MY COST RUNNETH OVER: DATA MINING TO REDUCE CONSTRUCTION COST OVERRUNS

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

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

Most construction projects overrun their budgets. Among the myriad of explanations giving for construction cost overruns is the lack of required information upon which to base accurate estimation. Much of the financial decisions made at the time of decision to build is thus made in an environment of uncertainty and oftentimes, guess work. In this paper, data mining is presented as a key business tool to transform existing data into key decision support systems to increase estimate reliability and accuracy within the construction industry. Using 1600 water infrastructure projects completed between 2004 and 2012 within the UK, cost predictive models were developed using a combination of data mining techniques such as factor analysis, optimal binning and scree tests. These were combined with the learning and generalising capabilities of artificial neural network to develop the final cost models. The best model achieved an average absolute percentage error of 3.67% with 87% of the validation predictions falling within an error range of ±5%. The models are now being deployed for use within the operations of the industry partner to provide real feedback for model improvement.

Keywords: artificial neural networks, cost estimation, cost overrun, data mining, decision support system.

INTRODUCTION

The business landscape is continually experiencing a growing recognition of information as a key competitive tool. Companies that are able to successfully collect, analyse and understand the information available to them are among the winners in this new information age (Huang *et al.* 2006). Available computing hardware and database technology allows for easy, efficient and reliable data storage and retrieval. Additionally, widespread use of networked computers and sophisticated database systems enables companies to pool their data together from across different geographical locations using data servers. However, 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 in different areas if they are able to make their data work for them using data mining.

As pointed out by Fayyad *et al.* (1996a), the real value of storing data lies in the ability to exploit useful trends and patterns in the data to meet business, operational, or scientific goals as well as for decision support and policy making. Present advances in

[1] D.Ahiaga-Dagbui@ed.ac.uk

Ahiaga-Dagbui DD and Smith SD (2013) My cost runneth over: data mining to reduce construction cost overruns *In:* Smith, S.D and Ahiaga-Dagbui, D.D (Eds) *Procs 29th Annual ARCOM Conference,* 2-4 September 2013, Reading, UK, Association of Researchers in Construction Management, 559-568.

the fields of data warehousing, artificial intelligence, statistics, data 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 (Bose and Mahapatra 2001). Data mining, knowledge discovery in databases, has been extensively used in fields such as business (Apte *et al.* 2002), finance (Kovalerchuk and Vityaev 2000) and medicine (Koh and Tan 2011). However, informal discussions with a number of construction companies suggest that very few take advantage of their data, transforming it into cutting edge business decision support tools. Against this backdrop, the authors have provided an overview of the field of data mining with some specific applications in construction management. The data mining methodology is then applied to the problem of cost estimation in the construction industry using Artificial Neural Networks (ANN). Final cost prediction models were developed using the vast project database of a major water utility provider in the UK. The aim was to convert the experience and knowledge imbedded in past projects into intelligence and decision support systems that could potentially improve the accuracy of construction cost estimation, thereby reducing the problem of cost overruns.

DATA MINING

Data mining 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. Data mining has been applied to detect money laundry and fraudulent transactions by Senator *et al.* (1995), investigate the effectiveness of sales campaigns by Ngai *et al.* (2009), intrusion detection in computer network administration by Julisch (2002) and for loan repayment assessment (see Lee *et al.* 2006).

Although it is yet to find extensive application in practice within the construction industry, construction management researchers have started investigating data mining's applicability to different problems. It has been applied to improving construction knowledge management (Yu and Lin 2006), estimating the productivity of construction equipment (Yang *et al.* 2003), study of occupational injuries (Cheng, Leu, *et al.* 2012), alternative dispute resolution (Fan and Li in press) and prediction of the compressive strength high performance concrete (Cheng, Chou, *et al.* 2012).

Data Mining Process

Data mining normally follows a generic process illustrated in Figure 1. It starts with the selection of relevant data from a data warehouse that contains information on organisation and business transactions of the firm (Ngai *et al.* 2009). 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 and avoid the curse of 'garbage-in-garbage-out'. Pre-processing typically involves steps such as removing of duplicate entries, sub-sampling, clustering, transformation, de-noising, normalisation or feature extraction (StatSoft Inc. 2011b). The authors however note the issue of unavailable of relevant data as a potential barriers to effective data mining in the construction industry as most firms do not have a culture of storing detailed information about the projects they undertake.

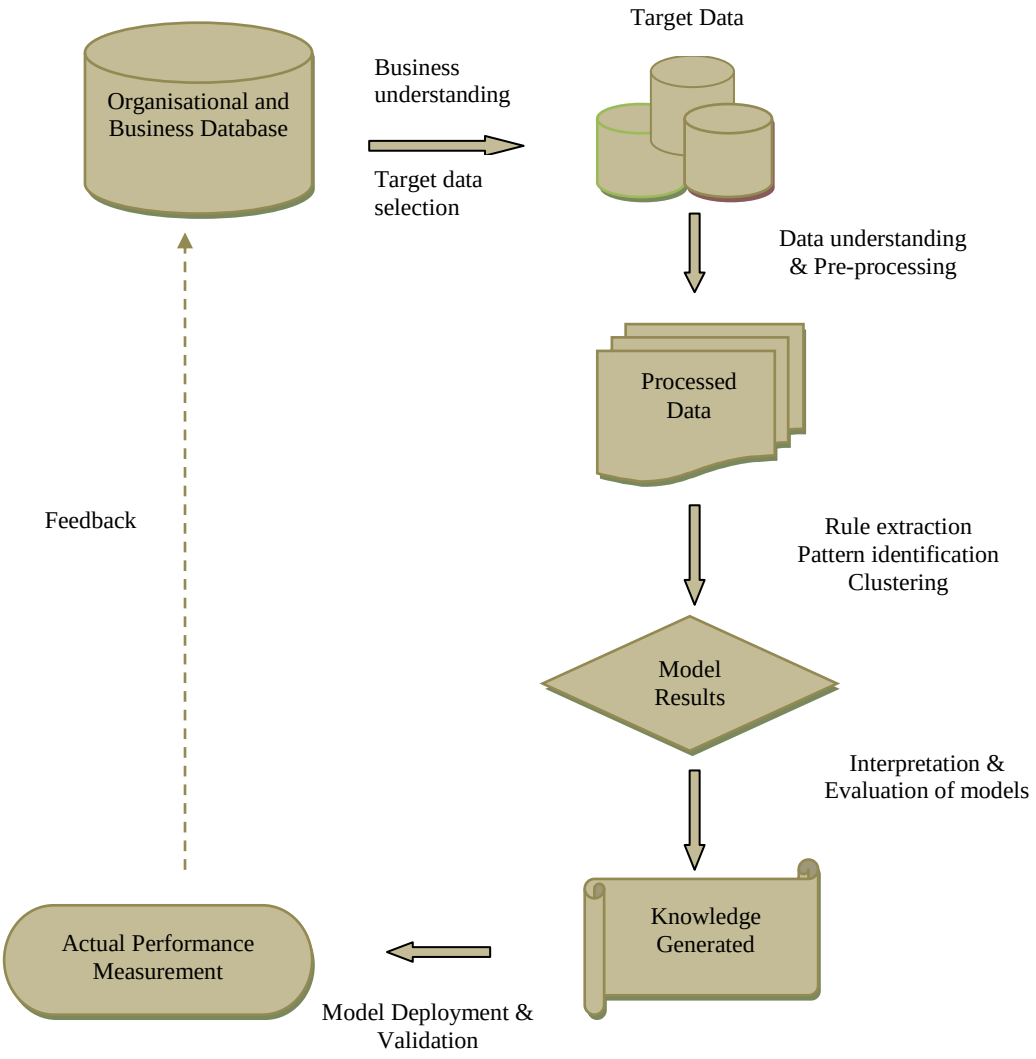

Figure 1: The generic data mining process

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. The type of modelling approach adopted would depend on the a number of factors, chief of which would normally be the type and quantity of data available, the aim of the modelling exercise and the predictive performance required (StatSoft Inc 2008). This is often an elaborate process, sometimes involving the use of competitive evaluation of different models and approaches and deciding on the best model by some sort of bagging system (voting or averaging) (StatSoft Inc. 2011a). Some of the available modelling techniques include case-based reasoning, principal component analysis, regression, decision trees, machine learning, genetic algorithm, fuzzy logic, as well as artificial neural networks, which has been used for the experimental part of this paper. The results from the data mining stage are then evaluated and presented into some meaningful form to aid business decision making. This step might involve graphical representation or visualisation of the model for easy communication. The knowledge generated is then validated by deploying the model in a real life situation to test the model's efficacy (Koh and Tan 2011).

It is important to note however that data mining in itself does not guarantee success when the models are deployed. For instance, if one seeks long enough in any database, it is possible to find patterns and seeming interrelations between variables which are

actually not valid (Fayyad *et al.* 1996b), resulting in model failure when deployed in real life. Also, no amount of data will allow for accurate prediction based on attributes that do not capture the required information. Success from any data mining venture is predicated on the availability of quality and quantity of data (StatSoft Inc. 2011a). The data must essentially contain data attributes that are relevant to the problem under investigation.

COST OVERRUNS

Cost performance on a construction project remains one of the main measures of the success of a construction project (Atkinson 1999; Chan and Chan 2004). However, estimating the final cost of construction projects can be extremely difficult due to the complex web of cost influencing factors that need to be considered (Ahiaga-Dagbui and Smith 2012) - type of project, likely design and scope changes, ground conditions, duration, type of client, tendering method- the list is endless. Trying to work out the cost influence of most of these variables at the inception stage of a project where cost targets are normally set can an exhaustive task, if not at all futile. Ignoring most of them altogether creates a perfect recipe for future cost overruns. Also, a high level of uncertainty surrounds most of these factors at the initial stages of the project (Jennings 2012).

Flyvbjerg *et al.* (2004) report that 9 out 10 infrastructure projects overrun their budgets and that infrastructure project have an 86% likelihood of exceeding their budgets. The on-going Edinburgh Trams project has already far exceeded its initial budget leading to significant scope reduction to curtail the ever-growing cost (Miller 2011; Railnews 2012). The recent 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 completed at £8.9 billion in 2010 (Gidson 2012; NAO 2012). These statistics have often led to extensive claims, disputes and lawsuits in some cases within the industry (Love *et al.* 2010).

Causes of overruns have been attributed to several sources including improperly managed risk and uncertainty (Okmen and Öztas 2010), scope creep (Love *et al.* 2011), optimism bias (Jennings 2012) to suspicions of foul-play and corruption (Wachs 1990; Flyvbjerg 2009). Another potential root cause of overruns is the lack of adequate information on which to base realistic and accurate estimates. Nicholas (2004) points out that estimators thus have to rely largely on their own experience and historical cost information when preparing initial estimates. Typically, an estimate can only be as good as the information it is based on so that, *ceteris paribus*, the level of accuracy of the estimates produced also increases as more information becomes available. Data mining is thus deemed as a possible way of capturing valuable information within historical data to support the estimation process at the initial stages of project definition.

DATA

The data used for the models in this paper were supplied by an industry partner with its primary operation in the delivery of water infrastructure and utility in the UK. The authors were granted access to the vast database of almost 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 £1000 - £30 million and durations from a short 3 months to 5 years.

The initial analysis involved drilling down into the database to find what might be useful in modelling final cost. First, cluster analysis and purposive sampling was used to create groups of project cases that were similar, without significant missing data or extreme values and representative of the entire dataset. One of the clusters containing about 1600 projects completed between 2004 and 2012, with cost range of between £4000 -15 million, comprising newly built, upgrade, repair or refurbishment projects was used for the models reported in this paper. 15 project cases were selected using stratified random sampling to be used for independent testing of the final models. The remaining data was then split in an 80:20% ratio for training and testing of the models, respectively.

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. 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. Further variable screening using scree test, factor analysis and optimal binning in Statistica 10 software was used to reduce the initial set of predictors to six[2]

Cost values were normalised to a 2012 baseline with base year 2000 using the infrastructure resources cost indices by the Building Cost Information Services (BCIS 2012). 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.

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 numerical in nature. 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 (Anderson 1995). ANN, an abstraction of the human brain with abilities to learn from experience and generalise based on acquired knowledge, is also able to cope with

[2] Initial list of predictors: 1. Delivery Partners - X, Y, Z 2. Purpose - wastewater, water, general 3. Scope of project - newbuild, upgrade, refurbishment, replacement 4. Target cost 5. Duration. 6. Operating region – North, South, East, West;

multicollinearity (Moselhi *et al.* 1991), a characteristic of construction data (Boussabaine and Elhag 1999). Neural networks has already been used to develop prototype models at an earlier stage of the this research (*see* Ahiaga-Dagbui and Smith 2012) and has also been applied to forecasting tender price (Emsley *et al.* 2002) and for identification and quantification of risk by McKim (1993). See Moselhi *et al.* (1991) for a review of neural network application in construction management research.

The final model was developed after an iterative process of fine-tuning the network parameters and/or inputs until acceptable error levels were achieved or when the model showed no further improvement. First, the automatic network search function of Statistica 10 software was used to optimise the search for the best network parameters, after which customized networks were developed using the optimal parameters identified. 5 activations functions[3] were used at this stage in both hidden and output layers, training 2000 multi-layer perceptron networks and retaining the 5 best for further analysis. The overall network performance was 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:

$$SOS = \sum (T_i - O_i)^2 \quad \ldots\ldots \text{Equation 2}$$

Where Oi is the predicted final cost of the ith data case (Output)

T_i is the actual final cost of the ith data case (Target).

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. The higher the *p*-value, the less reliable the observed correlations.

The retained networks are then validated using the 15 separate projects that were selected using stratified sampling at the beginning of the modelling exercise. See *Figure 2* for the overall performance of 7 of the retained networks. This plot allowed for a quick comparison of the average error achieved by the selected models. A sensitivity analysis was performed on each retained network to assess predictor's contribution to network performance. To do this, the model's predictive performance is measured while deleting one input factor at a time, starting from the least important, until the model showed no further improvement or begun to decay.

Table 1 shows the predictions and absolute percentage errors (APE) achieved by model 33, which as the best overall model. The average APE achieved by model 33 was 3.67% across the 15 validation cases. Its APEs ranged between 0.04% and 15.85%. It was observed that the worst performances of the model were achieved on projects with the smallest values in the validation set (cases 13 & 15). This might potentially be because a majority of the projects used for the model training had values in excess of £5 million. However, the real monetary errors on these predictions were deemed satisfactory as they were relatively small (about £3500 & £2500 for models 13 & 15 respectively). 87% of the validation predictions of the best model were within ±5% of the actual cost of the project.

[3] identity, logistic, tanh, exponential and sine activation functions

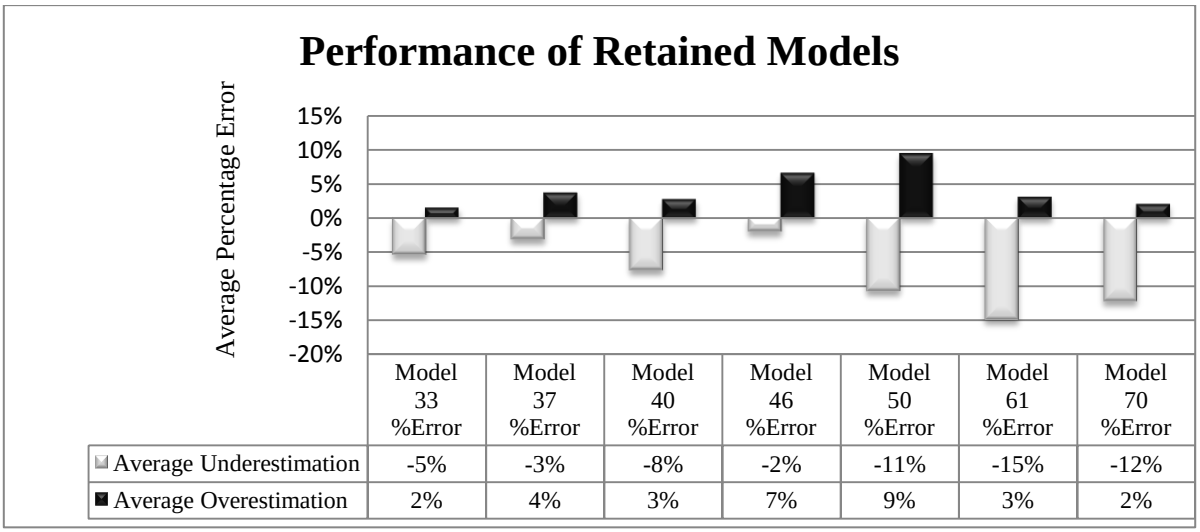

Figure 2: Performance of selected models

Table 1: Validation results of the best model (Model 33)

Validation Case	*Actual Final Cost*	*Final Cost predicted*	*Model Error*	*Model Absolute % Error*
1	£ 4,912,649	£ 5,120,943	-£ 208,294	4.24%
2	£ 1,617,225	£ 1,617,805	-£ 580	0.04%
3	£ 11,277,470	£ 10,743,624	£ 533,846	4.73%
4	£ 2,110,260	£ 2,136,125	-£ 25,865	1.23%
5	£ 5,398,965	£ 5,425,142	-£ 26,177	0.48%
6	£ 180,532	£ 181,214	-£ 681	0.38%
7	£ 2,572,564	£ 2,530,178	£ 42,386	1.65%
8	£ 1,440,593	£ 1,372,864	£ 67,729	4.70%
9	£ 3,842,258	£ 3,793,851	£ 48,407	1.26%
10	£ 4,194,219	£ 4,131,285	£ 62,934	1.50%
11	£ 375,170	£ 387,731	-£ 12,561	3.35%
12	£ 50,637	£ 51,502	-£ 865	1.71%
13	£ 24,479	£ 22,017	£ 2,462	10.06%
14	£ 858,112	£ 824,334	£ 33,779	3.94%
15	£ 21,798	£ 18,344	£ 3,454	15.85%
			Average Absolute % Error	**3.67%**

CONCLUSION

The authors make a case for using data mining in modern construction management as a key business tool to improve construction performance. This could essentially help construction firms to transform their data into cutting edge decision support systems for business improvement and gain competitive advantage. An overview of data mining and its methodology, as well as applications have been detailed in the paper. The method was then applied to the problem of final cost estimation of construction project using artificial neural networks. Cost estimation was chosen for this study as one of the main reasons cited for cost overruns is the lack of information at the initial stages of the project for accurate estimation. Data mining thus attempts to exploit

already existing information, in combination with what is known about the new project to make its forecasts of final cost. The best model in this paper achieved an average absolute percentage error of 3.67% with 87% of the validation predictions falling within an error range of ±5%. The authors 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 authors however identify a poor culture of data warehousing in the construction industry as one of the major challenges to effective data mining. For most construction companies, relevant data for modelling construction processes is sparse or even unavailable. Data mining depends heavily on the availability of business, operational and project data, stored in a meaningful and retrieval manner. Also, it is important to point out that the potential benefits of data mining are not overstated or lauded by researchers or practitioners as panaceas in themselves. Its limitations and potential pitfalls must always be clearly communicated to the end user.

REFERENCES

Ahiaga-Dagbui, D D and Smith, S D (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, 307-16.

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.

Atkinson, R (1999) Project management: cost, time and quality, two best guesses and a phenomenon, its time to accept other success criteria. *International Journal of Project Management*, **17**(6), 337-42.

BCIS (2012) BIS Construction Price and Cost Indices. In, http://www.bcis.co.uk: Building Cost Information Services, UK.

Bose, I and Mahapatra, R K (2001) Business data mining—a machine learning perspective. *Information & Management*, **39**(3), 211-25.

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.

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

Cheng, C-W, Leu, S-S, Cheng, Y-M, Wu, T-C and Lin, C-C (2012) Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. *Accident Analysis & Prevention*, **48**(0), 214-22.

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.

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.

Fan, H and Li, H (in press) Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques. *Automation in Construction*(0).

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

Fayyad, U, Piatetsky-Shapiro, G and Smyth, P (1996b) From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, **17**(3), 37-54.

Flyvbjerg, B (2009) Survival of the unfittest: why the worst infrastructure gets built—and what we can do about it. *Oxford Review of Economic Policy*, **25**(3), 344-67.

Flyvbjerg, B, Holm, M K S and Buhl, S (2004) What Causes Cost Overrun in Transport Infrastructure Projects? *Transport Reviews*, **24**(1), 3-18.

Gidson, O (2012) London 2012 Olympics will cost a total of £8.921bn. *The Guardian*. http://goo.gl/sxatK. *23 October 2012*

Huang, M-J, Tsou, Y-L and Lee, S-C (2006) Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. *Knowledge-Based Systems*, **19**(6), 396-403.

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.

Julisch, K (2002) Data mining for intrusion detection. *Applications of data mining in computer security*, 33-58.

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

Kovalerchuk, B and Vityaev, E (2000) Data mining in finance. In: USA: Kluwer Academic Publisher, Hingham MA.

Lee, T S, Chiu, C C, Chou, Y C and Lu, C J (2006) Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis*, **50**(4), 1113-30.

Love, P, Davis, P, Ellis, J and Cheung, S O (2010) Dispute causation: identification of pathogenic influences in construction. *Engineering, Construction and Architectural Management*, **17**(4), 404-23.

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.

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*. 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.

NAO (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.

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.

Railnews (2012) Edinburgh tram costs soar again. *Railnews*. http://goo.gl/M5uZ7

Senator, T E, Goldberg, H G, Wooton, J, Cottini, M A, Khan, A F U, Klinger, C D, Llamas, W M, Marrone, M P and Wong, R W H (1995) Financial Crimes Enforcement Network AI System (FAIS) Identifying Potential Money Laundering from Reports of Large Cash Transactions. *AI Magazine*, **16**(4), 21.

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

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

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

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

Yang, J, Edwards, D J and Love, P E D (2003) A computational intelligent fuzzy model approach for excavator cycle time simulation. *Automation in Construction*, **12**(6), 725-35.

Yu, W-d and Lin, H-w (2006) A VaFALCON neuro-fuzzy system for mining of incomplete construction databases. *Automation in Construction*, **15**(1), 20-32.