Interdependence between Structural Engineering and Construction Management
Edited by Ozevin, D., Ataei, H., Modares, M., Gurgun, A., Yazdani, S., and Singh, A.

ISBN: 978-0-9960437-6-2

ESTIMATING THE RUNNING COSTS OF COMMERCIAL BUILDINGS: ARTIFICIAL NEURAL NETWORK MODELING

DEVINDI GEEKIYANAGE and THANUJA RAMACHANDRA

Dept of Building Economics, University of Moratuwa, Moratuwa, Sri Lanka

Running costs of a building is a substantial share of its total life-cycle cost (LCC) and it ranges between 70-80% in commercial buildings. Despite its significant contribution to LCC, investors and construction industry practitioners tend to mostly rely on construction cost exclusively. Though the early stage estimation of running costs is limited due to the unavailability of historical cost data, several efforts have been taken to estimate the running costs of buildings using different cost estimation techniques. However, the prediction accuracy of those models is still challenged due to less quality and amount of data employed. This study, therefore, developed an artificial neural network (ANN) model for running costs estimation of commercial buildings with the use of building design variables. The study was quantitively approached and running costs data together with 13 building design variables were collected from 35 commercial buildings. The ANN model developed resulted in a 96.6% perfect correlation between the running cost and building design variables. The testing and validation of the model developed indicate that there is greater prediction accuracy. These findings will enable industry practitioners to make informed cost decisions on implications of running costs in commercial buildings at its early stages, eliminating excessive costs to be incurred during the operational phase.

Keywords: Cost modeling, Operations cost, Maintenance cost, Building design variables, Decision-making, LCC.

1 INTRODUCTION

Usually, costs incurred during the operational phase of a building responsible for a substantial share of its Life Cycle Cost (LCC). Some buildings have inherently higher running cost than others, such as commercial buildings. For example, the running costs of commercial buildings account for over 69% of the total LCC (Wang *et al.* 2014). Similarly, Wong *et al.* (2010) revealed that the running cost of an office building varies between 72 to 81% of its total LCC. Despite its contribution to the LCC structure, often running cost is given less focus in investment decision making and investors tend to mostly rely on initial cost alone.

A recent study on the review of existing models for LCC estimation revealed that there is no simple model for estimating the running cost of buildings to date (Krstić and Marenjak 2017). The application of available methods and models for the running cost estimation of buildings are also limited to the later stage of building life cycle as these models require an extensive set of operational cost data (Krstić and Marenjak 2017). For example, Al-Hajj and Horner (1998) have presented a running costs model for institutional buildings, with eleven cost elements and to an accuracy of 1.13%. Similarly, Kirkham *et al.* (2002) and El-Haram *et al.* (2002) have developed

WLCC models for hospital buildings where cost components such as facilities management costs, energy costs, maintenance costs, residual costs, and discount rate were determinants of WLCC. Early-stage supportive running cost estimation models are therefore essential as it provides implications of costs to be incurred during the operating phase of buildings at early design stages of building constructions.

Estimation of cost of a product, system, or service based on its determinants is a well-known and approved method for cost estimation over the years. For example, Kirkham *et al.* (1999) have developed an energy cost model for sports centres based on building design variables such as the number of users and floor area. However, Krstić and Marenjak (2017) stressed that these models are not based on adequate historical cost records and not based on the available cost structure, rather than standard cost structure. Authors further indicate that the models developed so far ignore some important factors such as the age, location, level of occupancy, and standards of operation.

Deciding through which type of building to include in a forecasting model is not the only problem. The choice of modelling technique is also important (Boussabaine *et al.* 1999). Among the statistical approaches, regression techniques deserve attention due to relative ease of implementing and requirement of less computational power than other statistical approaches (i.e. genetic algorithms, neural networks, support vectors machine) (Fumo and Biswas 2015). However, the application of purely parametric cost estimation methods is limited due to the lack of reliable historical cost data and building design variables, which have a direct influence on its LCC. In contrary, Boussabaine *et al.* (1999) opined that statistical models have been used for some time but in present, artificial intelligence is proposed as a more reliable and accurate modelling technique. Providing professionals with accurate forecasting techniques will enable them to make informed and reliable estimates of likely running cost in commercial buildings, as well as other forms of buildings. Therefore, this study introduces an early-stage supportive running cost estimation model for commercial buildings with use of the artificial neural network (ANN) modelling.

2 RESEARCH METHODS

The research was primarily approached quantitatively to develop early-stage supportive running cost estimation models for commercial buildings with the use of ANN modelling. The documents including architectural drawings, bills of quantities, historical cost records, and monthly utility bills were reviewed to collect the required data. The case buildings selected for the study was limited to 35 out of the population of 117 commercial buildings, which were recorded in Sri Lanka due to the time constraints and limited access to cost data. Generally, a sample size of more than 30 at 5% confidence level is sufficient for many types of research. Though it is said that a big sample of data is required to run an ANN, an ANN tool including a particular training-validation-test procedure for small datasets has been developed some years ago and recently refined in order to obtain not only realistic regression laws, but also reliable ones (One can refer to Pasini and Potestà (1995) and Pasini *et al.* (2001) for the fundamentals of this tool) (Pasini 2015). Accordingly, the commercial buildings selected for the study consists of 49% of office buildings and 37% of banks while remaining include educational institutes, retails, and multi-purpose (i.e. hotel + apartment) buildings. Further, a majority of the selected buildings (63%) consists of three to 12 while remaining 26% and 11% are 13 to 25 and above 25 storied buildings respectively.

Based upon statistical pre-analysis, 13 variables (i.e. building design variables), which are quantitative in nature and convertible (nominal data) were selected for predicting the running cost

of commercial buildings. The influence of variables on the running cost and ease of availability of data were the primary factors in the selection of the variables. Further, the running cost data were collected in accordance with the standards of BCIS, BS ISO 15686-5:2008 standard, and NRM3, for three consecutive financial years: 2014, 2015, and 2016.

Initially, the collected dataset was subjected to the 'Multiple imputation" technique to impute the missing values within the data set. Next, the target variable was normalized using the gross internal floor area and obtain the normalized target variable called running cost/sq. ft. The ANN model was developed with the aid of Neural Designer machine learning software and the Feedforward neural network with backpropagation training was administered as it is commonly used with linear activation function. Finally, the prediction accuracy of the developed model was evaluated with use of the mean absolute percentage error (MAPE) and Theil's U value.

3 DATA ANALYSIS AND FINDINGS

In order to proceed with the neural network analysis, there are three basic assumptions to be satisfied. Firstly, both dependent and independent variable s should be the continuous form of data. In this study, the dependent variable, which is running costs/sq. ft and independent variables including working days/week, working hours/day, building age, GIFA, net floor area, circulation area, height, number of floors, window area, Window-to-Floor-Ratio, and number of occupants are scale data. In addition, two dummy variables namely, the grouping of buildings (1=Detached, 2=Attached), and type of structure (1=Concrete, 2=Steel, 3=Pre-fabricated) were added to the analysis to represent the nominal data collected. Therefore, satisfied the first assumption. Next, the Shapiro-Wilk normality test was conducted to explore the normal distribution of residual values. As observed from Table 1, the significance of the standardized residual (ZRESI) is greater than 0.5 indicates that the ZRESI is normally distributed.

Table 1. Test of normality: Shapiro-Wilk.

	Statistic	**df**	**Sig.**
Standardized Residual	0.954	30	0.211

Next, the relationship between the dependent variable and the independent variables are needed to be linear, both for each independent variable and globally. Accordingly, a scatterplot analysis was conducted between each independent variable and the dependent variable and the charts derived are presented in Figure 1. As shown in the scatterplot matrix, five continuous independent variables namely GIFA, NFA, CA, building height, and the number of floors out of 11 have strong linear relationships with the dependent variable: running cost. Although other six independent variables don't represent strong linear relationships with the dependent variable as the points are more scattered and it is observed that the points are trying to gather along the diagonal. Therefore, it is concluded that all the independent variables have linear relationships with the dependent variable, thus satisfied the third assumption. As shown in the scatterplot matrix, five continuous independent variables namely GIFA, NFA, CA, building height, and the number of floors out of 11 have strong linear relationships with the dependent variable: running cost. Although other six independent variables don't represent strong linear relationships with the dependent variable as the points are more scattered and it is observed that the points are trying to gather along the diagonal. Therefore, it is concluded that all the independent variables have linear relationships with the dependent variable, thus satisfied the third assumption.

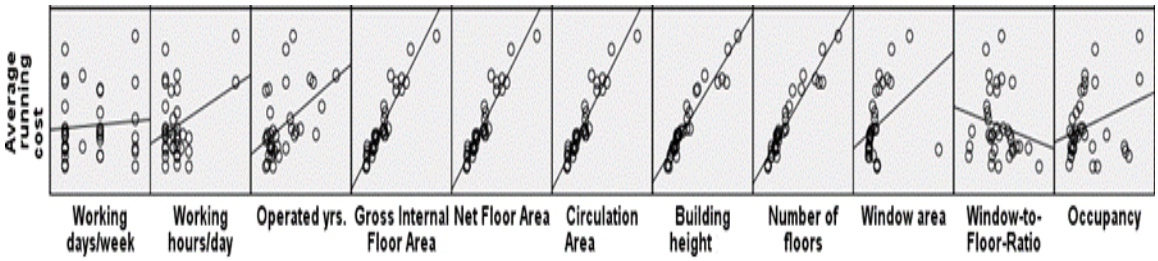

Figure 1. The relationship between the running cost and building design variables.

3.1 The Artificial Neural Network Model for Running Cost Estimation of Buildings

Initially, the pretreated data (with multiple imputation and normalization) was input to the "Neural Designer" software and the data set consists of 35 instances was itself divided into three types of instances such as 60% of training instances, 20% of selection instances, and 20% of testing instances. Next, the data set was further pretreated by setting all instances, which includes constant variables, repeated variables and univariate outliers and multivariate outliers as unused instances. Fortunately, the results derived indicate that there are no any constant or repeated variables in the data set and only 2 instances were set as unused due to outliers. Then, the neural network was developed and it represents the predictive model. Accordingly, the size of the scaling layer is 13, the number of inputs. The scaling method for this layer is the "MinimumMaximum". Further, the neural network was designed with three layers. Table 2 depicts the size of each layer and its corresponding activation function. The architecture of this neural network can be written as 13:10:7:1.

Table 2. The summary of the neural network.

	Input number	Neurons number	Activation function
1	13	10	Linear
2	10	7	Linear
3	7	1	Linear

The statistics of the parameters shown in Table 3 depict information about the complexity of the model. In general, it is desirable that all minimum, maximum, mean and standard deviation values are not very big as shown for the developed model.

Table 3. Parameters statistics of the natural network model.

	Minimum	Maximum	Mean	Deviation
Parameters	-0.990295	0.990173	-0.0601779	0.57255

The loss index plays an important role in the use of a neural network. It defines the task the neural network is required to do and provides a measure of the quality of the representation that it is required to learn. The normalized squared error (NSE) is used here as the error method. If the NSE has a value of unity then the neural network is predicting the data 'in the mean', while a value of zero means a perfect prediction of the data. In this network, the NSE is 3.28.

The procedure used to carry out the learning process is called training (or learning) strategy. The quasi-Newton method was applied as the training strategy of the neural network in this study in order to obtain the best possible loss. It is based on Newton's method but does not require

calculation of second derivatives. Instead, the quasi-Newton method computes an approximation of the inverse Hessian at each iteration of the algorithm, by only using gradient information. Accordingly, the initial value of the training loss is 3.18746, and the final value after 230 iterations is 0.00238913 whereas the initial value of the selection loss is 5.59346, and the final value after 230 iterations is 0.208116.

A standard method to test the loss of a model is to perform a linear regression analysis between the scaled neural network outputs and the corresponding targets for an independent testing subset. This analysis leads to three parameters for each output variable. The first two parameters, a and b, corresponding to the y-intercept and the slope of the best linear regression relating scaled outputs and targets. The third parameter, R^2, is the correlation coefficient between the scaled outputs and the targets. If we had a perfect fit (outputs exactly equal to targets), the slope would be 1, and the y-intercept would be 0. If the correlation coefficient is equal to 1, then there is a perfect correlation between the outputs from the neural network and the targets in the testing subset. Accordingly, Table 4 lists the linear regression parameters for the scaled output running cost/sq. ft.

Table 4. The linear regression parameters for the scaled output running cost/sq. ft.

Regression parameters	**Value**
Intercept	-0.0355
Slope	1.14
Correlation	0.966

The mathematical expression represented by the neural network inputs working days/week, working hours/day, attached/detached, age, gross internal floor area, net floor area, circulation area, height, no. of floors, type of structure, window area, window to floor ratio and occupancy to produce the output Running cost/sq. ft. For function regression problems, the information is propagated in a feed-forward fashion through the scaling layer, the perceptron layers and the unscaling layer.

3.2 Model Testing

The purpose of model testing is to evaluate the performance of the developed ANN model in estimating a functional form that relates the design variables of commercial buildings to the running cost. Table 5 presents the prediction accuracy of the developed ANN model with use of the MAPE and Theil's U statistic, which commonly used to measure the performance.

Table 5. Results of test statistics for model accuracy.

Test	**ANN**
MAPE	-4.9%
Theil's U value	0.049

As shown in Table 5, the average MAPE of the ANN model is -4.9%, indicates that the ANN model has been achieved a high accuracy. The Theil's U value for the ANN model is 0.049 (where the U value indicates greater accuracy as $U \rightarrow 0$). Further, the neural network model recorded a correlation of 0.966, this accuracy is better than that recorded by the so far developed parametric regression models for LCC estimation.

4 CONCLUSIONS

The paper has highlighted the importance of different modelling techniques for predicting cost to be incurred during the operation phase of buildings particularly, commercial. The level of MAPE for the ANN model can be considered acceptable in most real applications, depending on the phase of application of the model. It is clear from this limited experiment that ANN was able to extract a functional form (i.e., a function) that represents the problem under investigation. The study has also shown that ANN models may prove as a good alternative to parametric cost modelling. Within the limits of this study, ANN models have been shown to be able to model data that strongly exhibit noise and achieve reasonable accuracy.

Acknowledgments

This work was supported by the Senate Research Committee of University of Moratuwa under the Grant No. SRC/LT/2017/21.

References

Al-Hajj, A., and Horner, M. W., Modelling the Running Costs of Buildings, *Construction Management and Economics*, Taylor and Francis, 16(4), 459-470, 1998.

Boussabaine, A. H., Kirkham, R. J., and Grew, R. G., Estimating the Cost of Energy Usage in Sport Centres: A Comparative Modelling Approach, *15th Annual ARCOM Conference*, Hughes, W. (ed.), 481-488, Liverpool, UK, 1999.

El-Haram, M. A., Marenjak, S., and Horner, M. W., Development of A Generic Framework for Collecting Whole Life Cost Data for The Building Industry, *Quality in Maintenance Engineering*, Emerald, 8(2), 144-151, 2002.

Fumo, N., and Biswas, R. M. A., Regression Analysis for Prediction of Residential Energy Consumption, *Renewable and Sustainable Energy Reviews*, Elsevier, 47(C), 332-343, 2015.

Kirkham, R. J., Boussabaine, A. H., and Awwad, B. H., Probability Distributions of Facilities Management Costs for Whole Life Cycle Costing in Acute Care NHS Hospital Buildings, *Construction Management and Economics*, Taylor and Francis, 20(3), 251-261, 2002.

Kirkham, R. J., Boussabaine, A. H., Grew, R. G., and Sinclair, S. P., Forecasting the Running Costs of Sport and Leisure Centres, *08th International Conference on Durability of Building Materials and Components*, Lacasse, M. A., and Vanier, D. J. (eds.), 1728-1738. Institute for Research in Construction, 1999.

Krstić, H., and Marenjak, S., Maintenance and Operation Costs Model for University Buildings, *Technical Gazette*, 24(1), 193-200, 2017.

Pasini, A., Artificial Neural Networks for Small Dataset Analysis, *Journal of Thoracic Disease*, AME Publication Company, 7(5), 953, 2015.

Pasini. A., Pelino, V., and Potestà, S., A Neural Network Model for Visibility Nowcasting from Surface Observations: Results and Sensitivity to Physical Input Variables, *Journal of Geophysical Research*, American Geophysical Union, 106(D14), 14951-14959, July, 2001.

Pasini, A., and Potestà, S., Short-Range Visibility Forecast by Means of Neural-Network Modelling: A Case-Study, *II Nuovo Cimento C*, Springer, 18(5), 505-516 September, 1995.

Wang, N., Wei, K., and Sun, H., Whole Life Project Management Approach to Sustainability, *Management in Engineering*, American Society of Civil Engineers, 30(2), 246-255, March, 2014.

Wong, I. L, Perera, S., and Eames, P. C., Goal Directed Life Cycle Costing as A Method to Evaluate the Economic Feasibility of Office Buildings with Conventional and TI-Façades, *Construction Management and Economic*s, Taylor and Francis, 28(7), 715-735, 2010.