed for additional information which may prove to be useful in the study. Manual pedestrian counts, from 6:00 am to 6:00 pm, of people attracted to condominiums were performed for the research. From these data points, linear regression models were estimated. The assumptions and violations of the sample points were checked in order to come up with the best model possible. The linear regression model shown in this manuscript is a recreation of the original model derived from the dataset of Tecson, Sia, and Viray. After estimating the linear regression model, ANN models were estimated to facilitate the comparison of results.

In ANN using the back propagation algorithm, all the modeling happens in a *blackbox*. Unlike regression models where one can see the equations formulated, only the network architecture can be seen from ANN. An artificial neural network model is composed of a collection of nodes and links. The nodes act as processing units which receive information from adjacent units, process these information, and send these results to its adjacent nodes. The links indicate the flow of signal through the network. This algorithm is based on minimizing the error of the feed forward neural network output compared to the required output (Cai, Yin, and Xie, 2009; Dohnal) by adjusting the weights. In order to commence with model development, the researcher considers the number of inputs and outputs from which the number of nodes and layers depend. Shown below is a figure which illustrates how the different layers and components of an ANN are arranged. The first layer is the input layer which represents the different inputs to the model. For this study, the inputs used by the student researchers in their final linear regression model would also be used; i.e. the number of commercial establishments inside the condominium, the residential floor area in square meters, the number of building employees, and the average number of persons per unit. Dividing the dataset into three subgroups, training, validating, and test data is done next. After this, the data is normalized to avoid a variable having a dominating effect on other variables in the model. Aside from this, the researcher should also decide on the transfer functions between the input nodes, hidden layers, and the output node/s. After considering all of these, training may commence to generate the ANN. Note that, if one tries to predict something which is *outside* the database, then the ANN will not work since it has not been trained to deal with such cases. So, if one compares the two methods, least squares and ANN, one can easily determine the main difference; the former is very much grounded on assumptions and statistical principles.

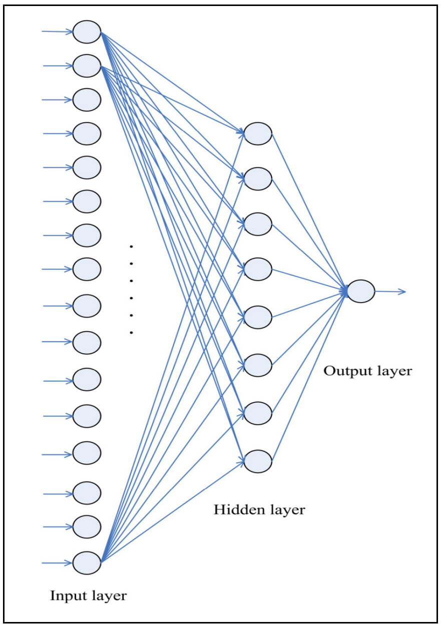

Figure 1. ANN Architecture

4. DATA AND RESULTS

The different dependent and independent variables used in this study are defined as follows.

Y1 = no. of people attracted to the condominium during the peak hour
X1 = the number of commercial establishments inside the condominium
X2 = the commercial floor area of the condominium in square meters
X3 = the residential floor area in square meters
X4 = the total number of floors of the condominium
X5 = unit occupancy rate, in percent
X6 = the number of building employees
X7 = parking occupancy rate, in percent
X8 = the average number of persons per unit
X9 = number of years in operation
X10 = monthly maintenance fee paid by owners of the units per square meter in pesos
X11 = number of entrances/exits

The original linear regression model estimated for the dataset is shown in the following table. It can be seen that the linear regression model offers a modest R square value implying its fair performance in predicting the number of people attracted to condominiums. The independent variables included in the model all show positive coefficients which agree with the expected signs since all these factors are expected to contribute to the increase in the number of people attracted. Also, all the factors are significant except for X8 when the 95% level of confidence is considered. The original researchers probably have enough reasons to include this factor in the model, despite having a p-value greater than 0.05. Overall, the model is acceptable to predict the number of people attracted.

Table 1. Model Replicated from the Previous Study

Regression Statistics	
Multiple R	*0.797925384*
R Square	*0.636684919*
Adjusted R Square	*0.578554506*
Standard Error	*63.55113347*
Observations	*30*

	Coefficients	*t Stat*	*P-value*
Intercept	*-1.371902328*	*-0.025535634*	*0.979830467*
X1	*3.769946331*	*3.906645341*	*0.000629443*
X3	*0.00185471*	*2.176149353*	*0.039200624*
X6	*2.296388599*	*2.422271474*	*0.023004964*
X8	*8.932877818*	*0.790191399*	*0.43684771*

The generated ANN models were estimated by using the same set of predictor variables used in the linear regression model shown in the previous table. It has already been shown that all these independent variables are predictive of the dependent variable, and that these independent variables, *i.e. independent variables in the final model*, are not highly correlated to each other, as one can see in the bivariate correlation, Table 2. This process is done in order to screen the possible significant factors to be incorporated in an ANN model. Correlated inputs somehow induce overfitting and reduce the generalizing capabilities of the model. The following table shows the bivariate Pearson-correlation coefficients for the independent variables included in the final linear regression model of the original researchers. This will be the basis of the ANN models to be generated. By inspecting the values, it can be inferred that no two variables are highly correlated to each other. This is good for ANN modeling since, correlated variables tend to result in less generalization for the model.

Table 2. Correlation Table

Correlations				
	X1	**X3**	**X6**	**X8**
X1	1	0.323865115	0.007833227	-0.063957465
X3	0.323865115	1	0.40383106	0.009479282
X6	0.007833227	0.40383106	1	-0.037908799
X8	-0.063957465	0.009479282	-0.037908799	1

The breakdown of the dataset into training, validating, and testing subgroups adopted for this research is the 60%-20%-20% distribution respectively. This is the default breakdown of the dataset into the three mentioned subgroups used in MATLAB. The training dataset is used for computing the gradient and updating the network weights and biases. Adding the validating subset helps detect over-fitting. If in case the accuracy in the training data set increases, but the accuracy in the validation data set stays the same or decreases, then there is a strong possibility of overfitting and that training should be immediately stopped. Values in the dataset have also been normalized, having values between -1 and 1 using the MAPMINMAX function of MATLAB. There are several other ways by which we may normalize the dataset, but for this case, we adopt what has been suggested by MATLAB.

In order to generate good fitting models, numerous MATLAB trials were performed. The best among the many trials was selected. The type of links which connect the inputs to the different nodes in the hidden layer are also chosen. These arrows represent the manner in which the input values are transformed into values that are to be used by the nodes in the hidden layer. For this study, the *tan-sig* transfer function is used since it is a commonly used transfer function in MATLAB. It is also the default transfer function used by the *newff* library in MATLAB. The *linear function* was used as the transfer function between the hidden layer and the output layer since it is best suited for function fitting problems according to MATLAB. The output layer contains the node which represents the value to be predicted, and for this case, it is the number of people attracted to the condominiums. Note that a single hidden layer is used for this study since the original researchers believed that only simple relationships exist between the dependent and independent variables. Only the number of nodes and not the number of hidden layer was varied for this study.

The algorithm used for training the dataset is the Levenberg-Marquardt algorithm. This algorithm is the fastest algorithm for training small and medium size datasets using feedforward neural networks (Yu and Wilamowski, 2010; Ranganathan, 2004). The learning rate controls the magnitude by which the weights are modified among the nodes during each of the training iterations. The learning rate used is the default value for MATLAB which is 0.01. In order to allow faster training of the dataset, a momentum coefficient is used. The value used is the default value of 0.9.

In order to assess the ANN model, several measures such as the mean squared error (MSE) and R were used. The MSE is chosen as the error metric of choice since it penalizes a model with distant errors and favors a network with none to few distant errors (Twomey and Smith, 1996). These will help determine the overall fit of our model to the dataset. The observed vs predicted plots will be generated in order to visualize how the model performs. The relative importance of each variable in the model is also computed so that it would be possible to determine which variables greatly affect the model. Lastly, the best ANN model generated, *having a single hidden layer*, is compared with the linear regression model generated by the students to determine which model fits the data better. Table 3 summarizes the various parameters assumed and varied to generate the ANN models.

Table 3. Summary of ANN Modeling

Inputs	***x1*** = number of commercial establishments inside the condominium ***x3*** = residential floor area in square meters ***x6*** = number of building employees ***x8*** = average number of persons per unit
Output	number of people attracted to condominiums
Hidden layer	1
Number of nodes in hidden layer	Varied from 3 to 9
Total sample	30
Training dataset	60%
Validating dataset	20%
Testing dataset	20%
Transfer function (input-hidden layer)	tan-sig
Transfer function (hidden layer-output)	linear
Training algorithm	Levenberg-Marquardt algorithm
Learning rate (0.0 – 1.0)	0.01
Momentum factor (0.01 to 0.9)	0.9
Error metric	MSE

The following table shows the different iterations that have been undertaken. There were 7 different models, having 4 inputs with 1 output, generated.

Table 4. Modeling Summary

	4 input variables, 1 hidden layer						
MODEL NAME	**4-3-1**	**4-4-1**	**4-5-1**	**4-6-1**	**4-7-1**	**4-8-1**	**4-9-1**
Number of nodes in hidden layer	3	4	5	6	7	8	9
Transfer Function	tan-sig	tan-sig	tan-sig	tan-sig	tan-sig	tan-sig	tan-sig
Overall R	0.75	0.78	0.78	0.78	0.68	0.76	0.77
MSE	4.9E+03	3.7E+03	4.4E+03	4.6E+03	6.6E+03	4.3E+03	3.7E+03

From the criteria established, the best model would be chosen based on the R value and MSE. Remember that a higher value of R, *close to 1.0*, is desired since it means that the model has a good fit with the data set. A lower value of MSE is also warranted since it denotes the amount of deviation of the predicted values from the observed values. It shows that there are minimal differences between predicted and target values. By close inspection of Table 4, it can be verified that the highest value of R and lowest value of MSE occurred for MODEL 4-4-1. This means that this model is the best model generated *among all the models generated having 4 inputs with a single layer and a single output*. The architecture of MODEL 4-4-1 is shown below.

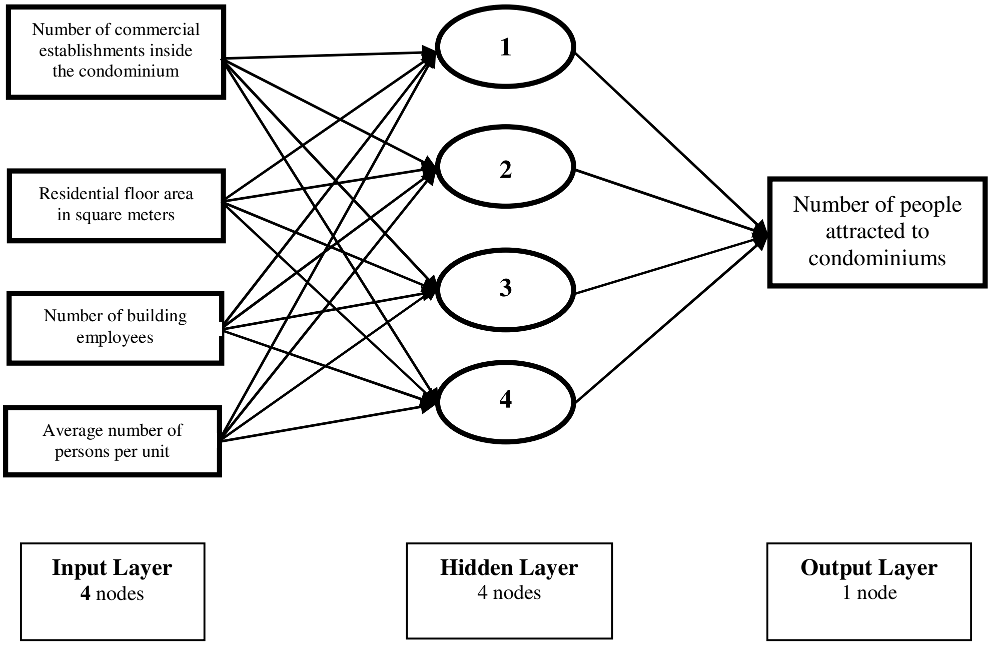
Number of commercial establishments inside the condominium
Residential floor area in square meters
Number of building employees
Average number of persons per unit
1
2
3
4
Number of people attracted to condominiums
Input Layer
4 nodes
Hidden Layer
4 nodes
Output Layer
1 node

Figure 2. MODEL 4-4-1 Architecture

The following figure shows how the model performed, in terms of MSE, while in training, validating, and testing stages. From this figure, it is evident that the MSE decreases as the model is trained, and at 8 epochs, the training was stopped since the MSE of the validation dataset reached its lowest point. This indicates that the model is ready for testing. This figure also shows how the model performed against the testing dataset. From this plot, it can be seen that all the three curves are close to each other, indicating that the model performed consistently among the three datasets.

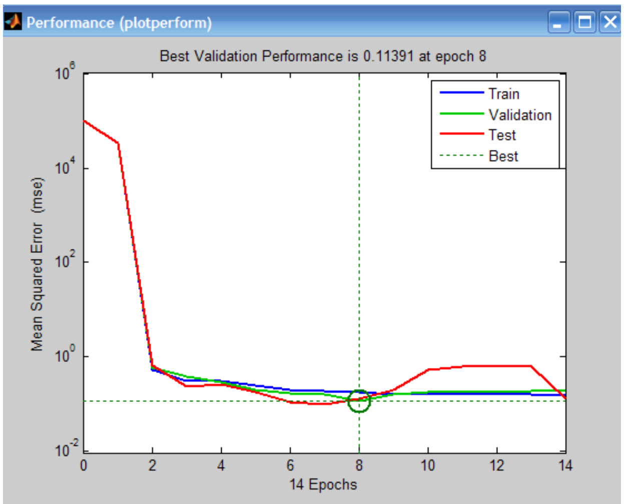

Figure 3. MODEL 4-4-1 Performance

To illustrate this better, Figure 4 shows a plot of the predicted values against the observed values. For a model to have a good fit on the dataset, the following plot should show a clear pattern of values tracing the diagonal. This means that the predicted values fit the observed values accurately. As for the figure below, it is clear that the model performed well, but not exceptionally well, due to an unremarkable R value of 0.78333. A model which perfectly fits the data has an R value of 1.0. For this case, there is a moderately high R value and the plot shows some marked deviations from the diagonal broken line shown.

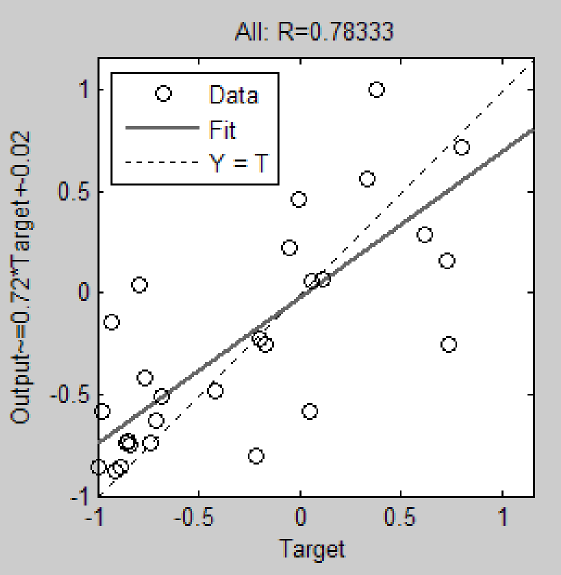

Figure 4. MODEL 4-4-1 Output vs Target Values

By looking at Figure 5, it can be said that at low target values, the model tends to overestimate, while at high target values the model tends to underestimate. From this figure, it can be seen how the model tried to replicate the observed values. Somehow the model was able to trace the values of the targets, though not perfect. The magnitudes of these deviations are further shown in Figure 6.

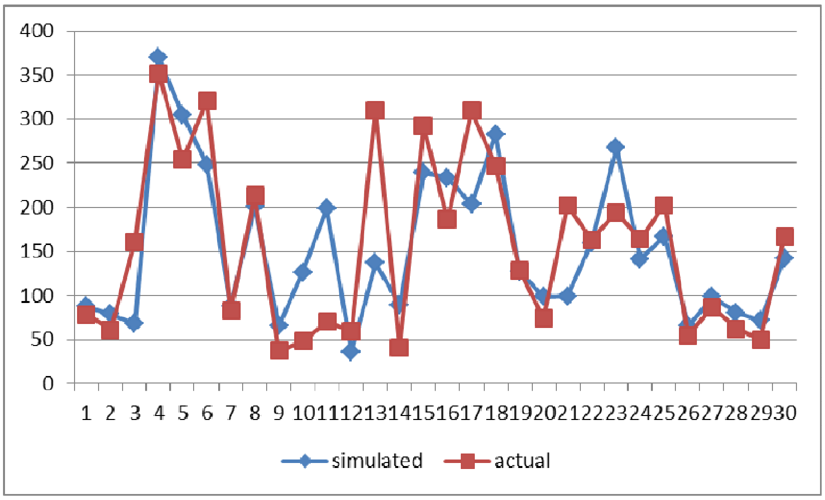

Figure 5. MODEL 4-4-1 Simulated and Actual Values

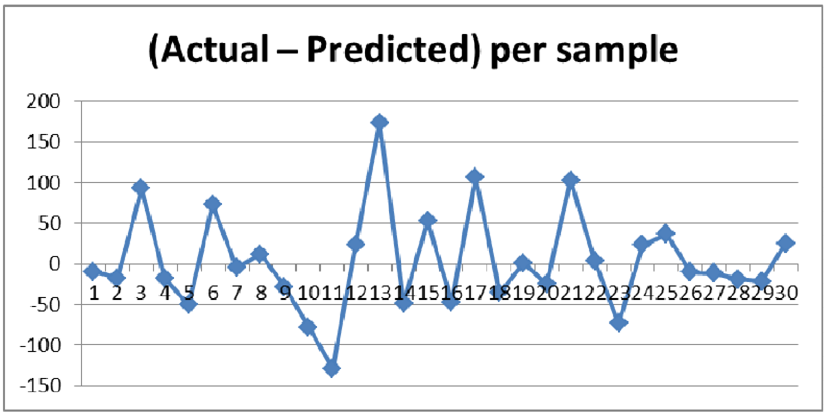

Figure 6. MODEL 4-4-1 Error per sample

Now that MODEL 4-4-1 has been chosen as the best model, the individual input factors are examined according to their individual relative importance in the model. The relative importance of each of the input factors indicates the relative contribution of the input variables in predicting the output (Paliwal and Kumar, 2011). From Figure 7, it can be verified that the number of building employees is the most important input factor followed by the residential floor area, number of commercial establishments, and the average number of persons per unit. This is somewhat similar to the result of linear regression analysis shown in Table 1 except for *X1*, the number of commercial establishment in the condominium. From the results of the linear regression analysis, it can be seen that the order of importance is X1, X6, X3, and X8 respectively.

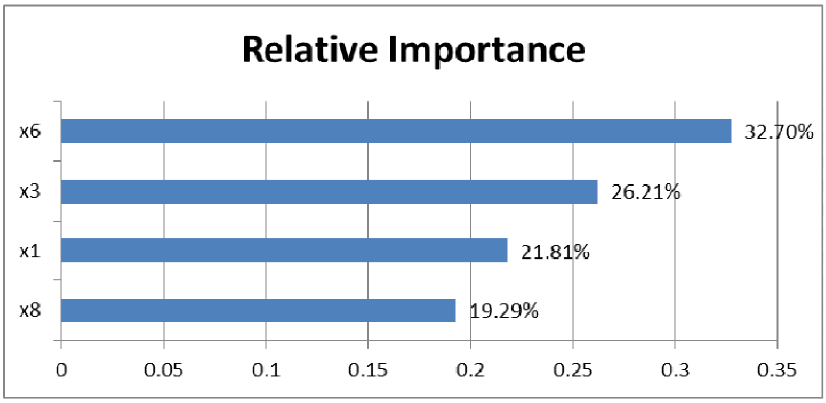

Figure 7. MODEL 4-4-1 Relative Importance of Inputs

The results of the linear regression model may now be compared with the ANN model. The linear regression model has a higher R value of 0.87 when compared to the 0.78 of ANN. This somehow suggests that the regression model explains a larger amount of variability in the dataset than the ANN model. However, it should be noted that the ANN model generated here is not really the *best* model one can possible generate. If there were more training pairs, then the results could have been different.

An interesting aspect of ANN is that sensitivity analysis may be performed for the dataset. It can show how the output, the number of people attracted to condominiums, changes when the value of an input variable changes, while leaving all the other variables unchanged at their mean[21]. This type of analysis is very helpful since it gives us the impact on the output of a particular change in an input variable. Important input factors will yield significant changes in the output when the inputs are varied (NeuroSolutions, 2000). Therefore in order to generate a good model, the input factors with the largest sensitivity values should be retained in the network. Conversely, inputs that that produce no change should be eliminated since it only produces longer training times and possible poor generalizations. Figure 8 shows the sensitivity analysis that was generated by the NeuroSolutions for Excel Software while the magnitudes are shown in Table 5. It shows that the output is most sensitive to X6, then X3, X1, and X8. A change of 1 building employee creates a corresponding 41.54 change in the number of persons attracted to the condominiums. The same can be extended to the other input factors as seen in Table 5. This result complements the result in relative importance for the ANN model, see Figure 7.

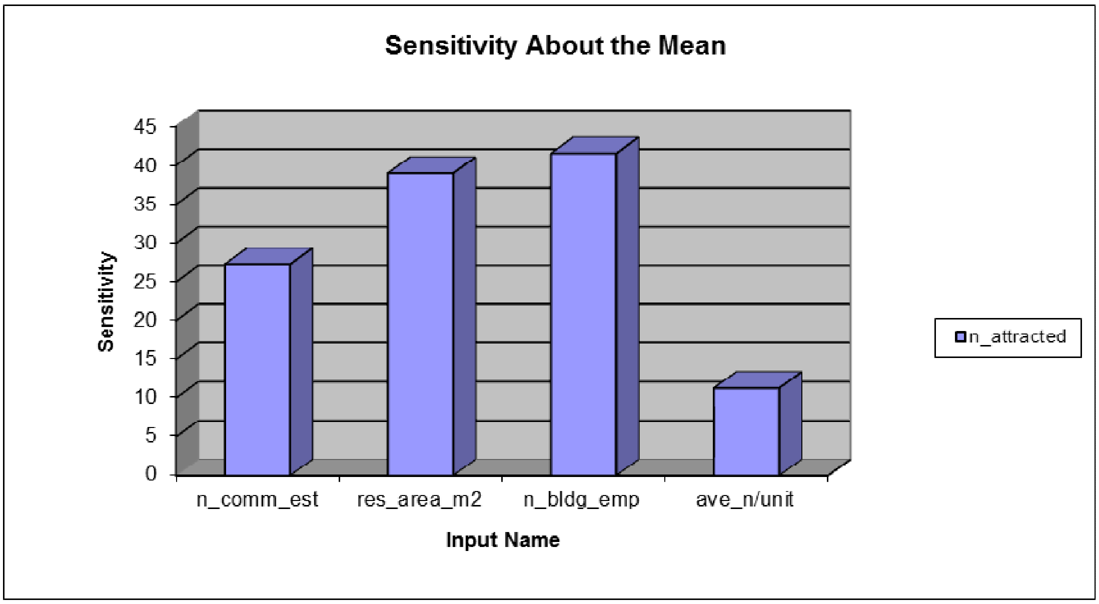

Figure 8. Sensitivity Analysis (*NeuroSolutions*)

Table 5. Sensitivity Analysis Values (*NeuroSolutions*)

Sensitivity	***Code***	***n_attracted***
n_comm_est	X1	27.21895126
res_area_m2	X3	38.97346513
n_bldg_emp	X6	41.54752837
ave_n/unit	X8	11.28883283

4. CONCLUSION

This paper aimed to estimate the number of people attracted to the condominiums and identify important predictor variables that influence the number of people attracted to condominiums. These were done by performing both linear regression and ANN to the dataset which contains 30 samples from Tecson, Sia, and Viray. From the results of linear regression and ANN modelling, it can be said that all the four independent variables included, *number of commercial establishments inside the condominium, residential floor area in square meters, number of building employees and the average number of persons per unit*, are significant to the different models generated. The R-values indicate that these four predictors have substantially accounted for the variation in the dependent variable, for both the ANN and the linear regression models. The ANN model did well in simulating the data with an R value of 0.78 but the linear regression model performed better, having an R value of 0.87. However, it should be noted that the best ANN model developed, ANN 4-4-1, lacks training data which could result in a less satisfactory model. It should also be stated that complex networks, when applied to a small dataset could possibly result in overfitting. The application of the ANN method has proven that it is a feasible technique in estimating the number of people attracted to condominiums. However, this result should be interpreted with caution since the ANN model has limitations in terms of the data it may take. The ANN model also lacks one of the strong points of linear regression; i.e. causality. Nevertheless, this paper was able to illustrate how to generate and analyze ANN models along with its counterpart, multiple linear regression. In order to generate a better model for ANN, it is suggested that a larger database is needed so that better generalizations may be made, therefore increasing R value resulting in a better fit and less error. Aside from these, there are other ways to cope with small datasets as offered by MATLAB which may be explored for future studies.

REFERENCES

Cai , M., Yin , Y., and Xie, M., Prediction of hourly air pollutant concentrations near urban arterials using artificial neural network approach, **Transportation Research Part D**, pp. 32–41, 2009

Dohnal, J., **Using of Levenberg-Marquadt Method in Identification by Neural Networks**

Fan, H. S. and Lam S. H., *Parking Generation of Commercial Developments in Singapore*, **Journal of Transportation Engineering**, May 1997, pp.238-242.

Kalyoncuoglu, S. F. and Tigdemir, M., *An alternative approach for modelling and simulation of traffic data: artificial neural networks*, **Simulation Modelling Practice and Theory 12**, pp. 351–362, 2004

Karlaftis , M.G. and Vlahogianni , E.I., Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights, **Transportation Research Part C**, pp.387–399, 2011.

Mathew, T. V.and Krishna Rao, K. V., **Introduction to Transportation Engineering**

Nathans, L., Oswald, F., and Nimon K., Interpreting Multiple Linear Regression: A Guidebook of Variable Importance, Practical Assessment Research and Evaluatin, Volume 17, Number 9, April 2012 ISSN 1531-7714

NeuroSolutions Newsletter, June 2000.

Orquina, C. A., *A Study on Parking Occupancy of Residential Condominiums in Metro Manila*, UP-NCTS Thesis, July 2001.

Paliwal, M and Kumar, A., Assessing the contribution of variables in feed forward neural network, **Elsevier** 2011.

Pyndick, R. and Rubinfeld, D.**Econometric Models and Economic Forecasts** 1997

Ranganathan , A., **The Levenberg-Marquardt Algorithm**, 2004

Shahin, M. A., Maier, H. R., and Jaksa, M. B., Investigation into the Robustness of Artificial Neural Networks for a Case Study in Civil Engineering

SPSS User's Guide.

Tecson, M.R., Sia, R., and Viray, P., Application of Artificial Neural Network Using Back Propagation Network To Trip Attraction of Condominiums in Metro Manila, **DLSU**, 2003

Twomey, J.M. and Smith, A., Chapter 4 in ***Artificial Neural Networks for Civil Engineers: Fundamentals and Applications**,* ASCE Press 1996.

Washington, S.P., Karlaftis, M.G., and Mannering, F.L. ***Statistical and Econometric Methods for Transportation Data Analysis***. Chapman & Hall/CRC,Washington D.C., 2003.

Yam, Richard C., M., Whitfield, R., C., and Chung, R., W., F., *Forecasting Traffic Generation in Public Housing Estates*, **Journal of Transportation Engineering**, July 2000, pp. 358-361.

Yu, H., Wilamowski, B. M., **Levenberg-Marquardt Training**, Auburn University, 2010