NEURAL NETWORKS FOR MODELLING THE FINAL TARGET COST OF WATER PROJECTS

Dominic D Ahiaga-Dagbui[1] and Simon D Smith[2]

School of Engineering, University of Edinburgh, EH9 3JL, Scotland, UK.

Producing reasonably accurate cost estimates at the planning stage of a project is important for the subsequent success of the project. The estimator has to be able to make judgement on the cost influence of a number of factors including site conditions, procurement, risks, price changes, likely scope changes or type of contract. This can shroud the estimation process in uncertainty, which has often resulted in project cost overruns. The knowledge acquisition, generalisation and forecasting capabilities of Artificial Neural Networks (ANN) are explored in this pilot study to build final cost estimation models that incorporate the cost effect of some of the factors mentioned above. Data was collected on ninety-eight water-related construction projects completed in Scotland between 2007-2011. Separate cost models were developed for normalised target cost and log of target costs. Variable transformation and weight decay regularisation were then explored to improve the final model's performance. As a prototype of a wider research, the final model's performance was very satisfactory, demonstrating ANN ability to capture the interactions between the predictor variables and final cost. Ten input variables, all readily available or measurable at the planning stages for the project, were used within a Multilayer Perceptron Architecture and a Quasi-Newton training algorithm.

Keywords: cost estimation, cost modelling, neural networks.

INTRODUCTION

Cost estimation is a heavily experience-based process, and involves the evaluation of several complex relationships of cost-influencing factors, largely based on professional judgement (Alex *et al.* 2010). A thorough cost estimation exercise would involve the evaluation of the cost effect of factors such as site restrictions, ground conditions, contract type, location of the project, procurement method, etc. However, preliminary investigations show that this is rarely the case, most likely due to the difficulties of quantifying the cost implications of these factors. The classical way of accounting for the cost effect of these variables is using the so-called contingency fund (Baccarini 2005), which unfortunately has mostly failed to keep construction projects within budget.

[1] d.ahiaga-dagbui@ed.ac.uk

[2] simon.smith@ed.ac.uk

Ahiaga-Dagbui, DD and Smith, SD (2012) Neural networks for modelling the final target cost of water projects *In*: Smith, S.D (Ed) *Procs 28th Annual ARCOM Conference*, 3-5 September 2012, Edinburgh, UK, Association of Researchers in Construction Management, 307-316.

Traditional cost estimation i.e. estimating the cost of labour and materials and making allowance for profits and overheads for individual construction items, is deterministic by nature (Okmen *et al.* 2010) and largely insufficient in reaching the actual final cost of a project. The approach largely neglects and poorly deals with uncertainty and their correlation effects on cost (Oztas *et al.* 2005). It is also difficult to account for the cost effect of some of the variables mentioned above using the traditional cost estimation method.

The aim of this experimental research, which is part of a larger research in integrating risk and cost modelling, is to explore the use of Artificial Neural Networks (ANN), as a data mining technique for developing cost forecast models of construction projects. ANN is employed to model the relationships between qualitative factors that have an impact on construction cost and quantifiable items that represent different cost centres in the bills of quantities. The paper provides an overview of cost estimation, estimation accuracy and cost models as well as neural network theory and applications. Details about the development of a predictive model for final target cost of water projects are detailed, with conclusions and recommendations for further research.

COST PLANNING AND ESTIMATION

Effective cost planning relates the design of construction projects to their cost, so that while taking full account of quality, risks, likely scope changes, utility and appearance, the cost of a project is planned to be within the economic limit of expenditure (Kirkham *et al.* 2007). This stage in a project life-cycle is particularly crucial as decisions made during the early stages of the development process carry more far-reaching economic consequences than the relatively limited decisions which can be made later in the process. This initial process may also influence the client’s decision on whether or not to progress with the project. The cost planning process leads to the generation of a reliable initial project budget that sets up a cost control system to ensure that client expectations are met. For many clients, completing the project within this initial budget is a paramount determinant of client satisfaction. Despite the great importance of cost estimation, it is undeniably not simple nor straightforward because of the lack of information in the early stages of the project (Hegazy 2002).

Cost estimation, the determination of quantity and cost required to construct a facility or to furnish a service (Westney 1992), forms the crux of the cost planning exercise. The approach used for cost estimation normally varies from the early strategic phase of a project to the construction phase and will depend on a number of other factors including level of accuracy required, the speed estimation required, experience level of the estimator and the level of information available at the time of estimate. Accurate estimation of future cost however, is a difficult task (Nicholas 2004), if not an elusive aim. This can mostly be attributed to the fact that cost estimation, which must not be confused with budgeting, occurs at the conception phase of the project, before many of the cost influencing factors about the project are available even to the client (Hegazy 2002).

COST MODELS

Ferry *et al.* (1999) also describe cost models as the symbolic representation of a system, expressing the content of that system in terms of the factors which influence its costs. The models may be in the form of mathematical equations (eg. Regression models) or a set of defined steps to estimate the cost of a particular item (eg. Storey enclosure method). Cost models can be very useful in strategic level decisions such as bid/not to bid decisions, with potential saving of time and effort on non-viable projects. They are furthermore appealing because of current harsh economic climate with tough competition and limited resources. However, the production of reasonably accurate, acceptable and timely parametric cost estimates can be a difficult task. For example, using only 4 different parameters for a project and considering three alternative values for each, and varying one at a time will produce 81 different project solutions or alternatives. This can be done rather rapidly using an computer-based model but will undoubted be a laborious task using traditional cost estimation (Sequeira 1999). The time, effort and resource level required for this task would mostly be unjustifiable at planning stages of a project, perhaps a strong suggestion that detailed cost estimates at strategic level are often far from the optimal solutions because of time and resource constrains.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks, henceforth referred to as neural networks (NN) with artificial implied, is an analogy-based, non-parametric information-processing system that has performance characteristics similar to a biological neural network of the brain (Anderson *et al.* 1992). They retain two features of the biological neural network: the ability to learn from experience and make generalisations based on this acquired knowledge (Haykin 1994).

Neural networks are structured to provide the capability to solve problems without the benefits of an expert and without the need of programming. They can seek patterns in data that are not obvious (Anderson and McNeill 1992) and are particularly suited for complex, hard-to-learn problems where no formal underlying theories or classical mathematical and traditional procedures exist (Adeli 2001). NNs are fundamentally different from algorithmic computing and statistical methods like regression in one way- they learn inductively by examples and then are able to generalise solutions (Flood *et al.* 1994). Modelling techniques including regression analysis, case-based reasoning and fuzzy logic analysis find it difficult dealing with problems such as imprecision, incomplete and uncertainty of data and other variables affecting costs and implicit combinatorial effects and inter-relationships of cost variables (Flood and Kartam 1994), areas where NN is often at its best.

Applications of neural networks

Neural network has been used successfully for foreign exchange prediction (Shi *et al.* 2011, Khashei *et al.* 2012); medical diagnosis (Dreiseitl *et al.* 2009); flight and robot control (NASA 2003, Lee *et al.* 2010); and loan applicant assessment (Malhotra *et al.* 2003). Earliest construction industry application of neural networks can be traced back to 1989 by Adeli and Yeh (1989) on engineering design and machine learning. It has since been used in construction management for estimating the cost of highway projects (Wilmot *et al.* 2005, Pewdum *et al.* 2009); predicting the cost of water and sewer installations (Alex *et al.* 2010) and building projects (Emsley *et al.* 2002); mark-up estimation (Li *et al.* 1999); risk quantification (McKim 1993); and tender price forecast (Boussabaine *et al.* 1999). Neural Network application bibliographies have been provided by Adeli (2001) for Civil Engineering and Moselhi *et al.* (1991) for construction management research.

Training the Neural Network

A neural network, like the human brain, learns from experience (Hinton 1992). Experience here refers to past data within the domain of the problem under study. The aim of any training regime is to help the network to continuously reduce the error of its predictions by varying the weights between its connections (Setyawati *et al.* 2003). Examples of the training set are presented to the network in its input layer. These are then transferred to the hidden layer by some form of activation function, normally a linear activation function. Random weights are applied to these input values in the hidden layer and then their cumulative weighted values transferred to the output layer. If the training algorithm adopted is a supervised one, the result of the training, called the output, is compared to the target (the expected real value) at the output layer and the error (difference between the output and the expected value, normally measured as the root mean squared error RMSE) is computed. This is then sent as feedback to the network and an error function is used to try and minimize the value of the error in the next cycle of training. The most common form of learning is the back propagation method, which is a supervised learning method (Setyawati, Creese and Sahirman 2003).

Neural Network Problems

Neural networks do exact their own demands however. NN are data-hungry, and performance is largely dependent on plenteous, representative and reliable data (Anderson and McNeill 1992). Another major criticism of the NN approach to data modelling is that it offers little explanation on the relationships between the variables it is modelling (Boussabaine *et al.* 1997, Hair *et al.* 1998). The technique is still disregarded by some researchers, referring to it as a 'black-box' technique because the network parameters do not offer casual explanations, making it difficult to elucidate what is learnt from the neural network model (Paliwal *et al.* 2011). To these criticisms, some have argued that it might be preferable to focus on how well a neural network model produces its results, rather than how it produces it (Hair *et al.* 1998). It is envisaged that further research into framework and internal processes within the neural network will offer better explanatory insight into the influence of independent variables in the modelling process.

DATA

Data was collected on ninety-eight water projects completed in Scotland between 2007 and 2011. The nature of the projects were rather varied, ranging from construction of water mains, water treatment plants, Combined Sewer Overflows (CSOs), installation of manholes or water pumps and upgrades and repairs to sewers. All the projects were target cost contracts with values between £9,000-£14million and durations from 1-22months.

MODEL DEVELOPMENT

The modelling process involved investigating the performance of different network topologies and parameters in predicting the final cost of the projects. It was carried out using the Statistica 10 software, in the stages detailed below:

Data Pre-processing

The aim of data pre-processing is to structure and present the data to the model in the most suitable way as well as offer the modeller the chance to get to know the data thoroughly. For this research, extreme values and outliers were either re-coded or deleted from the sample set and missing values replaced with the mean or mode. Input errors were corrected and all cost values were normalised to 2010 with the base year 1995 using the BIS cost indices. Invariant variables, such as procurement option, payment method, fluctuation measure and type of client, were removed from the variable set as they would only increase the model complexity and yet offer no useful information for model performance. Finally, categorical variables such as type of project, need for project, etc. were coded using the one-of-N coding, resulting in 4 sub-variables for type of soil for example (Good, Moderate, Poor, Not Applicable). Twenty-eight sub-variables resulted out of the initial 11 input variables. This coding allowed the model to infer importance on its own without the modeller imposing weightings or subjective ratings to the variables. Ninety project cases remained after the pre-processing stage and were then passed on for the modelling proper.

Phase One: TC and FTC

At this stage, the model was developed using the raw normalised estimated target cost (TC) and final target cost (FTC). Two different network architectures, the Multilayer Perceptron (MLP) and the Radial Basis Function (RBF), were experimented initially. RBF models the relationship between inputs and targets in a 2 phases: it first performs a probability distribution of the inputs before the searching for relationships between the input and output space in the next stage (StatSoft Inc. 2011a). MLPs on the other hand model using just the second stage of the RBF. As expected, the MLP models were superior to the RBF networks for this regression problem and so the rest of the modelling was carried out using just MLPs.

The network was set to train 200 different models, iterating between 1-50 nodes in a single hidden layer using a data split of 75:15:10% for training, testing and validation sample sets respectively. The three best networks were retained and examined for further improvement. The validation set was not used in the training of the model so can be considered as an independent verification of the model’s ability to generalise on new data. Five different transfer functions- logistic, tanH, negative exponential, identity and sine were each tested. These transfer functions are used to squash the data range of the processing signals to values normally between 0 to 1 or -1 to +1 since the neural network algorithms are most sensitive to inputs within a small range. Gradient descent, Conjugate descent and Quasi-Newton (BFGS) training algorithms were also experimented for all the models. Early stopping, the process of halting training when the test error stops decreasing, was used to prevent memorising or over-fitting the dataset in order to improve generalization. Over-fitted models perform very well on training and testing data, but fail to generalise satisfactorily when new ‘unseen’ cases are used to validate their performance.

Overall performance of the network is measured using the correlation coefficient between predicted and output values as well as the Sum of Squares (SOS) of errors. SOS is defined here as:

$$SOS = \sum (O_i - T_i)^2 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\text{(eqn. 1)}$$

Where Oi is the prediction (network outputs)
T_i is the target (actual value) of the *i*th data case.

The higher the SOS value, the poorer the network at generalisation, whereas the higher the correlation coefficient, the better the network. The p-values of the correlation coefficients were also computed to measure their statistical significance as a test of whether the observed correlations were achieved by fluke. The higher the *p*-value, the less reliable the correlations observed.

The results from the best network for this phase was rather unsatisfactory as the errors observed were very high (See table 1.0), most likely due to the use of the raw data for the modelling. The best network at this stage was an MLP with 25 input variables, 31 nodes in the hidden layer. It was trained using a BFGS training algorithm, tanH and logistic activation functions in the hidden and output layers respectively.

Table 1: Network Performance: TC and FTC

Index	Net. name	Test perf.	Validation perf.	Test error	Validation error
1A	MLP 25-31-1	0.917	0.990	8.830E+10	2.096E+10
2A	MLP 25-37-1	0.928	0.988	8.163E+10	8.964E+09
3A	MLP 25-50-1	0.921	0.987	8.555E+10	1.348E+10

Phase two: LogTC and logFTC

The common log values of TC and FTC were then used for the next phase as it has been suggested that data transformation can significantly improve performance of NN models (Shi 2000). The 3 best networks were retained after training 200 different networks using the same parameters as above. The results showed significant improvement in the error values but slightly deteriorated in correlation (see table 2.0). This can be attributed to the fact that log of TC and FTC reduced the cost inputs to a smaller range, making them more sensitive to the training algorithms of neural networks. The common log of the target costs most likely made it easier for the network to learn the relationships between the variables than in the previous phase.

Table 2: Network performance: Log TC and logFTC

Index	Net. name	Test perf.	Validation perf.	Test error	Validation error
1B	MLP 25-29-1	0.925	0.933	0.091	0.131
2B	MLP 25-48-1	0.918	0.932	0.100	0.125
3B	MLP 25-16-1	0.893	0.936	0.174	0.134

Phase three: log FTC and logTC with Weight Decay

The effect of using weight decay regularisation in the hidden and output layers was then investigated. This was an attempt to encourage the network to develop smaller weights to further reduce the problem of over-fitting, thereby potentially improving generalization performance of the network. Weight decay modifies the network's error function to penalize large weights - the result is an error function that compromises between performance and weight size (StatSoft Inc. 2011b). The results showed a further improvement in both the error and correlation coefficient for the validation samples. The validation performance of the best network was now 0.968 with a p-value of 0.00 and an SOS of 0.062. The number of neurons in the hidden layer had also reduced from 29 in the best model to 19 when weight decay was applied.

Evidently, the model was getting better in predicting the final cost of projects based when the learning reinforcement technique of weight decay was used.

Table 3: Network Performance with Weight decay regularisation

	Net. name	Test	Validation	p-value	Test	Validation	Training
		Perf.	Perf.	p<0.050	error	error	algorithm
1C	MLP 25-19-1	0.983	0.968	0.00	0.092	0.062	BFGS 89
2C	MLP 25-22-1	0.929	0.958	0.00	0.065	0.064	BFGS 26
3C	MLP 25-22-1	0.948	0.949	0.00	0.066	0.098	BFGS 56

A relative importance table below shows each variable's contribution to the model's generalisation abilities. At this stage, table four is indicative of the relative influence of the various inputs on the outturn cost. It gives the contractor important information on which factors need most attention during the tendering stage, especially in terms of final cost. The client/contractor would then be able to simulate the effect of changing these factors within the model to see its direct likely impact on the final cost. The SOS of residuals for the full model is computed and compared to that of the reduced model when each predictor is removed from the neural network. The variables are then arranged in order of importance according to the change in performance noticed when they were removed. The initial estimated target cost was the most important factor, as could be expected, and site access contributed very little to the model. Duration of the projects was unexpectedly ranked 7th in the relative importance table. In general, longer project durations tend to cost more than shorter ones. The observation here might be due to the poor representation of the number of projects across the range of durations used in the model building. More than 65% of the project cases were completed within four (4) months which would make the model biased towards projects within this class. This may mean that the model in its current form might not be a good predictor for projects with durations in excess of 4 four months. The high ranking of project frequency, tendering strategy and contractor's need for the project indicates the attention that has to be given these factors when preparing tender documents.

Table 4: Relative Importance of Variables

Factor	Weighting	Ranking
logTC	5.91	1
Project Frequency	2.55	2
Tendering Strategy	2.52	3
Need for Project	2.00	4
Ground Condition	1.45	5
Project Type	1.38	6
Duration	1.20	7
Location	1.16	8
Soil Type	1.05	9
Site Access	1.00	10

CONCLUSIONS

Artificial Neural Network is used to develop a cost estimation model for water projects in this paper. Their ability to capture and generalise non-linear relationships are exploited to detect the interactions in qualitative variables like tendering method, contractors need for the project, location, site access and project type in developing cost models to predict the final target cost of water projects. The use of weight decay regularisation to encourage the development of a parsimonious network to improve the model's performance and reliability was also investigated. This showed significant promise for future analysis if combined with other techniques like pruning and sensitivity analysis of predictor variables. As a prototype of a wider research, the results achieved are very satisfactory and will potentially be improved with a larger dataset in this on-going research

The developed models have several potential applications in industry and construction management. The model can easily be converted to a desktop package that construction professionals could use in rapid prediction of final cost of projects using only factors that are readily available or measurable at planning stage of the project. It is also very useful at the design stage of a project when information is incomplete and detailed designs are not available. The use of the model could also greatly reduce the time and resources spent on estimation as well as provide a benchmark to compare detailed estimates. It will further allow the generation of various alternative solutions for a construction project using 'what if' analysis for the purposes of comparison.

REFERENCES

Adeli, H (2001) Neural networks in civil engineering: 1989-2000. *Computer-Aided Civil and Infrastructure Engineering*, **16**(2), 126-42.

Adeli, H and Yeh, C (1989) Perceptron learning in engineering design. *Computer-Aided Civil and Infrastructure Engineering*, **4**(4), 247-56.

Alex, D P, Al Hussein, M, Bouferguene, A and Siri Fernando, P (2010) Artificial Neural Network Model for Cost Estimation: City of Edmonton's Water and Sewer Installation Services. *Journal of Construction Engineering and Management*, **136**, 745.

Anderson, D and McNeill, G (1992) Artificial neural networks technology. *A DACS (Data & Analysis Center for Software) State-of-the-Art Report, Contract Number F30602-89-C-0082*, 87.

Baccarini, D (2005) Understanding project cost contingency - A survey. In: *Queensland University of Technology Research Week*, Sidwell, A C, Ed.), Brisbane, Queensland: Queensland University of Technology.

Boussabaine, A and Elhag, T (1997) A neurofuzzy model for predicting cost and duration of construction projects. *RICS Research (9 p.). The Royal Institution of Chartered Surveyors*.

Boussabaine, H and Elhag, T (1999) Tender Price Estimation Using ANN Methods, EPSRC Research Grant (GR/K/85001). In, Liverpool, UK: School of Architecture & Building Engineering, University of Liverpool.

Dreiseitl, S, Binder, M, Hable, K and Kittler, H (2009) Computer versus human diagnosis of melanoma: evaluation of the feasibility of an automated diagnostic system in a prospective clinical trial. *Melanoma research*, **19**(3), 180.

Emsley, M W, Lowe, D J, Duff, A, Harding, A and Hickson, A (2002) Data modelling and the application of a neural netowrk approach to the prediction of total construction costs. *Construction Management and Economics*, **20**, 465-72.

Ferry, D J, Brandon, P S and Ferry, J D (1999) *Cost planning of buildings*. Vol. 7, Oxford, UK: Blackwell Science Ltd.

Flood, I and Kartam, N (1994) Neural networks in civil engineering. I: Principles and understanding. *Journal of Computing in Civil Engineering*, **8**(2), 131-48.

Hair, J, Tatham, R, Anderson, R and Black, W (1998) *Multivariate Data Analysis (5th Edition)*. Prentice Hall.

Haykin, S (1994) *Neural networks: a comprehensive foundation.* Prentice Hall PTR Upper Saddle River, NJ, USA.

Hegazy, T (2002) *Computer-based construction project management*. Upper Saddle River, NJ: Prentice Hall Inc.

Hinton, G E (1992) How neural networks learn from experience. *Scientific American*, **267**(3), 144-51.

Khashei, M and Bijari, H (2012) Exchange rate forecasting better with hybrid artificial neural networks models. *Journal of Mathematical and Computational Science*, **1**(1).

Kirkham, R and Brandon, P S (2007) *Ferry and Brandon's Cost Planning of Buildings*. 8th ed. John Wiley & Sons.

Lee, C T and Tsai, C C (2010) Nonlinear adaptive aggressive control using recurrent neural networks for a small scale helicopter. *Mechatronics*, **20**(4), 474-84.

Li, H and Love, P E D (1999) Combining rule-based expert systems and artificial neural networks for mark-up estimation. *Construction Management and Economics*, **17**(2), 169-76.

Malhotra, R and Malhotra, D (2003) Evaluating consumer loans using neural networks. *Omega*, **31**(2), 83-96.

McKim, R A (1993) Neural networks and identification and estimation of risk. *AACE International Transactions*(15287106), P.5.1-P.5.1.

Moselhi, O, Hegazy, T and Fazio, P (1991) Neural networks as tools in construction. *Journal of Construction Engineering and Management*, **117**(4), 606-25.

NASA. *NASA Neural Network Project Passes Milestone* Dryden Flight Research Center, NASA, 2003 [cited 6th December, 2011. Available from http://www.nasa.gov/centers/dryden/news/NewsReleases/2003/03-49.html.

Nicholas, J M (2004) *Project management for business and engineering: Principles and practice*. Second ed. MA, USA; Oxford, UK: Elsevier Butterworth–Heinemann.

Okmen, O and Öztas, A (2010) Construction cost analysis under uncertainty with correlated cost risk analysis model. *Construction Management and Economics*, **28**(2), 203-12.

Oztas, A and Okmen, O (2005) Judgmental risk analysis process development in construction projects. *Building and Environment*, **40**(9), 1244-54.

Paliwal, M and Kumar, U A (2011) Assessing the contribution of variables in feed forward neural network. *Applied Soft Computing*, **11**(4), 3690-6.

Pewdum, W, Rujirayanyong, T and Sooksatra, V (2009) Forecasting final budget and duration of highway construction projects. *Engineering, Construction and Architectural Management*, **16**(6), 544-57.

Sequeira, I (1999) *Neural network based cost estimation*, Masters, Department of Building, Civil and Environmental Engineering, Concordia University.

Setyawati, B R, Creese, R C and Sahirman, S (2003) Neural Networks for Cost Estimation (Part 2). *AACE International Transactions*, 1-.

Shi, C, Wang, H, Yin, F and Ru, Z (2011) ARIMA and neural network prediction of foreign exchange reserves. *In*. IEEE, Vol. 2, 986-9.

Shi, J J (2000) Reducing prediction error by transforming input data for neural networks. *Journal of Computing in Civil Engineering*, **14**, 109.

StatSoft Inc. (2011a) *STATISTICA 10 (data analysis software system), www.statsoft.com*, Version 10.

StatSoft Inc. (2011b) Electronic Statistics Textbook. In, OK Tulsa: StatSoft, .

Westney, R E (1992) *Computerized management of multiple small projects.* Marcel Dekker, Inc.

Wilmot, C G and Mei, B (2005) Neural network modeling of highway construction costs. *Journal of Construction Engineering and Management*, **131**, 765.