DEALING WITH CONSTRUCTION COST OVERRUNS USING DATA MINING

Citation:

Ahiaga-Dagbui, D.D and Smith, S.D (2014) Dealing with construction cost overruns using data mining, *Construction Management Economics* (***accepted June 2014, in press***).

Corresponding email: d.ahiaga-dagbui@ed.ac.uk or domdagbui@yahoo.com

DEALING WITH CONSTRUCTION COST OVERRUNS USING DATA MINING

One of the main aims of any construction client is to procure a project within the limits of a predefined budget. However, most construction projects routinely to overrun their cost estimates. Existing theories on construction cost overrun suggest a number of causes ranging from technical difficulties, optimism bias, managerial incompetence and strategic misrepresentation. However, much of the budgetary decision making process in the early stages of a project is carried out in an environment of high uncertainty with little available information for accurate estimation. Using nonparametric bootstrapping and ensemble modelling in artificial neural networks, final project cost forecasting models were developed with 1600 completed projects in this experimental research. This helped to extract information embedded in data on completed construction projects, in an attempt to address the problem of dearth of information in the early stages of a project. 92% of the 100 validation predictions were within ±10% of the actual final cost of the project whiles 77% were within ±5% of actual final cost. This indicates the model's ability to generalise satisfactorily when validated with new data. The models are being deployed within the operations of the industry partner involved in this research to help increase the reliability and accuracy of initial cost estimates.

Keywords: artificial neural networks, bootstrapping, cost overrun, data mining, ensemble modelling.

INTRODUCTION

The main concern of a construction client is to procure a facility that is able to meet its functional requirements, of the required quality, and delivered within an acceptable budget and timeframe. The cost aspect of these key performance indicators would seem to rank highest most times, especially in difficult financial periods such as the present. The estimates prepared at the initial stages of the project can play several roles - they can form the basis of cost benefit analysis, for selection of potential delivery partners, to support a to-build-or-not-to-build decision, and very often as a benchmark for future performance measure. As suggested by Kirkham and Brandon (2007) therefore, effective cost estimation must relate the design of the constructed facilities 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 limits of expenditure. This stage in a project life-cycle is particularly crucial as decisions made during the early stages of the development process carry far more reaching economic consequences than the relatively limited decisions which can be made later in the process. Effective cost estimation is therefore so vital, it can seal a project's financial fate, Nicolas (2004) notes.

However, in spite of the importance of cost estimation, it is undeniably neither simple nor straightforward because of the lack of information in the early stages of the project, Hegazy (2002) observes. Many projects consistently fail to meet initially set cost limits due to a number of causes ranging from the inability to accurately identify and quantify risk (Akintoye 2000), error in estimation (Jennings 2012), design changes and scope creep (Odeck 2004, Love *et al.* 2011) and even suspicions of foul-play and corruption (Wachs 1990, Flyvbjerg *et al.* 2002).

Developments in the business landscape however suggest a growing recognition of information as a key competitive tool. A vast amount of data is continuously generated by construction business transactions. As per due diligence or contractual requirements, most construction firms maintain copious information on each project undertaken. The amount of data generated by these firms presents both a challenge and opportunity - a challenge to traditional methods of data analysis since the data are often complex, and of course, voluminous. On the other hand, construction firms stand a chance of gaining competitive edge and performance improvement by making their data work for them using detailed data mining. Fayyad *et al.* (1996) noted that the real value of storing data lies in the ability to exploit useful trends and patterns in the data to meet business or operational goals as well as for decision support and policy making. Advances in the fields of data warehousing, artificial intelligence, statistics, visualisation techniques and machine learning now make it possible for data to be transformed into valuable asset by automating laborious but rewarding knowledge discovery in databases.

Data mining, simply described here as the analytical process of knowledge discovery in large databases, has found extensively application in industries such as business (Cf. Apte *et al.* 2002) and medicine (Cf. Koh and Tan 2011). However, discussions with a number of construction companies during this research suggest that very few take advantage of the data available to them to develop business decision support tools. At best, their analysis is usually limited to basic sample statistics of averages or standard deviations. Against this backdrop, we collaborated with a major UK water infrastructure provider to investigate the use of data mining techniques to develop cost models that can be applied during the early estimation stages for more reliable cost forecasting. As already pointed out, a lack of information for reliable estimation has been identified as one of the main causes of cost growth in construction. It is hoped that data mining might help to convert historical data on projects into decision support systems, to partly address the problem of insufficient information for reliable estimation at early stages of a project. The problem of cost growth and its causes are examined in the next section of the paper, followed by an overview of data mining and its applications. The data mining methodology was then applied to the problem of cost estimation in the construction industry using Artificial Neural Networks (ANN). Some practical implications of the research have been identified in the conclusions along with some possible barriers to effective data mining within the construction industry.

COST OVERRUNS

Chan and Chan (2004) conducted a critical analysis on existing literature on construction benchmarking and proposed a framework of both qualitative and quantitative descriptors to evaluate the success of a construction project. They validated their framework using three hospital projects and noted that cost performance on a construction project remains one of the main measures of success even though there were other emerging qualitative measures like health and safety and environmental performance. We have previously investigated cost overruns on construction projects as part of a wider research into the potential use of artificial neural networks for construction cost estimation (A and B 2012). We attempted to model final cost using non-traditional cost factors such as project location, access to site and procurement method. It became obvious that estimating the final cost of projects can be extremely difficult due to the complex web of cost influencing factors that need to be considered. For a thorough and reliable estimate of final cost, the estimator has to be able to take into consideration factors such as the type of project, likely design and scope changes, risk and uncertainty, effect of policy and regulatory conditions, duration of project, type of client, ground conditions or tendering method. Trying to work out the influence of most of these variables at the inception stage of a project when cost targets are set, can be an exhaustive task, if not at all futile. Ignoring most of these factors altogether creates a recipe for eventual cost growth, disputes, law suits and even project termination in some cases. Jennings (2012) employed a longitudinal 'process tracking' approach to examine the dynamics between risk, optimism and uncertainty in construction and how these interact with the phenomenon of cost overruns using a case study of the 2012 London Olympic games. Jennings found that a high level of uncertainty surrounds the cost estimation exercise especially in the initial stages of the project, thus making it difficult to produce reliable cost estimates. What is then resorted to, in most cases, is the use of some arbitrary percentages, the so-called contingency funds, which unfortunately has mostly failed to keep construction projects within budget.

The on-going Edinburgh Trams project in Scotland for example has already far exceeded its initial budget leading to significant scope reduction to curtail the ever-growing cost (Miller 2011, Railnews 2012). The 2012 London Olympics bid was awarded at circa £2.4 billion in 2005. This was adjusted to about £9.3 billion in 2007 after significant scope changes. The project was eventually completed at £8.9 billion in 2010 (Cf. National Audit Office 2012). The UK Government commissioned report in 1998 on construction industry performance indicated that over 50% of projects overspent their budget (Egan 1998). A similar report around the same time in the US suggested that about 77% of projects exceed their budget, sometimes to the tune of over 200% (General Accounting Office 1997). In more recent years, Flyvbjerg *et al.* (2002) sampled 258 infrastructure projects worth US$90 billion from 20 different countries and found that 90% of the projects experienced budget escalation and that infrastructure projects in particular have an 86% likelihood of exceeding their initial estimates. Alex *et al*. (2010) report up to 60% discrepancy between actual and estimated costs of over the 800 water and sewer projects examined in their research. Flyvbjerg *et al.* (2004) thus concluded that little learning seemed to be taking place within the industry over time.

Cost growth in the construction industry has been attributed to a number of sources including technical error in design or estimation, managerial incompetence, risk and uncertainty, suspicions of foul play, deception and delusion, and even corruption. Akintoye and MacLeod (1997) conducted a questionnaire survey of general contractors and project managers in the UK construction industry to ascertain their perception of risk and uncertainty as well as their use of various risk management techniques. They concluded that risk management practice was largely experience and judgement based and that formal risk management techniques such as Monte Carlo simulation or stochastic dominance were seldom used due to doubts on their suitability and lack of knowledge and understanding of these methods. Hitherto, the industry still seems to struggle to deal with identifying and quantifying the impact of risk events. This may probably be due to the nature of the industry- it is fragmented, complex, each project spans several years, is constructed in an environment open to the weather inclement and has many different parties with varying business interests. Flanagan and Norman (1993) suggest that the task of risk management in most cases is so poorly performed, that far too much risk is passively retained, an assertion supported by Jennings' (2012) recent case study of the possible sources of cost growth on the 2012 London Olympic project.

Flyvbjerg *et al.* (2002, 2005), as well as Wach (1989, 1990) point to optimism bias and strategic misrepresentation, or delusion and deception in other words, as possible causes of cost growth particularly on large publicly funded projects. Flyvbjerg *et al* (2002) conducted a desk study analysis of the cost performance of 258 transportation projects worth US$90 billion and categorised the sources of cost overruns on construction projects into four groups: technical (error), psychological, economical and political. They concluded that cost escalation could not be adequately explained by estimation error, but more likely by strategic misrepresentation- an intentional attempt to mislead. They observed that 9 out 10 of the projects experienced significant cost escalation over their construction period and that there was evidence of a systematic bias in the cost estimates as the overruns experienced did not appear to be randomly distributed. Flyvbjerg *et al* controversially concluded that the cost estimates used to decide whether projects should be given the go-ahead were 'highly and systematically misleading', strongly suggesting foul play by project promoters.

Further developments of the strategic misrepresentation perspective by Flyvbjerg led to theories based on optimism bias, after Weinstein (1980). Optimism bias can be explained as the cognitive disposition to evaluate possible negative future events in a fairer light than suggested by inference from the base rates. Flyvbjerg (2007) draws on this concept and suggested that decision making in policy and infrastructure planning is flawed by the planning fallacy that we know, or at least are in control of all possible chain of events from project inception to completion, thereby leading to unjustifiable confidence in the prospects of the project and unrealistic estimates. While strategic misrepresentation is often intentional, according to Flyvbjerg *et al*, optimism bias is not. They make this distinction between the two concepts with the terms 'deception' and 'delusion' respectively (Flyvbjerg 2007). It might be easy to reckon how strategic misrepresentation and optimism bias work in tandem with business competition embedded in the lowest-bidder culture to often create an unrealistic low cost target of projects at the pre-construction phase of projects.

Another school of thought on cost overruns, referred to here as the Evolution Theorists, include Love *et al* (2011) as well as Gil and Ludrigan (2012). They argue that projects essentially evolve significantly between conception and completion so that it might be misleading in most cases to make a direct comparison between the costs at start and end of the project. Their thesis statement is straightforward- projects change, and when they do, they often come with increasing costs. Love *et al.* (2011) provide a rebuttal to Flyvbjerg's perspective on cost overruns, instead suggesting that the industry 'move beyond strategic misrepresentation and optimism bias' to embrace a more holistic understanding of the phenomenon that includes some level of the process and the social construct. They introduce the concept of 'pathogens' for example, the many events and actions that could not be accounted for at the initial stages of the project that eventually add-on to expected cost as the main drivers of cost growth. They further argue that Flyvbjerg's analyses are maybe too simplistic and not generalisable to all projects undertaken within the industry. Their argument would seem sustainable, especially on small, privately funded projects that do not have strong political or public interest. Besides, it is difficult to draw valid distinctions, along a continuum of motivation, from reasonable and justifiable optimism, through over-confidence and delusion, culpable error, to deliberate deceit using just statistical analysis, the method adopted in Flyvbjerg's works. Ahiaga-Dagbui and Smith (2014) provide a more detailed discussion on other possible causes of overruns including technical and managerial difficulties and poor estimation, as well as the dynamics between cost growth and cognitive dispositions such as Prospect Theory (Kahneman and Tversky 1979) and Kruger-Dunning effects (Kruger and Dunning 1999).

It is important to note here though that much of the current literature and media furore on cost growth seem to over simplify the rather complex causes of overruns on construction projects. As already noted, most construction projects, especially publicly funded capital intensive projects tend to go through a long gestation period after project conception during which many changes to scope and accompanying costs occur- sometimes the initial scheme bears little resemblance to the constructed asset. The estimated cost at project inception often fails to take into consideration a lot of details and information, largely because much of these are not yet available or uncertain. It seems erroneous therefore to make a direct comparison between the initial 'estimate' A (£40million on the Scottish Parliament, for example) and its final completion cost B (circa £414million) - the two schemes were very different (Cf. Audit Scotland 2004, Fraser 2004). More robust explanations of growth perhaps need to factor-in process and product, as well as sources of changes to scope. Flyvbjerg's works make a direct comparison between costs A and B, and wherever B>A, overruns are reported. It might be simplistic though, as pointed out by Love *et al.*, but probably justifiable as estimate A is usually the estimate used to get project approval when publicly funded projects are being appraised. It is important to bear in mind that it often practically difficult to discontinue a project once considerable amount of money has already been spent to get it started. This is referred to as the sunk cost effect by Arkes and Blumer (1985) or the phenomenon of escalation of commitment by Staw (1981). It may therefore be crucial for the industry to find more effective ways of project approval that better deals with underestimation of true cost and the setting of unrealistic cost targets.

Alex *et al* (2010) reviewed the cost performance on more than 800 construction projects of the Canada's Drainage and Maintenance Department and observed a discrepancy of up to 60% between estimated and actual final cost of projects completed between 1999 and 2004. They partly attributed this problem to the fact that the Department's estimation process was heavily experienced based, relying largely on professional judgement, just as observed by Akintoye and MacLeod (1997). A potential downside of experienced-based estimation is the difficulty in thoroughly evaluating the complex relationships between the many cost influencing variables already identified in this paper, or its inability to quickly generate different cost alternatives in a sort of what-if analysis. Furthermore, as noted by Okmen and Öztas (2010) in their research on cost analysis within an environment of uncertainty, traditional cost estimation i.e. the estimation of the cost of labour and materials and making allowance for profits and overheads for individual construction items, is deterministic by nature. It therefore largely neglects and poorly deals with uncertainties and their correlation effects on cost, thereby deemed inadequate in reaching a reliable and realistic final cost. As an alternative to traditional estimation approaches, data mining, using the learning and generalisation algorithms within artificial neural networks in combination with statistical bootstrapping and ensemble modelling is used to develop final cost models in this paper. The aim here is an attempt at circumventing the problems posed by uncertainty and lack of information in estimation in the early stages of a project.

DATA MINING

Data mining, otherwise referred to as Knowledge Discovery in Databases (KDD), is an analytic process for exploring large amounts of data in search of consistent patterns, correlations and/or systematic relationships between variables, and to then validate the findings by applying the detected patterns to new subsets of data (StatSoft Inc 2008). Data mining attempts to scour databases to discover hidden patterns and relationships in order to find predictive information for business improvement. Questions that traditionally required extensive hands-on analysis, experts and time, can potentially be quickly answered from a firm's existing data.

Goldberg and Senator (1998) report the use of pattern discovery techniques by the Financial Crimes Enforcement Network (FinCEN) of the United States Department of Treasury since 1993 to detect potential money laundering and fraudulent transactions from the analysis of about 200,000 large cash transactions per week. Using input factors such as age, housing, and job title and account balance, Huang *et al.* (2007) developed a support vector machine credit scoring model to assess loan applicant's credit worthiness in an attempt to limit a financing firm's exposure towards default. Hoffman *et al* (1997) have also explored the use of data visualization and mining techniques for DNA sequencing in the area of cell biology. Ngai *et al.* (2009) provide a comprehensive review of data mining applications in customer relationship management area, classifying these applications into four groups of customer identification, attraction, retention and development. One-to-one marketing and loyalty programs targeted towards customer retention seem to receive the most attention from researchers.

Although data mining is yet to find extensive application in practice within the construction industry, construction management researchers have been investigating its applicability to different problem areas. Using some of the concepts of data mining and the theory of inventive problem solving, Zhang *et al.* (2009) developed a value engineering knowledge management system (VE-KMS) that collects, retains and re-uses knowledge from previous value engineering exercises in an attempt to streamline future exercises, making them more systematic, organised and problem-focussed. Cheng *et al* (2012) also developed EFSIMT, a hybrid fuzzy logic, support vector machines and genetic algorithm inference model to predict the compressive strength of high performance concrete using input factors such as the aggregate ratio, additives and working conditions. This kind of model allows for a more reliable prediction of the strength of a particular mix for design and quality control purposes as concrete strength is generally affected by a lot of factors. There is generally a higher rate of occupational injuries in the construction industry than industries like manufacturing for example (Cf. Larsson and Field 2002). This might possibly be because of the dynamic and hazardous environment of a typical construction site. Liao and Perng (2008) thus employ the use of association rule based data mining to identify the characteristics of occupational injuries reported between 1999 and 2004 in the Taiwan construction industry. Wet-weather related injuries and fatalities were particularly significant in their study.

Data Mining Process

Data mining normally follows a generic process of business and data understanding, data preparation, modelling proper, evaluation of models, and deployment. It starts with the selection of relevant data from a data warehouse that contains information on organisation and business transactions of the firm. The selected data set is then pre-processed before actual data mining commences. The pre-processing stage ensures that the data are structured and presented to the model in the most suitable way as well as offer the modeller the chance to get to know the data thoroughly. Pre-processing typically involves steps such as removing of duplicate entries, sub-sampling, clustering, transformation, de-noising, normalisation or feature extraction.

The next stage involves the actual modelling, where one or a combination of data mining techniques is applied to scour down the dataset to extract useful knowledge. This process can sometimes be an elaborate process involving the use of competitive evaluation of different models and approaches and deciding on the best model by some sort of bagging system (StatSoft Inc. 2011). Table 1 provides a framework for selecting a particular data mining technique. The type of modelling technique adopted depends on a number of factors, including the aim of the modelling exercise, the predictive performance required and the type of data available. Each modelling technique can also be evaluated in terms of its characteristics. For example, regarding 'interpretability', while regression models generate an equation whose physical properties can be easily interpreted in terms of the variables used in explaining the phenomenon under study (Hair *et al.* 1998). Neural networks, on the other hand, do not produce any equation and have thus been derided as 'black-boxes' by some researchers including Sarle (1994). However their power and ability to model complex non-linear relationships between predictors make them particularly desirable for hard-to-learn problems and where *a priori* judgements about variable relationships cannot be justified (Adeli 2001).

The results from the data mining stage are then evaluated and presented into some meaningful form to aid business decision making. The knowledge generated is then validated by deploying the model in a real life situation to test the model's efficacy.

Table 1- Framework for selecting a data mining technique

Data mining category	**Data mining requirement**	**Data mining technique**	**Technique characteristics**
• Regression • Clustering • Classification • Visualisation • Summarisation	• Prediction • Pattern discovery • Surveillance • Performance • Measurement • Business • Understanding	• Regression • Support Vector Machine (SVM) • Self-Organising maps • Genetic algorithm, etc	• Flexibility • Accuracy (Precision) • Power • "Interpretability" • Ease of deployment

DATA

The data mining process described in the previous section of this paper is now applied to cost estimation within a partnering major water infrastructure client in the UK. The aim here is in two folds - to develop decision support systems from existing data to complement the existing estimation process within our collaborating organisation and also to investigate ways of circumventing the problem of lack of information for reliable estimation at the early stages of a project. Many crucial business decisions have to be made at this stage including tender evaluations, contract award, project feasibility or securing loans to finance the project. Our collaborating organisation typically has three stage of estimation before inviting bids from contractors. The third stage estimate, Gate Three, is usually based on about 50-60% completed scope design and is used for evaluation of tenders after which detailed design is carried out by the selected contractor in a sort of design and build contract framework. The estimates produced by the models developed in this paper thus allows the organisation to forecast its total likely commitment before tendering and before definitive estimates are available.

The data collection process involved an initial shadowing of the tendering and estimation procedure within the organisation. We were thus allowed to be quasi members of the tendering team of the company on some of its projects to observe how the estimates were produced. It was also an opportunity to gain a first-hand understanding of how the data to be used of the modelling was generated and what different variables meant. The initial dataset contained over 5000 projects completed between 2000 and 2012. The scope of these projects varied from construction of major water treatment plants to minor repairs and upgrade. Project values ranged from a mere £1000 to £30 million and durations from 3 months to 5 years. The initial analysis involved drilling down into the database to find what might be useful in modelling final cost. To ensure some level of homogeneity in the data, K-means cluster analysis was used to create clusters of project cases based on duration and cost. V-fold cross-validation with Mahalanobis distance was used to search for optimum number of clusters between 2 and 10 clusters. This distance measure was preferred to the popular square Euclidean distance because it helps account for the variance of each variable as well as the covariance between cost and duration of the project cases. The cases to be used in the modelling also had to be without significant missing data and somewhat representative of the entire dataset. One of the clusters containing about 1600 projects completed between 2004 and 2012 was used for the models reported in this paper. One hundred of these project cases were selected using stratified random sampling with cost as the strata variable to be used for independent second stage validation of the final models. Stratified random sampling was used because this would hopefully allow for the selection of cases that are representative of the entire range of possible cases within the dataset. The remaining data was then split in a 70:15:15% ratio for training, testing and first stage validation respectively. Further details on the dataset used for the modelling is found in Table 2.

Table 2- Overview of data used for model development

Size	*Types of project*	*Type of organisation*	*Cost range*	*Duration range*	*Year span*
c.1600	Water mains, manholes, combined sewer	Client	£4000 to	1 month to	2004 to

	overflows, repairs, upgrades		£15million	5 years	2012

Data Pre-processing

The pre-processing stage ensures that the data are structured and presented to the model in the most suitable way as well as offer the modeller the chance to get to understand the data thoroughly. Cost values were normalised to a 2012 baseline using the infrastructure resources cost indices by the Building Cost Information Services with a base year 2000. This allowed for cost values to be somewhat comparable across different years. Numerical predictors were further standardized to *zScores* using

$$zScore = \frac{x_i - \mu}{\sigma} \qquad \text{Equation 1}$$

Where: *zScore* is the standardized value of a numerical input, x_i

µ is the mean of the numerical predictor

σ is the standard deviation of the numerical predictor

Since neural networks was to be used for the actual modelling exercise, standardizing either input or target variable into a smaller range of variability would potentially aid the effective learning of the neural net whiles improving the numerical condition of the optimization problem (StatSoft Inc 2008). If one input has a range of 0 to 1, while another has a range of 0 to 30 million, as was the case in the data that were used in this analysis, the net will expend most of its effort learning the second input to the possible exclusion of the first. All categorical variables were coded using a binary coding system.

The next stage involved deciding which predictors to use in the modelling exercise. It was easy to remove predictors such as project manager, project ID or year of completion from the set of predictors on precursory examination as they were likely not to be good predictors when the model is used in practice. Table 3 contains details on the set of initial predictors used at the beginning of the modelling.

Table 3 - Initial list of variables for model development

		Category			
	Type of data	**1**	**2**	**3**	**4**
	Project Information				
1	Tendering Strategy	Open competitive	Selective competitive	Negotiated	Serial
2	Procurement Option	Design-bid-build	Design and build	Management types	Partnering
	Site Information				
3	Ground Condition	Contaminated	Non-contaminated	Made-up	-
4	Type of Soil	Good	Moderate	Poor	-
	Other Information				
5	Delivery Partner*	X	Y	Z	-
6	Scope of Project	New-build	Upgrade	Refurbishment	Replacement
7	Purpose of Project	Wastewater	Water	General	-
8	Operating Region	North	South	East	West

1. *Other factors include project duration (months) and awarded target cost (£). Model output was final cost at completion (£).*
2. **indicated as X, Y and Z for confidentiality reasons*

COST MODEL DEVELOPMENT

Data visualisation using scatter and mean plots in the earlier stages of the modelling suggested non-linear relationships between most of the variables and final cost. Also, most of the predictors are categorical, rather than the usual numeric type. It was thus decided to use Artificial Neural Networks (ANN) for the actual modelling because of their ability to cope with non-linear relationships and categorical variables (Cf. Anderson 1995). ANN is an abstraction of the human brain with abilities to learn from experience and generalise based on acquired knowledge (Moselhi *et al.* 1991). It is also able to cope with multicollinearity, a statistical condition where two or more variables are highly correlated or dependent on each thereby resulting in spurious predictions when both of those variables are included in the model (Marsh *et al.* 2004). Neural networks has previously been applied to forecasting tender price (Elhag and Boussabaine 1998, Emsley *et al.* 2002) and for quantification of risk in construction by McKim (1993). See Moselhi *et al.* (1991) for a review of neural network application in construction management research.

Standard Models

The cost models were developed using an iterative process of fine-tuning the network parameters and inputs until acceptable error levels were achieved or when the model showed no further improvement. The model training began with a search for optimal model parameters. This was done in a trial and error manner to begin with, training several networks and examining them for possible performance improvement using the input factors in Table 3 and Cost at Completion as model output. Two different network architectures, the Multilayer Perceptron (MLP) and the Radial Basis Function (RBF), were experimented at this stage. 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 2008). MLPs on the other hand model using just the second stage of the RBF. The MLP models were superior to the RBF networks and so the rest of the modelling was carried out using just MLPs. It was found that the best trial results were achieved with MLPs with a single hidden layer having between 3-10 nodes.

Consequently, using a custom range of 3-10 hidden nodes in 1 hidden layer, a dataset size split of 70% for training, 15% for testing and another 15% for first stage validation, 1000 networks were trained, retaining the best 10 performing networks for further examination. These 10 networks were selected based on their overall performance, measured using the correlation coefficient between predicted and output values as well as the Mean Sum of Squares (MSE) of errors. MSE is defined here as:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (O_i - T_i)^2 \qquad \text{Equation 2}$$

Where: *Oi* is the predicted final cost of the *i*th data case (Output); T_i is the actual final cost of the *i*th data case (Target) and n is the sample size.

The higher the MSE 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. The higher the *p*-value, the less reliable the observed correlations. The best 10 retained networks were then further validated using the 100 independent validation cases that were selected using the stratified sampling at the beginning of the modelling exercise.

Five different activation functions, i.e. identity, logistic, tanh, exponential and the sine functions were iterated in both hidden and output layers, using gradient descent, conjugate descent and Quasi-Newton (BFGS) training algorithms. See Fausett (1994) and Gurney (1997) for the fundamentals for neural network architectures, algorithms, or Skapura (1996) for a practical guide to developing neural network models. Early stopping, the process of halting training when the model 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.

Redundant predictors, those that do not add new information to the model because they basically contain the same information at another level with other variables, were detected using spearman correlations, bi-variate histograms or cross-tabulation. These were tendering strategy, procurement option and type of soil. This is likely due to the invariant nature of these predictors as most of the project were procured through design and build contracts with a mix of open-competitive and negotiated tendering strategies. Type of soil was found to be linearly dependent on ground condition, thereby not making any additional contribution to the model's output.

All the best 10 models identified at this stage had 12 input nodes in the input layer, between 3 and 7 nodes in a single hidden layer with one output, i.e. final cost. They had either a tanh or logistic activation function between their input and hidden layers with an identity transfer function in the output layer.

Bootstrapping

Bootstrapping is a general technique, attributed to Efron (1992), for estimating sampling distributions that allow for treating the observed data as though it were the entire (discrete) statistical population (StatSoft Inc. 2011). It provides an avenue for using subsamples from a sample in a manner that addresses the variability and uncertainty in statistical inferences. Traditional approaches to statistical inference are based on the assumption of normality in the data distribution. This is reasonable and largely accepted but where this assumption is wrong, Efron (1992) warns that the corresponding sampling distribution of the statistic may be seriously questionable. In contrast, nonparametric bootstrapping provides a way to estimate a statistic of population without explicitly deriving the sample distribution. During the development of the models presented so far, the dataset was divided into three subsets for training, testing and validation. On a closer examination, this might be regrettable, as not all the data gets used for training, testing or validation, and therefore some level of information within the entire dataset is lost in the learning process. If bootstrapping is employed, a different split of data is used each time for training or testing so as to glean as much information as possible from the entire dataset.

Statisticians disagree though on the number of bootstrap samples (BS) necessary to produce reliable results. Most textbooks suggest choosing a sufficiently large enough bootstrap sample size without specific guidance on an optimum size. Efron and Tibshirani (1993), as well as Pattengale *et al.* (2009) however suggest that an minimum of 100 or a maximum of 500 BS is generally sufficient in most cases. Bootstrapping was thus applied to the dataset in this manner - 600 different training, validation, testing BS sample sets were generated by perturbing the entire dataset for each model using sampling *with* replacement over a uniform probability distribution. This should ensure that as many data cases as possible get used in the training, validation or testing samples sets. With the same inputs, neural network architectures, activation functions, hidden layers and nodes used in the case of the standard sample models developed in the previous section, 1000 neural network models were then trained and tested, retaining the best 10 performing models just as before. The 10 retained models were then further validated using the 100 separate validation cases just as was done previously.

Figure 2 shows the performance of the best 10 models from both the standard and bootstrapped models validated with the 100 validation cases. It is obvious that bootstrapped models far outperform the standard models. Whiles the bootstrapped models overestimated actual final cost by about 4% on average, the standard models overestimated by 8.35% on average. Furthermore, the bootstrapped models underestimated actual final cost with an average error of about -6%, whereas the standard models averaged about -10%. This performance improvement is likely due to the fact that by using the 600 bootstrapped sample sets, the models were afforded a wider learning space than the standard models. The bootstrapped models were then carried forward for further analysis discussed below.

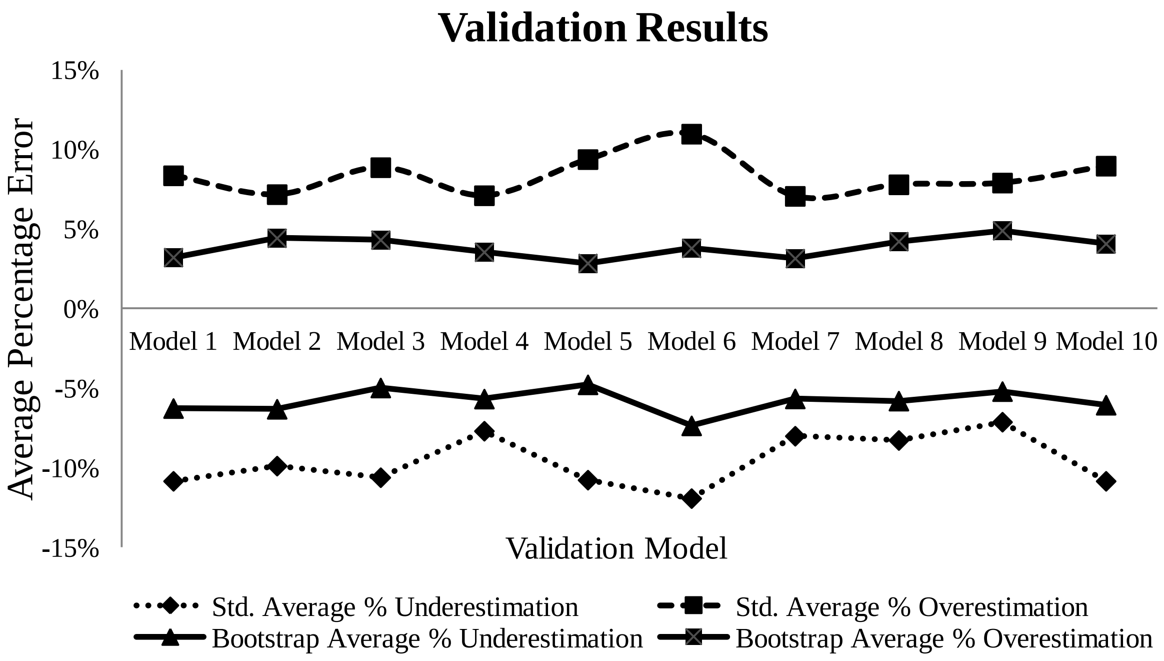

Figure 1 - Validation Results (Standard Models vs Bootstrapping)

Ensemble Network

All modelling techniques are prone to two main types of error, bias and variance, largely because models essentially try to reduce complicated problems into simple forms and then attempt to solve the ‘reduced’ problem using an imperfect finite dataset. Bias is the average error any particular model will make across different datasets whereas variance reflects the sensitivity of the model to a particular choice of dataset (StatSoft Inc. 2011). The use of ensembles can improve the results that are produced from individual models by combining them in a way that achieves some sort of compromise between variance and bias. Also known as Committee Methods (Cf Oza 2006), ensembles attempt to leverage the power of multiple models to achieve better prediction accuracy than any of the individual models could on their own. It is perhaps a way of consulting a 'committee of several experts', the 10 different bootstrapped models in this case, before reaching a final decision either by averaging, voting or by 'winner-takes-all', whichever is most appropriate (see Jordan and Jacobs 1994, Breiman 1996). The result, at least in theory, is a model (the ensemble) that is more consistent in its predictions and on average, at least as good as the individual networks from which it was built. A weighted average algorithm was thus applied to combine the 10 best bootstrapped models to trade off bias and variance to improve performance.

Table 4 compares the performance of the ensemble model with the bootstrapped models and the standard models. It is obvious that significant improvement has been achieved by applying the ensemble technique to the 10 bootstrapped models.

Table 4- Summary of results (Standard, Bootstrapped & Ensemble Models)

Model	*Average percentage error*	
	Overestimate	Underestimate
Standard models	+8.35%	-9.6%
Bootstrapped models	+3.84%	-5.81%
Ensemble model	+2.33%	-3.83%

In Table 5, details of a sample of 20 results out of the 100 validation cases used to test the ensemble model are highlighted. It shows a comparison between the ensemble final cost prediction and the actual final cost of the project, with a measure of the actual monetary error observed.

Table 5- Sample results from ensemble model validation

Case	Actual final cost (£,000)	Ensemble prediction (£,000)	Ensemble error (£,000)	*Ensemble absolute % error (£)*
1	4,846	4,990	(144)	2.97
2	1,586	1,590	(4)	0.25
3	24,986	23,760	1,226	4.91
4	11,143	10,934	209	1.88
5	5,328	5,765	(437)	8.20
6	3,787	3,723	64	1.69
7	17,346	16,967	379	2.18
8	4,136	4033	103	2.49
9	3,117	2994	123	3.95
10	1,000	939	61	6.10
11	1,773	1674	99	5.58
12	3,779	3600	179	4.74
13	209	192	17	8.13
14	3,960	3810	150	3.79
15	294	300	(6)	2.04
16	2,296	2220	76	3.31
17	2,104	2038	66	3.14
18	248	247	1	0.40
19	208	192	16	7.69
20	201	197	4	1.99

Table 6 shows a summary the performance of the ensemble model for all the 100 validation cases. 92% of the 100 validation predictions were within ±10% of the actual final cost of the project with 77% within a ±5% of actual final cost. Only 8 out of the 100 validation had predictions beyond ±10% of the final cost of the project case.

Table 6- Summary of validation performance of ensemble model

Percentage Error	*Number of cases*	*Percentage of total validation set*
Within ±5%	77	77%
±5% < x > ±10%	15	15%
Beyond ± 10%	8	8%
Total	100	100%

CONCLUSION

A lot of project and cost information is usually generated on any one particular construction project. If this is done in a meaningful and retrieval manner for a number of projects over time, a vast database of potentially valuable asset results. This can be converted into valuable decision support systems using data mining methodologies. The possibilities are that these decision support systems could help construction practitioners in making better informed and reliable decisions as well as reduce the time and resources spent in reaching these decisions.

Cost growth, attributed to a number of causes including the unavailability and uncertainty of necessary information for reliable estimation at the early stages of a project, remains one the major problems in the construction industry. We make a case for using data mining in modern construction management as a key business tool to help transform information embedded in construction data into decision support systems that can complement traditional estimation methods for more reliable final cost forecasting. Using a combination of nonparametric bootstrapping and ensemble modelling in artificial neural networks, cost models were developed to estimate the final construction cost of water infrastructure projects. 92% of the 100 validation predictions were within ±10% of the actual final cost of the project with 77% within ±5% of actual final cost. We are now exploring avenues of transforming the models into standalone desktop applications for deployment within the operations of the industry partner that collaborated in this research.

The models developed in this paper will be particularly useful at the pre-contract stage of the partnering construction firm in this research as it will provide a benchmark for evaluating submitted tenders. They could further allow the quick generation of various alternative solutions for a construction project using what if analysis for the purposes of comparison. The method and approach adopted to develop the models can be extended to even more detailed estimation so long as relevant data can be acquired. It must be pointed out that reliable cost planning and estimation forms only one aspect of dealing with cost growth in construction. A more holistic approach must include effective project governance and client leadership, accountability and measures of cost control. Also, an effective data mining exercise does depend heavily on both quantity and quality of data. Companies that want to employ data mining techniques thus have to be intentional in how they collect and store their data, making sure it contains relevant business and operational data to solve the problem at hand.

REFERENCES

A and B (2012) Reference details withheld for anonymity during refereeing.

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

Ahiaga-Dagbui, D D and Smith, S D (2014) Rethinking construction cost overruns: Cognition, learning and estimation. *Journal of Financial Management of Property and Construction*, **19**(1), 38-54.

Akintoye, A (2000) Analysis of factors influencing project cost estimating practice. *Construction Management & Economics*, **18**(1), 77-89.

Akintoye, A S and MacLeod, M J (1997) Risk analysis and management in construction. *International Journal of Project Management*, **15**(1), 31-8.

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, J A (1995) *An Introduction to neural networks*. Cambridge, Massachusetts: MIT Press.

Apte, C, Liu, B, Pednault, E P D and Smyth, P (2002) Business applications of data mining. *Communications of the ACM*, **45**(8), 49-53.

Arkes, H R and Blumer, C (1985) The psychology of sunk cost. *Organizational Behavior and Human Decision Processes*, **35**(1), 124-40.

Audit Scotland (2004) *'Management of Holyrood Building Project' (Audit Report prepared for the Auditor General of Scotland)*, Audit Scotland.

Breiman, L (1996) Bagging predictors. *Machine learning*, **24**(2), 123-40.

Chan, A P and Chan, A P (2004) Key performance indicators for measuring construction success. *Benchmarking: An International Journal*, **11**(2), 203-21.

Cheng, M-Y, Chou, J-S, Roy, A F V and Wu, Y-W (2012) High-performance concrete compressive strength prediction using time-weighted evolutionary fuzzy support vector machines inference model. *Automation in Construction*, **28**(0), 106-15.

Efron, B (1992) Bootstrap methods: Another look at the jackknife, *Breakthroughs in Statistics*, pp. 569-93: Springer.

Efron, B and Tibshirani, R (1993) *An introduction to the bootstrap*. Vol. 57, New York: Chapman and Hall.

Egan, J (1998) *Rethinking construction: the report of the Construction Task Force to the Deputy Prime Minister, John Prescott, on the scope for improving the quality and efficiency of UK construction*, London: Department of the Environment, Transport and the Regions Construction Task Force.

Elhag, T M S and Boussabaine, A H (1998) An artificial neural system for cost estimation of construction projects. In: *14th Annual ARCOM Conference*, Hughes, W, Ed., University of Reading: Association of Researchers in Construction Management, 219-26.

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.

Fausett, L V (1994) *Fundamentals of neural networks: architectures, algorithms, and applications*. Prentice-Hall Englewood Cliffs, NJ.

Fayyad, U, Piatetsky-Shapiro, G and Smyth, P (1996) The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM*, **39**(11), 27-34.

Flanagan, R and Norman, G (1993) *Risk Management and Construction*. Oxford: Blackwell Science Ltd.

Flyvbjerg, B (2005) Design by deception: The politics of megaproject approval. *Harvard Design Magazine*, **22**, 50-9.

Flyvbjerg, B (2007) Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, **16**(1), 3-21.

Flyvbjerg, B, Holm, M K S and Buhl, S L (2002) Understanding costs in public works projects: Error or lie? *Journal of the American Planning Association*, **68**(279-295).

Flyvbjerg, B, Holm, M K S and Buhl, S L (2004) What causes cost overrun in transport infrastructure projects? *Transport Reviews*, **24**(1), 3-18.

Fraser (2004) *Holyrood Enquiry- A report by the Rt Hon Lord Fraser of Carmyllie QC on the construction of the Holyrood Building Project presented to the First Minister and Presiding Officer* (SP Paper No. 205), www.holyroodinquiry.org: Scottish Parliamentary Corporate Body 2004.

General Accounting Office (1997) *Transportation infrastructure- managing the costs of large-dollar highway projects*, Washington DC: United States General Accounting Office (GAO).

Gil, N and Lundrigan, C (2012) The leadership and governance of megaprojects. In: *CID Technical Report No. 3/2012*: Centre for Infrastructure Development (CID), Manchester Business School, The University of Manchester, 18.

Goldberg, E G and Senator, T E (1998) The FinCEN AI System: Finding financial crimes in a large database of cash transactions. *In:* Jennings, N R and Woodridge, M J (Eds.), *Agent Technology: Foundations, Applications and Markets*. Berlin: Springer.

Gurney, K (1997) *An introduction to neural networks*. London: UCL Press.

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

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

Hoffman, P, Grinstein, G, Marx, K, Grosse, I and Stanley, E (1997) DNA visual and analytic data mining. *In, Visualization '97., Proceedings*, 24-24 Oct. 1997, 437-41.

Huang, C L, Chen, M C and Wang, C J (2007) Credit scoring with a data mining approach based on support vector machines. *Expert Systems with Applications*, **33**(4), 847-56.

Jennings, W (2012) Why costs overrun: risk, optimism and uncertainty in budgeting for the London 2012 Olympic Games. *Construction Management and Economics*, **30**(6), 455-62.

Jordan, M I and Jacobs, R A (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**(2), 181-214.

Kahneman, D and Tversky, A (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, **47**(2), 263-91.

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

Koh, H C and Tan, G (2011) Data mining applications in healthcare. *Journal of Healthcare Information Management—Vol*, **19**(2), 65.

Kruger, J and Dunning, D (1999) Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, **77**(6), 1121-34.

Larsson, T J and Field, B (2002) The distribution of occupational injury risks in the Victorian construction industry. *Safety Science*, **40**(5), 439-56.

Liao, C-W and Perng, Y-H (2008) Data mining for occupational injuries in the Taiwan construction industry. *Safety Science*, **46**(7), 1091-102.

Love, P E D, Edwards, D J and Irani, Z (2011) Moving beyond optimism bias and strategic misrepresentation: An explanation for social infrastructure project cost overruns.

Marsh, H W, Dowson, M, Pietsch, J and Walker, R (2004) Why Multicollinearity Matters: A Reexamination of Relations Between Self-Efficacy, Self-Concept, and Achievement. *Journal of Educational Psychology*, **96**(3), 518.

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

Miller, D (2011) Edinburgh Trams: Half a line at double the cost. *BBC, 18th February, 2013, http://goo.gl/mfr96*.

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

National Audit Office (2012) *The London 2012 Olympic Games and Paralympic Games: post-Games review* HC 794- Session 2012-13, National Audit Office, UK.

Ngai, E W T, Xiu, L and Chau, D (2009) Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, **36**(2), 2592-602.

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

Odeck, J (2004) Cost overruns in road construction—what are their sizes and determinants? *Transport Policy*, **11**(1), 43-53.

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.

Oza, N C (2006) Ensemble data mining methods. *Encyclopedia of Data Warehousing and Mining, Idea Group Inc*, 448-53.

Pattengale, N D, Alipour, M, Bininda-Emonds, O R, Moret, B M and Stamatakis, A (2009) How many bootstrap replicates are necessary? *In*, *Research in Computational Molecular Biology*. Springer, 184-200.

Railnews (2012) Edinburgh tram costs soar again. In: *Railnews (14th June 2012), http://goo.gl/M5uZ7*.

Sarle, W S (1994) Neural Networks and Statistical Models. In: *Proceedings of the Nineteenth Annual SAS Users Group International Conference*.

Skapura, D M (1996) *Building neural networks*. USA: ACM Press.

StatSoft Inc (2008) A Short Course in Data Mining. In: StatSoft, Inc.

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

Staw, B M (1981) The escalation of commitment to a course of action. *Academy of management Review*, **6**(4), 577-87.

Wachs, M (1989) When planners lie with numbers. *Journal of the American Planning Association*, **55**(4), 476–9.

Wachs, M (1990) Ethics and advocacy in forecasting for public policy. *Business and Professional Ethics Journal*, **9**(1-2), 141–57.

Weinstein, N D (1980) Unrealistic optimism about future life events. *Journal of personality and social psychology*, **39**(5), 806.

Zhang, X, Mao, X and AbouRizk, S M (2009) Developing a knowledge management system for improved value engineering practices in the construction industry. *Automation in Construction*, **18**(6), 777-89.