IMPLEMENTATION OF A NEURAL NETWORK MODEL FOR THE COMPARISON OF THE COST OF DIFFERENT PROCUREMENT APPROACHES

Anthony Harding, David Lowe, Adam Hickson, Margaret Emsley and Roy Duff

Department of Building Engineering, UMIST, PO Box 88, Manchester, M60 1QD, UK

The choice of procurement system for a building project is a very significant one. However, there is currently very little comparative cost data to inform the selection of procurement system, especially the total cost to the client.

This paper reports on a model that will make such a comparison possible. This model, which is currently under development at UMIST, is designed to consider 39 project variables, including the choice of procurement system, and estimate the final cost of the project using a neural network. The suitability of a neural network to model this problem has already been established by a pilot study. The advantage of using this type of model is that it permits comparisons between the different procurement methods to be made within the context of that particular project, rather than within projects as a whole.

The classification and representation of the different variables to be considered within this model are discussed. In addition to this, the implementation of factor/cluster analysis to reduce the number of variables required by the neural network, and hence increase its accuracy, is explained. Furthermore, the level of confidence of the model and its implications for the implementation of a “What if?” analysis are discussed. This analysis would allow the client to assess how changing certain variables, including procurement, might affect the final cost of the project.

Keywords: cost modelling, early stage estimating, neural networks, procurement.

INTRODUCTION

The selection of the procurement route can have a significant influence on the final cost of a building project. However, Masterman (1994) found no evidence, in two studies of clients’ selection procedures, of cost differences being considered when choosing a building procurement system. This discovery provided the motivation for an EPSRC sponsored research project, begun at UMIST in June 1996, which aimed to investigate and analyse the comparative costs of using different building procurement systems as well as the techniques which might be used to do this.

The specific objectives of this feasibility study were to: evaluate the data currently available in the collaborating firms; identify the most appropriate analytical and predictive model(s); define any further data required to produce satisfactory analyses; establish the availability of this data; and, carry out preliminary model testing.

The results of this study (Duff *et al.* 1998) showed the potential benefit of the development of a neural network model for the comparison of the costs of different procurement routes. Subsequent to the feasibility study a second project, also funded by the EPSRC, is now underway at UMIST. The purpose of this project is to develop

Harding, A, Lowe, D, Hickson, A, Emsley, M and Duff, A R (1999) Implementation of a neural network model for the comparison of the cost of different procurement approaches. *In:* Hughes, W (Ed.), *15th Annual ARCOM Conference*, 15-17 September 1999, Liverpool John Moores University. Association of Researchers in Construction Management, Vol. 2, 763-72

a predictive neural network model of the cost to the client which would permit the comparison of the cost of different procurement routes.

Previous research

The Building Economic Development Committee (1974) made one of the earliest comparisons between the cost of different procurement routes. They produced a guide in which cost performance on projects designed by different types of consultant was measured by comparing the percentage of projects completed within 5% of the estimated cost. The results were restricted to a comparison of conventional and design and build methods of procurement and the validity of the analysis is, of course, dependent upon the relative accuracy of the estimates.

The Department of Industry and the Department of the Environment (1982) made a subjective comparison between traditional, design and build and management procurement routes. The guide they produced compared them against three cost criteria. Differences were not quantified but subjectively evaluated as “higher” or “lower” than the traditional procurement route.

Brandon 1988) suggested that the cost of the procurement route be taken into account by means of an addition. The additions range from 0% for the conventional and design and build systems to 15% for construction management, apparently based on experience.

Design and build (Franks 1983, Rowlinson 1986), management contracting (CCMI 1985, Sidwell 1983) and construction management (Moore 1984, Olashore 1986, Reading University 1991) systems have been the subject of much individual examination. Nevertheless, it is only recently that any objective evaluation of comparative costs has emerged. In a recent report on design and build (Reading University 1996), the Centre for Strategic Studies in Construction has concluded that the design and build procurement system is “at least 13% cheaper than traditional procurement”. The analyses were carried out using multiple regression and identified 11 variables, including choice of procurement system, which together explained 51% of the variability in project cost.

Masterman (1994) investigated procurement system selection methods and found that references to the financial aspects of selection are limited to clients’ strategic objectives, such as whether uncertainty on price is acceptable, price competition is important, or a firm price is required before work can commence. The actual cost of the choice of procurement cost was not objectively evaluated, simply because there was no means of doing so.

The aim of the research is therefore to provide a predictive model of the cost of different procurement routes using the 39 variables identified in the pilot study. While earlier research was aimed at establishing a general cost difference, it is recognized in this research that the difference in cost is not necessarily a blanket figure, independent of other cost significant variables. This means that while one procurement route may be cheaper for one type of project, it cannot be assumed that that same route will be cheaper for another type of project.

THE MODEL

Previous research has tended to assume that any differences in the cost of different procurement routes would be relatively constant from project to project. However, in the absence of any empirical data to support it, it was felt that this assumption could not be made. Therefore, the relative cost of different procurement routes would have to be determined specifically for each project. The simplest way of determining this cost difference between procurement routes is to create a model that is able to predict the cost of the project from all the cost significant variables, including procurement. This model can then be used to evaluate the expected costs of the project under different procurement routes, and the differences observed.

In order to accurately evaluate the differences in cost between the different procurement routes for any particular project, the model is required to model the complex and little understood interrelationships which exist between all the cost significant variables at this early stage of the project. Because the relationships between the different variables are so difficult to identify and quantify, it was decided to use techniques which are able to infer the relationships between the cost significant variables using existing data. Two such modelling techniques were identified in the pilot study: linear regression and neural networks. The neural network model was found to be the most appropriate technique.

The advantages of neural networks are as follows.

- Neural networks, unlike linear regression, are able to model interdependencies between input data which will inevitably occur when considering construction cost significant variables. For example, the model variables - such as number of storeys, gross floor area and number of lifts - will almost certainly be correlated.

- Neural networks can deal more readily with non-linear relationships. More significantly, these relationships are determined implicitly by the model, and do not therefore require them to be specified.

- Neural networks can, more effectively than regression, handle incomplete data sets. This is important because it is difficult to guarantee that complete data sets will be always be available, and an subsequent use of the developed model can take place before all the parameters of the project have been determined.

Requirements of the neural network

One of the advantages of using a neural network is that it can, theoretically, model any function, provided it is presented with enough data and given enough time for training. In the real world, however, collecting very large amounts of data can be difficult at best and impossible at worst, and the amount of time available to train the neural network can be restricted.

In order to ensure that sufficient data is available a simple rule of thumb was employed: that 10 to 20 times as many data sets as input variables are required. The target size set was 500 projects. This is at the lower end of the spectrum of values suggested by the rule of thumb (400-800), but was chosen because of the difficulty anticipated in collecting such quantities of data and because of the time restrictions on the project. When the model is implemented in industry, it is likely that many more data sets will become available. This will make for a more accurate model of reality.

In addition to the number of data sets it is also necessary to consider how the variables are represented as inputs to the model. This can affect both the accuracy of the neural network and the ease with which it can be trained. As a general rule it is best that a single input is used for a variable only when that variable has some meaning as a single variable (Tarassenko 1998). That is to say, if the value of the variable increases then it must represent the increase of some factor which influences the outcome of the model. This is not usually a problem for variables which are just real numbers, such as the floor area, but is a problem where a variable represents a selection from a number of categories (categorical variables). For example, with site access there is clearly an order between *unrestricted*, *restricted* and *highly restricted*, inasmuch as the restriction to access increases, it will be expected to cause the cost to increase. Therefore this variable can be represented by a single input. Each of the three categories maps to a different value of the input, and the value corresponding to *restricted* (the middle category) lies between the values corresponding to the other two categories.

Where such an order either does not exist or is not known, and there is a choice between more than two options, this approach will not work. The most common way of dealing with such variables is to use a series of binary inputs to represent the choices, where each choice has its own input, which is set to 1 when the variable's value is that choice, and all the variable's other inputs are set to 0. It is possible to reduce the number of binary inputs required by having one choice represented by all the inputs being 0. However, even when this strategy is used the number of binary variables required by the network is still $n - 1$, where n is the number of choices. If many variables are represented in this way then it can result in a significant increase in the number of inputs, which can have a detrimental effect on the accuracy of the neural network model. It is therefore necessary to try to find a means of representing variables as a single input wherever possible.

Representation of output variables

The output of the model is the whole total cost of the building to the client. There are two approaches to obtaining this figure. The first is to model the entire cost of the building as one output. The second is to model two figures: the final account sum and the additional costs to the client. Modelling the two values this way could allow the client to see how the costs will be broken between the actual construction of the building and the various additional fees which will be required. However, having two outputs from the neural network could make it less accurate than if there were only one. Therefore, a third option would be to have two neural networks, one for each of the two costs, and the accuracy of the three methods compared.

Representation of input variables

Each of the variables has been analysed in order to determine the best way of representing that variable to the neural network. The way these variables will be represented falls into four main groups: continuous variables, categorical variables which can be coded as a single input, categorical variables for which the best representation cannot be determined at this stage, and categorical variables which must be represented as a series of binary inputs.

The first of these groups comprises those variables which are real numbers, which are just a single continuous value, rather than variables which represent a choice between a number of options. These variables were: *Project duration*; *Gross internal floor area*; *Wall area*, *IT installation*; *Air conditioning*. Note that *IT installation* and *Air conditioning* are included because they can be specified as % coverage. In addition to these continuous variables there were also a few integer variables: *Number of lifts*; *Number of stories*. These behave in the same way as continuous variables, because they do not represent one of a choice of options, but a quantity.

Where the largest and smallest values of these variables differ by more than one order of magnitude it may be required to use the logarithm of that value as an input. This ensures that the range of values is more evenly distributed. It may also be possible to increase the speed with which the network is trained by standardising the data. Perhaps the easiest way of doing this is to subtract the mean of the input values for that variable and divide by the standard deviation, yielding a set of inputs with a mean of 0 and standard deviation of 1.

The rest of the variables represent one of a choice of categories. Where this choice is one between two categories, only one binary variable is required, but when there are more than two it is necessary to consider whether the categories can be placed in some kind of order of influence. For some variables, obtaining such an order was simple. The location of the project site, for example, is divided into three categories: *Central Urban*; *Other Urban*; and *Rural*. It is clear that there is a progression here from the crowded town centre situation to a rural one. Therefore *Central Urban* may be given a value of 0, *Other Urban* a value of 1 and *Rural* a value of 2. The variables which could be classified in this way were: *Quality*; *Access to site*; *Site topography*; *Type of location*; *Roof construction*; *Shape complexity*. Additional to this were variables which had only two categories: *Type of contract (fixed/fluctuating)*; *Nature of site (greenfield/brownfield).*

There were a great many more variables for which no such order was immediately apparent, and identification of some kind of order was more difficult. One of these variables was the *internal wall finishes*. In reality, the variable is a choice of different material combinations which will make up the finish. However, in order to resolve it to a variable whose values are more readily comparable the value of the input was set to be the standard cost per m^2 of that finish. This provides an order to the categories which is proportional to how much the finish is expected to impinge upon the final cost of the building. This technique of measuring each choice by some like for like cost measure was found to be suitable for a great many inputs: *Roof finishes*; *External walls*; *Stairs*; *Windows/external doors*; *Internal walls/partitions*; *Internal wall finishes*; *Floor finishes*; *Ceiling finishes*; *Fittings*; *Sanitary appliances*; *Disposal installation*; *Mechanical installation*; *Electrical installation*; *Special installations*.

Most of the input variables to the project have fallen into the first two groups. However, there are still a few variables with categories for which an order could not be identified. This was for either of two reasons. The first reason was because the actual differences in cost between the possible choices were uncertain. For example, there is not as much consensus on the comparative costs of the frame type in the literature than for, say, finishes. A further problem is that there is a possibility of the cheapest value of a variable like the frame type varying from building to building. If this is found to be the case then a binary input coding system would be required. Therefore it is necessary to be open to both the possibility of representing the variable as binary inputs and the possibility of representing it as a single variable. Then the most appropriate coding can be selected after careful analysis of the data. All the following variables belong to this group: *Type of substructure*; *Type of frame*; *Upper floor construction*; *Form of contract*.

In addition to these variables, there was a group of variables for which a binary input coding would be essential. These were *Tendering strategy*, *Purpose (speculative/bespoke/PFI)*, and *Procurement*. It was expected that the effect of changing the value of one of these variables would vary from project to project.

Therefore it would not be possible to identify any kind of order which would be consistent from project to project, and representation by binary inputs was required.

REDUCING THE VARIABLES

One of the constraints on the current research is that the data set of previous projects will be limited to around 500 projects. However, it is known that the accuracy of the neural network as a model of reality tends to be inversely proportional to the number of input variables. Therefore, if it is possible to remove variables which are of very little significance in the model, then the number of variables actually input to the neural network can be reduced and its accuracy can be increased.

Factor analysis

Factor analysis is a statistical technique in which differences in scores on several variables are represent by a few dimensions or constructs (Kline 1993). Factors can be defined as a linear combination of variables, and are believed to suggest underlying processes that have created the correlations among variables (Tabachnick and Fidell 1989). One goal, therefore, of factor analysis, applicable to this project, is representing relationships among sets of variables parsimoniously. Manly (1986) considers the technique to be useful for gaining insight into the structure of multivariate data. However, he refers to it as “something of an art” and not as objective as most statistical methods.

Comrey (1973) states that the particular variables selected for a study will have a critical bearing on the kinds of results obtained. This includes not only the selection of appropriate variables but also the measurement characteristics of the data variables and the subjects from which the data is obtained. The variables should ideally be continuous, or, if measured by means of an ordinal scale, be an acceptable approximation of continuous. While it is possible to conduct factor analysis using dichotomous data, the prospects of distorted results are reduced as the number of categories increases. The distribution of the data should be reasonably normal, avoiding bi-modality or extreme skewness (Comrey 1973). Tabachnick and Fidell (1989) suggest that, provided factor analysis is used descriptively as a convenient way to summarize the relationships in a large set of observed variables, assumptions regarding the distributions of variables are not in force. However, when statistical inference is used to decide the number of factors, multivariate normality is assumed. It is essential, therefore, to exclude, from the factor analysis, those variables for which a binary input coding was required.

Communality is the sum of squared loadings for a variable across factors. It is usually taken to indicate how much a variable has in common with a hypothetical set of factors. Comrey (1973) considers communalities to represent the extent of overlap between the variables and factors. They give the proportion of variance in the variables that can be accounted for by the scores in the factors. The greater the communality, the better the factors account for the variance of a specific variable (Kline 1993). Child (1990) suggests that where the communality is too low (for example, 0.3 or less) one could be justified in eliminating that item in a re-analysis. Low communality values coupled with insignificant correlation coefficients between the variable and the output, would justify the removal of variables from the model on two counts. The first is that it does not contribute to the underlying dimensions of the data. The second is that it has a little influence on the output of the model.

Methods of condensation
There are two basic models: component analysis and common factor analysis. Components are real as they can be directly derived from the data of a study. Common factors are hypothetical because they are estimated from the data. In principal components, unities are placed in the diagonal, the separation of common and unique variance is imperfect as the variables themselves, including their uniqueness, determine the factors. Common factor analysis attempts to separate out unique variance. As unique variance (noise) has no scientific interest it is clearly useful to separate out common and unique variance. Further the fact that the factors are hypothetical rather than real is an advantage as a factor may account for the correlations among variables without being entirely defined by them.

Method of computing factor scores
Factor scores can be generated and used within the neural network in lieu of the raw data. This would reduce the number of inputs required by the neural network, and hence increase the accuracy to which the network may be trained. Factor scores are estimates of the scores subjects would have received on each of the factors had they been measured directly. There are three commonly used methods for computing factor scores: the regression method, Bartlett Method and Anderson-Rubin method. Tabachnick and Fidell (1989) recommend the use of the Anderson-Rubin approach when uncorrelated scores are required and otherwise the regression approach as it better understood and widely available.

Cluster Analysis

Alternatively, if on investigation of the variables it is discovered that they do not conform sufficiently to normality then cluster analysis will be applied. Cluster analysis is a non-parametric alternative to factor analysis, which attempts to identify relatively homogeneous groups of variables. Unlike factor analysis, the variables can include interval, count and binary data. Again, as with factor analysis new variables based on cluster membership can be generated for inclusion within the neural network.

IMPLEMENTATION

One of the features of early stage cost models which have been proposed in recent years is that the industry has not made extensive use of them. Fortune (1995) showed that newer techniques, such as expert systems and statistical models, have not received widespread use.

The model will serve no purpose if it is not implemented. It is therefore necessary to observe the reasons for previous models' not having achieved widespread use, and to address the issues raised.

Confidence

Perhaps the principal reason for the rejection of such techniques in the past has been that the industry has not been sufficiently confident in the results from these models. Fortune identified two reasons for this. The first is ignorance: estimators did not have any understanding of the technique. The second is that estimators are suspicious of any technique which does not allow them to use their professional judgement. As many of these techniques aim to be objective, they do not allow for any such judgement.

In order to be able to subjectively evaluate the results of such a model, and hence input professional judgement, the estimator must know how much confidence he may place in the results. This can be addressed in two ways.

The first is by providing the user with a quantitative indication of how much confidence may be placed in the results. In the first instance, this could be provided by the results of the testing of the model using measures of fit such as R^2 or confidence limits. This would provide the user with a rough indication of how much the cost of a project may be expected to vary from the initial estimate provided by the model.

The second and more useful would be for the neural network to be stochastic, rather than deterministic. Such a network would generate a cost distribution with mean and standard deviation, rather than a single value. This would not only provide the user with an indication of the cost, but also the cost risk associated with the project. While techniques exist for creating this type of network, it is likely that such a network will be developed in a third or subsequent phase of the research.

What if?

Another reason for past models not having been implemented is that, in the past, users have simply been given a “black box” model. The industry is quite rightly very suspicious of such models which have not been proven in a real life situation, and whose role in the industry is not fully understood.

The value of this predictive model as an early stage cost estimator will need to be proved in a real situation before it will be accepted. After such a proving the model will also be more robust as more data will have been made available to train the neural network. Despite the need to be proved as a predictive model, however, the model will still be applicable at he end of this phase of the research.

It is important to recognize the difference between the model generating a satisfactory cost forecast, and predicting relative differences in cost. Predicting the final cost of the project is difficult. The neural network must learn all the subtle differences between the behaviours of different of projects. However, in order to generate a comparative estimate, the neural network is only required to learn more general characteristics of projects: whether changes to different characteristics affect the project positively or negatively, and by how much. Thus whether the final value obtained is the actual cost of that project is not as critical.

Therefore, even if the model is not sufficiently accurate to operate as an accurate predictive model of the actual cost of a project, the model could still provide an objective indication of the comparative costs of different options. Used in this way, the model would provide a “What if?” analysis. The “What if?” analysis enables the estimator to assess how changing certain characteristics of the building affects the project cost. For example, it would be able to evaluate the cost of increasing the floor area, adding air conditioning, or changing the frame type. This is, of course, the primary motivation for the research, as this comparative capability will allow the comparison of the cost of different procurement routes.

Conclusions

No techniques currently exist which compare the total cost of different procurement routes. In order to address this problem a predictive model is under development at UMIST which uses a neural network to estimate the total construction cost to the client of a building. The variables have been identified and data is currently being collected.

In order to ensure that the model is as accurate a representation of reality as possible, the representation of the cost significant variables to the neural network has been considered. Additionally, it has been shown how factor analysis may be used to reduce the inputs to the neural network and hence increase its accuracy further.

This predictive model will, once developed, be able to compare not only the cost of different procurement routes, but also other project strategic choices available to the client. Furthermore, once tested in industry and provided with more data there is no reason why this model should not become a well used robust model for early stage cost estimation.

REFERENCES

Brandon, P.S. *et al.* (1988) *Expert systems: the strategic planning of construction projects.* London: RICS.

Building Economic Development Committee (1974) *Before you build: what a client needs to know about the construction industry*. London: NEDO.

C.C.M.I. (1985) Survey on management contracting. London: C.C.M.I.

Child, D. (1990) *The essentials of factor analysis*. London: Cassell.

Comrey, A.L. (1973) *A first course in factor analysis*. New York: Academic Press.

Department of Industry and Department of Environment (1982) *The United Kingdom construction industry: a guide to methods of obtaining a new industrial building in the UK*. London: Invest in Britain Bureau.

Duff, R., Emsley, M., Gregory, M., Lowe, D. and Masterman, J. (1998) Development of a model of total building procurement costs for construction clients. *In:* Hughes, W. (ed.) *Procs. 14th annual ARCOM conference*. University of Reading, 9-11 September. Reading: ARCOM. **1**: 210–218.

Franks, J. (1983) Using the design and build system. *Building Trades Journal*. p. 14.

Fortune, C. and Lees, M. (1995) Strategic cost modelling and the role of judgement. *In: Procs RICS construction and building research conference*, Heriot-Watt University, Edinburgh, 107–113.

Kline, P. (1993) *The handbook of psychological testing*. London: Routledge.

Manly, B.F.J. (1986) *Multivariate statistical methods: a primer*. London: Chapman and Hall.

Masterman, J.W.E. (1994) *A study of the bases upon which clients of the construction industry choose their building procurement systems*. Unpublished PhD thesis, UMIST, Manchester.

Moore, R.F. (1984) *Response to change: the development of non-traditional forms of contracting*. Occasional paper No 31. Ascot: Chartered Institute of Building.

Olashore, O.B. (1986) *Management contracting and construction management: a comparative analysis*. Unpublished MSc dissertation, Brunel University.

Reading University (1991) *Construction Management Forum, Report and Guidance*. Reading: Centre for Strategic Studies in Construction.

Reading University (1996) *Designing and building a world-class industry*. Reading: Centre for Strategic Studies in Construction.

Rowlinson, S. and Langford, D. (1986) The contractor as designer. Paper to IASBE workshop: *Conference on the organization of the design process*, Zurich.

Tabachnick, B.G. and Fidell, L.S. (1989) *Using multivariate statistics*. 2ed. New York: Harper Collins.

Tarassenko, L. (1998) *A guide to neural computing applications.* London: Arnold.