THE UNIVERSITY *of* EDINBURGH

RETHINKING CONSTRUCTION COST OVERRUNS:

An artificial neural network approach to construction cost estimation

by

Dominic D Ahiaga-Dagbui, BSc (Hons), MSc, AFHEA

A thesis submitted in fulfilment of the requirements for the award of

PhD in Construction Project Management

at

The University of Edinburgh

2014

ABSTRACT

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 usually ranks highest. 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. Construction projects therefore have routinely overrun their estimates.

Cost overrun has been attributed to a number of sources including technical error in design, managerial incompetence, risk and uncertainty, suspicions of foul play and even corruption. Furthermore, even though it is accepted that factors such as tendering method, location of project, procurement method or size of project have an effect on likely final cost of a project, it is difficult to establish their measured financial impact. Estimators thus have to rely largely on experience and intuition when preparing initial estimates, often neglecting most of these factors in the final cost build-up. The decision-to-build for most projects is therefore largely based on unrealistic estimates that would inevitably be exceeded.

The main aim of this research is to re-examine the sources of cost overrun on construction projects and to develop final cost estimation models that could help in reaching more reliable final cost estimates at the tendering stage of the project.

The research identified two predominant schools of thought on the sources of overruns – referred to here as the PsychoStrategists and Evolution Theorists. Another finding was that there is no unanimity on the reference point from which cost performance could be assessed, leading to a large disparity in the size of overruns reported. Another misunderstanding relates to the term "cost overrun" itself.

The experimental part of the research, conducted in collaboration with two industry partners, used a combination of non-parametric bootstrapping and ensemble modelling with artificial neural networks to develop final project cost models based on about 1,600 water infrastructure projects. 92% of the validation predictions were within ±10% of the actual final cost of the project. The models will be particularly useful at the pre-contract stage as they will provide a benchmark for evaluating submitted tenders and also allow the quick generation of various alternative solutions for a construction project using what-if scenarios.

The original contribution of the study is a fresh thinking of construction "cost overruns", now proposed to be more appropriately known as "cost growth" based on a synthesises of the two schools of thought into a conceptual model. The second contribution is the development of novel models of construction cost estimation utilising artificial neural networks coupled with bootstrapping and ensemble modelling.

DECLARATION

I, Dominic D Ahiaga-Dagbui, declare that the work contained in this thesis has not been submitted for any other degree or professional qualification. It represents my own work, except where duly referenced, carried out under the supervision of Dr Simon D. Smith between August 2011 and July 2014.

24th July 2014

DEDICATION

‘Ebenezer,

Thus far have you brought me’

ACKNOWLEDGEMENT

I am really grateful for the support, supervision and friendship of
Dr Simon D. Smith - You did more than just your job

I am also indepted to the industry collaborators in this research,
without whose involvement the thesis would not have been completed
in its current form

My family – the best cheerleaders ever!

Grateful for the prayers and support of friends, especially in the iMC at
Central - you guys rock!

Kel, your words of encouragement and prayers kept me going especially
when the going got tough.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

"Somewhere, something incredible is waiting to be known."

Carl Sagan

(Astronomer, Writer and Scientist, 1934-1996)

1.1. BACKGROUND

Nine out of ten construction projects overrun their budget. Infrastructure projects have an 86% probability of outrunning their set cost targets. The size of these overruns can on average be as high as 45% for rail projects, 34% for bridges and tunnels and 20% for road projects. These are some of the staggering conclusions of the seminal works by Flyvbjerg *et al.* (2002, 2004).

Love *et al* (2012) report overruns of up to 70% more than the initial estimate while Odeck (2004) reports a cost overrun up to 183% of original cost. The total cost of 20 projects assessed by the Auditor General of Western Australia (2012) was A$6.2 billion, an astonishing A$3.3 billion (114%) more than the total original approved budget estimates. At least four of these projects were expected to experience overruns well beyond 200% of the original cost. These statistics might suggest that overruns are the normative, rather than the exception in the construction industry.

Overruns occur irrespective of the type of project - roads, bridges, rails, houses, schools and tunnels all suffer the same fate (Hinze *et al.* 1992, Love *et al.* 2014). Is it a small or mega project? Size really does not matter either (Bordat *et al.* 2004, Love *et al.* 2014). Flyvberg *et al*.'s (2002) database of 285 projects studied had cost values ranging from a small $1.5 million to a mega $8.5 billion.

Interestingly, this endemic phenomenon is not limited to developing countries that might arguably be stereotypically characterised by poor management, ineffective project delivery or corruption. In happens in the UK, Europe, Australia, the Americas, Africa, Asia. Everywhere. It is global.

Sounds like a sweeping generalisation? Perhaps. Well, maybe not. Boston's Central Artery in the US, dubbed the "Big Dig", was to cost US$2.6 billion but was completed at US$14.8 billion (Gelinas 2007). The New Children Hospital in Australia was approved at A$207 million. It

was delivered 365% over budget at A$962 million (Auditor General of Western Australia 2012). Edinburgh's recently completed tram project costs a reported £776 million instead of the initially estimated £375 million (City of Edinburgh Council 2014). Depending on which figures used, every Olympic Game since 1960 has experienced cost overruns - 179% on average (Flyvbjerg and Stewart 2012).

Unfortunately, it would seem that construction projects tend to make the news headlines, not for being remarkable engineering accomplishments that will support and stimulate economic growth and social integration of communities, but rather for being poorly managed and grossly over budget. The industry may have well earned itself the unenviable repute of delivering projects late and over budget, again and again, leaving clients dissatisfied and the tax-payer often out of pocket (Egan 1998, Audit Scotland 2004, Auditor General of Western Australia 2012).

Oddly though, Flyvbjerg *et al.* (2002) observed that the size of overruns have not improved over the 70 years that they studied. The trend continues to the present day with the Edinburgh Trams and the World Cup stadiums in Brazil still making the news headlines currently (City of Edinburgh Council 2014, Stadium Database 2014). Flyvbjerg *et al* (ibid, pp 290) thus controversially concluded that "no learning that would improve cost estimate accuracy seems to take place."

Why are cost overruns so prevalent in the industry? Why has there not been much improvement in the reliability of initial cost estimates over the years? Surely the industry has become a lot better at managing projects. Procurement systems have greatly evolved from traditional adversarial design-bid-build to different forms of collaborative and relationship contracts. There are more measures now for accountability and cost control for project procurement. Information Technology for construction has also improved significantly with the advent of Computer Aided Designs (CAD) and Building Information Modelling (BIM). There are now online collaborative platforms for effective

communication, design, visualisation, simulation, control and coordination of the entire construction process. There is growing take-up of digital 3D design and even 4D models that integrate the spatial and temporal aspects of a project to understand, predict, evaluate and manage even the most complex projects. Most of these IT systems support project cost estimation as well as allow for the use of estimation software and advanced costing methods like feature-based estimation, Monte-Carlo simulations, genetic algorithms or fuzzy logic.

It is against this backdrop that this research seeks to revisit the problem of cost overruns, to prompt a rethink of its sources, scale and provide some avenues for dealing with the problem.

1.2. PROBLEM STATEMENT

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 and very often as a benchmark for future performance measure. 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. Effective cost estimation is therefore so vital; it can seal a project's financial fate.

However, construction and infrastructure projects have, historically, cost more at the completion of the construction phase than was anticipated at the conception phase. This causes concern for clients as they are unable to forecast their total financial commitment for the project and often have to secure extra funds as well as find themselves

suffering from significant reputational detriments. It is also a concern for virtually every other stakeholder on a construction project particularly financers, contractors, designers and project operators. Understandably, projects overrunning their budgets cause disquiet to the tax payer when public money is used to finance the project. Perhaps the only beneficiaries of such events are the media, particularly the print media[1].

It is not difficult to appreciate the need for a solution to this problem, for greater estimate reliability at all stages of a project and for greater assurance that initial cost expectations are met. But, as the subsequent discussions in this thesis will show, the problem extends to the definition of what "budget overrun" actually means. There is also a misunderstanding of the actual size of overruns as well as the sources of these overruns on projects.

[1] For the Edinburgh public, the cost overruns of their Tram project and the Scottish Parliament have been a regular feature of both of the city's newspapers, who have delighted in revealing the latest catastrophes and misfortunes, giving a concerned view, while of course reveling in the greater distribution of print. Some of the news headlines included Edinburgh Tram "out of control again", "Edinburgh's tram fiasco...", "No end to Holyrood bills even when it's finished" and "Holyrood saga shatters Scots Illusions".

1.3. AIMS AND OBJECTIVES

The aims of this study, alongside the objectives to achieve these aims are stated as follows:

1. *To provide a better conceptual understanding of "cost overruns".*
   Specific objectives:
   a) ascertain through a critical review of the literature, the factors that contribute to the difference between the initially estimated cost and the resulting final costs at project completion;
   b) explore the different theoretical schools of thought on the cost overruns;
   c) synthesise the different schools of thought into a holistic conceptual model to help properly understand overrun.

2. *To develop final cost estimation models to forecast likely total cost of projects based on historical cost and project details of completed project.*
   Specific objectives:
   a) identify and collect a reliable dataset for the cost modelling process;
   b) establish a neural network modelling protocol for developing the cost models;
   c) validate the models using new project cases.

1.4. RESEARCH QUESTIONS

A preliminary review of the literature on cost overruns and estimation led to the following research questions that will guide the research in order to achieve the stated aims:

1. *Is the current understanding of construction cost overruns adequate?*

   An initial literature review suggests that there is no unanimity on the reference point for measuring assessing cost performance on construction projects thus leading to a large disparity in the size or range of overruns reported.

2. *What are the predominant schools of thought on the sources of construction cost overruns?*

   There also seems to be conflating theoretical explanations on the sources of cost overruns in the literature. While one school seems to explore the issue from an organisational and strategic point of view, the other tackles overruns from a technical and engineering perspective.

3. *Is there a conceptual difference between cost underestimation and cost overruns?*

   This question is closely linked to the previous. There does not seem to be distinction between these two terms when cost escalation is being discussed, especially by the media. If not correctly understood, they might lead to misleading conclusions and misplaced accusations when projects seem to outrun their budgets.

4. *Is neural networks an appropriate method of estimating the cost of construction projects?*

   While neural networks has been widely used for problems like foreign exchange prediction, medical diagnosis, flight and robot control and loan applicant assessment, it is yet to find widespread use in construction management research, particularly for cost estimation. Why is this the case? Initial answers include the large data requirements of neural networks.

Requisite data was not a problem for this research as a large database was secured. The effectiveness of using neural networks for final cost estimation will thus be explored as part of this research.

1.5. OVERVIEW OF RESEARCH APPROACH

The critical considerations when choosing a research approach often revolve around the type of problem under study and how to maximise the chances of adequately answering the research question(s) of the study. The research reported in this thesis largely adopts a quantitative approach with some elements of qualitative approach especially in the early stages of the research. A thorough exploration and critique of existing literature on construction cost overruns laid the basis on which to conduct a *quasi* experimental model development to estimate likely final cost.

After defining the aims and objectives, the research begun with a thorough review and critique of existing literature and theories on the cost overrun phenomenon. An experimental approach was then designed to provide the framework for data collection, analyses and subsequent model validation. The results of the model development were then discussed before reaching conclusions and recommendations for achieving more reliable cost estimates in the early stages of construction projects. A greater discussion on the research approach, with appropriate consideration of methodology is provided in Chapter Three.

1.6. RESEARCH CONTRIBUTION

The contributions of this research to the field of construction project management are closely aligned with the research aims described previously. But the reader of this thesis deserves to understand the intentions, the argument and the original contribution of the work, and while the research is yet to be demonstrated and disseminated in the remainder of this thesis, there is no harm in stating the contributions in advance. These are summarised as:

- **New understanding of construction cost overruns**
  Construction cost is not a new research area in any sense and has been investigated by many outstanding researchers in the past. But this current research has demonstrated conflation and confusion of existing understanding of cost overruns from two predominant schools of thought. The contribution here is a fresh rethinking of construction 'cost overruns', now proposed to be more appropriately known as 'cost growth'.

- **Novel models of construction cost estimation utilising Artificial Neural Networks coupled with Bootstrapping and Ensemble Modelling**
  As far as can be ascertained this combination of modelling approaches has rarely been used in any modelling problem and never in the field of construction cost. It is possibly naïve to expect such subtle modelling approaches to be embraced and utilised immediately by practitioners but the academic field is healthy and active – the work here undoubtedly makes an original contribution to it.

1.7. THESIS STRUCTURE

The thesis is organised into five chapters as illustrated in Figure 1.

1.7.1. Chapter One - Introduction

This chapter sets out the background and context of the research, laying the basis for the problem statement along with the aims and objectives of the research. The research questions have also been stated in this chapter.

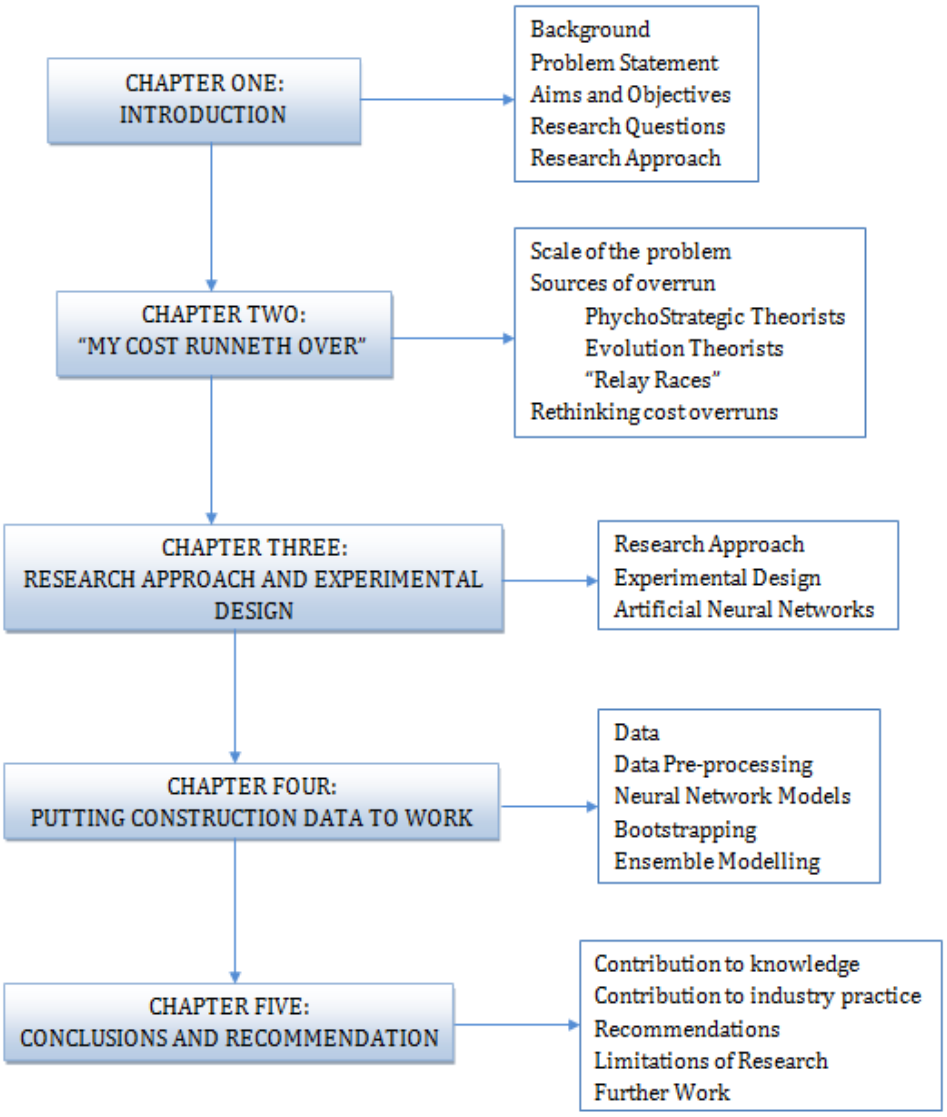


Figure 1: Thesis Structure

1.7.2. Chapter Two - "My Cost Runneth Over"

This chapter forms the critical spine of the research where the thesis statement is clearly identified. It briefly presents an overview of the scale of cost overruns in construction before critically evaluating the theoretical explanations of the causes of the phenomenon. Two notable schools of thought, referred to in this thesis as the PhychoStrategists and Evolution Theorists will be presented.

A conceptual model that attempts to balance the tension between these two schools is later presented in a section captioned Rethinking Cost Overruns, forging a platform that will allow for focussed development of more detailed cost forecasting models covered in the later parts of this thesis.

1.7.3. Chapter Three - Research Approach and Experimental Design

This section presents the considerations for adopting the research approach used as well as the experimental framework that will guide the modelling aspect of the research. The background, applications, strengths and weakness of artificial neural network are also evaluated in this chapter.

1.7.4. Chapter Four - Putting Construction Data to Work

This is the experimental chapter of the thesis where the data collected from two construction organisations will be 'put to work' to develop cost models for forecasting final cost. A small dataset of 98 completed projects will be used to develop trial models and experiment with different modelling strategies. Neural networks, data bootstrapping and ensemble modelling will then be used to explore a larger dataset of 1,600 projects in search of consistent patterns, correlations and systematic relationships between variables. The aim here is straightforward: to make data 'work' for construction organisations by extracting useful information embedded in them to predict cost.

The data analysis and the results achieved have intentionally not been split into different chapters so as to improve the readability of the thesis.

1.7.5. Chapter Five - Conclusions and Recommendation

In the closing chapter, conclusions will be made based on the findings from the literature review and data modelling chapters. The aims and objectives set out in the initial chapter of the thesis will be revisited and a judgement made on whether these have been achieved, and to what extent.

Finally, theoretical and practical contributions of the research will be usefully summarised while detailing implications of the findings for the procurement and management of projects. Some considerations of future research will then be provided along with limitations of the current research.

1.7.6. References

The list of all works cited in the thesis.

1.7.7. Appendix

Supplementary materials including publications and data collection form.

CHAPTER TWO

"MY COST RUNNETH OVER"

When we mean to build,

We first survey the plot, then draw the model;

And when we see the figure of the house,

Then must we rate the cost of the erection;

Which if we find outweighs ability,

What do we then but draw anew the model

In fewer offices, or at last desist

To build at all?

~ Shakespeare, Henry IV, Part 2

2.0 INTRODUCTION

Effective cost planning relates the design of buildings to their cost, so that while taking full account of quality, risks, likely scope changes, utility and appearance, the cost of a project is planned to be within the economic limit of expenditure. 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. Despite the importance of cost estimation, it is undeniably not neither simple nor straightforward. To achieve satisfactory accuracy and reliability of estimates, the estimator has to be able to reckon all future chain of events from project inception to its eventual delivery.

Apart from costing materials and labour, the estimator also has to take into consideration factors such as the type of project, likely design and scope changes, ground conditions, duration, type of client, tendering method. 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, would be an exhaustive task, if not at all futile. Ignoring most of them altogether creates a perfect recipe for future cost overrun.

This chapter will present a thorough discussion of the problem of construction cost overruns, more appropriately termed cost growth in some sections. First, a brief overview of the scale of cost growth is presented before discussing the theoretical explanations of the causes of the phenomenon, mainly from the perspective of two notable schools of thought, referred to in this thesis as the PhychoStrategists and Evolution Theorists. A conceptual model that attempts to balance the tension between these two schools is later presented in a section captioned *Rethinking Cost Overruns*, creating a platform that will allow for focussed development of more detailed cost forecasting models covered in the later parts of this thesis.

2.1. CONSTRUCTION COST OVERRUN: THE SCALE OF THE PROBLEM

The statistics on construction cost overruns have been well documented in the literature, official government publications and popular media. This section only presents an overview of the scale of the problem as a prelude to the discussion on how and why cost overruns actually occur.

The Auditor General of Western Australia assessed the management and performance of 20 capital intensive projects including sports venues, schools and hospitals, undertaken in Australia. The expected cost of all the projects at the time was A$6.157 billion, a staggering A$3.275 billion (114%) more than the total original approved budget estimates. 15 of the 20 projects were expected to exceed their original approved budgets, of which four were expected to exceed their budgets by more than 200% (Auditor General of Western Australia 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 Edinburgh Trams project in Scotland exceeded its initial budget leading to significant scope reduction to curtail the ever-growing cost (Miller 2011, Railnews 2012). The project, was initially expected to cost about £375 million, but was completed three years late at a reported £776 million (City of Edinburgh Council 2014).

The City of Boston's Central Artery project (popularly referred to as the Big Dig) was to cost US$2.6 billion but was completed at US$14.8 billion and 7 years late in 2006 (Gelinas 2007).

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).

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.

Table 1 shows some other examples of projects that have gone over budget. These statistics have often led to extensive claims, disputes and lawsuits in some cases within the industry.

Table 1: Some Examples of Cost Overrun in Construction Projects

Project	*Estimated Cost (in millions)*	*Final Cost (in millions)*	*% Overrun*
Sydney Opera House	A$7	A$102	1357
Nat West Tower	£15	£115	667
Thames Barrier Project	£23	£461	1904
Scottish Parliament	£195*	£414	112
British Library	£142	£511	260

**September 2000 estimate. Initially stated cost was about £40 million Source: Audit Scotland (2004)*

2.2. SOURCES OF COST OVERRUN

Many contracts are signed every day for some form of building work. It may be a completely new building, refurbishment or maintenance project. Some projects are simple, others complex; still, some take a couple of weeks and others, several years. Yet, they all have one thing in common - they can all go wrong. Causes of cost growth have been attributed to several sources including unidentified or improperly managed risk and uncertainty (Okmen and Öztas 2010), scope creep and rework (Love *et al.* 2005), optimism bias (Lovallo and Kahneman

2003, Flyvbjerg 2008) to suspicions of foul-play and corruption (Wachs 1990, Flyvbjerg 2009).

Without duplicating the extensive literature available on the subject or trying to provide an exhaustive list of the causes of cost overrun, this section of the thesis is only concerned with a synthesis of the mainstream arguments found in the literature. The review will draw particularly on the works of some of the contemporary authorities on the subject, such as Peter Love and Bent Flyvbjerg, to provide a holistic understanding of the cost overrun phenomenon.

2.2.1. Risk and Uncertainty

The terms risk and uncertainty are often used interchangeably although they do not necessarily mean the same thing. In an early seminal work, Knight (1921), describes "risk" as a word ordinarily used in a loose way to refer to any sort of uncertainty viewed from the standpoint of the unfavourable contingency but later insists that "*uncertainty must be taken in a sense radically distinct from the familiar notion of risk"*. More recently, Ross and Williams (2013) describe risk as "*the threat or possibility that an action or event will adversely or beneficially affect an organisation's ability to achieve its objective."* They further qualified this by suggesting that risk is "*the consequence of a hazard, measured as the likelihood of the hazard and its severity"*, should that hazard occur. Uncertainty is generally considered as a situation where the likelihood or measure of exposure of an event is unknown. As this thesis is concerned with construction cost, the bounded and simple definition of risk as the *measure of exposure and likelihood to financial loss* will be used.

The nature of construction projects makes them particularly prone to the effects of risk and uncertainty – they are unique, complex and dynamic; each project has many parties with differing business and project objectives; projects are exposed to the weather (not in a controlled environment); ground conditions are largely unpredictable

and total project duration can spread over several years. It is no surprise then that risk has been heavily cited as one of the main causes of failure to meet cost targets on construction projects (Akintoye and MacLeod 1997, Creedy 2006). Arguably, the construction industry is perhaps one of the most risk prone industries, with project cost being one of main areas susceptible to its effects. Almost all types of risk (including scope changes, inclement weather, unsuitable ground conditions, contractual arrangements, disputes, client's cash flow problems, etc.) present some financial ramifications.

Even though risk and uncertainty seems to pervade the construction industry, both Baccarini (2005) and Burger (2003) suggest that all too often, they are either ignored or dealt with in a completely arbitrary manner using rules-of-thumb or percentages - the so-called contingency fund. Flanagan and Norman (1993) thus assert that the task of risk management is often so poorly performed, that far too much risk is passively retained, ultimately resulting in cost escalation during project delivery.

However, can a process that combines intuitive judgement and forecasting of future events ever be precise or unbiased? A qualified "no" is probably the answer to that question, according to Kahneman and Tversky (1979), formulators of *Prospect Theory* – decision making under risk and uncertainty. This theory suggests that with little or equivocal information, people tend to make decisions based on the likely gains, or loss, of a venture, and not necessarily based on the real outcome of the decision. Kahneman, a Noble Prize winner for his works on decision making and behavioural economics, delineates decision making and the illusion of understanding, stating that we often exhibit an excessive confidence in what we believe we think we know about any situation, and that our inability to acknowledge the full extent of our ignorance and the uncertainty of the world we live in makes us prone to overestimate how much we really understand (Kahneman 2011). We generally tend to disregard or underestimate the likelihood

or severity of possible risk events. Kahneman's theory holds profound extensions for decision making in the construction industry, especially for large public projects where the cost of risk and uncertainty are particular heightened. It would also provide large support for Bent Flyvbjerg's recent works on strategic misrepresentation and optimism bias.

2.2.2. Strategic Misrepresentation

Some authorities on the subject of cost overrun, including Flyvbjerg *et al.* (2002, 2005) and Wach (1989, 1990), collectively referred to in this thesis as the PsychoStrategists[1], propose more depressing explanations to the phenomenon of cost overruns. They suggest that strategic misrepresentation, the deliberate distortion or misstatement of the amount of time and resources necessary to deliver the project, is possibly the main source of cost overrun, 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 compared the cost of projects at the time of the decision-to-build to the cost at completion and found apparent discrepancies in cost forecasts for transportation infrastructure projects to be on average 45% for rail, 34% for bridges and tunnels, 20% for roads. Nine out of ten projects in their sample outrun their cost targets with infrastructure projects having an 86% probability of overrunning their cost targets.

[1] This term PsychoStrategists is coined here to collectively refer to the proponents of the *psychological contributors* (e.g. optimism bias) and *business strategy* (e.g. strategic misrepresention) as the main sources of cost overruns.

They also observed that cost overruns in their sample were not randomly distributed but were systematic, leading them to rather controversially conclude that the cost estimates used to decide whether projects should be given the go-ahead were "highly and systematically misleading"(Flyvbjerg *et al.* 2002: page 279), with strong claims of foul play by project promoters. In order to get a project approved, Flyvbjerg claims that sponsors and estimators, especially on public works, tend to intentionally underestimate the true cost of the project.

Accordingly, only projects that fit the formula,

"*Underestimated costs + overestimated benefits = funding*"

are able to secure approval, and therefore funding (Flyvbjerg 2009). They posit that

> *"by routinely overestimating benefits and underestimating costs, promoters make their projects look good on paper, which helps get them approved and built"* (Flyvbjerg *et al.* 2005).

Wach (1989) was even more forthright in his paper '*When planners lie with numbers*' and later advocated for better ethics in forecasting for public works (Wachs 1990).

There was strong evidence in support of strategic misrepresentation on the Scottish Parliament. This is subject of the paper *"Exploring escalation of commitment in Constuction project management: Case study of the Scottish Parliament project"* (see Ahiaga-Dagbui and Smith 2014c) [attached as Appendix A4]. Five weeks after their election in 1999, the new Members of the Scottish Parliament (MSPs) had to vote on whether or not to continue the project. At this stage, Alex Salmond MSP, leader of the main opposition party wrote to Sir David Steel MSP, the Presiding Officer of the Scottish Parliament, requesting that the project be

suspended and that an estimate of possible cancellation cost be produced "*in order to properly debate the future of the Holyrood project or other alternatives*"[1]. He wrote in a follow-up letter,

> *"It is now possible that we may have to consider cancelling the Holyrood project; in the circumstances it is essential that no further actions should be taken which would add to the cost of cancellation if this were the decision which Parliament reached."*[2]

Faced with the dire prospect of possible project cancellation, civil servants in the Scottish Office, led by the Project Sponsor, decided to cover-up the fact that costs were going to be significantly higher than what the MSPs were to vote upon. In a classic example of strategic misrepresentation, the Project Sponsor did not include an extra £27million for risk in the estimates submitted to the MSPs.

The proposed vote urging a termination of the project was defeated by only three votes. Alex Salmond MSP, later told the public enquiry that followed the controversies surrounding the project that the vote was based on false information, adding that

> *"it is inconceivable that had the proper information been given to the members of the Scottish Parliament, that there wouldn't have been at least a delay for taking stock and reassessment... the figures, the facts, the timeline shows that when the Parliament were told they were inheriting a project of £109 million, it was actually well*

[1] Documentary evidence number MS/1/083, submitted to the Public enquiry following the controversies that surrounded the delivery of the project in 2003.

[2] Documentary evidence MS/1/084

over £200 million and was totally out of control... Parliamentarians being misled and misinformed is a very serious issue indeed."[1]

Lord Fraser, who chaired the enquiry, backs a case for strategic misrepresentation on the Holyrood Project by stating:

"*As at the point of hand-over, where there is a very tight vote in the Parliament on whether to proceed with this particular project or not, that figure was specifically kept away from them. It looks rather as though, those who were involved in this were determined to keep the figure down as low as possible, even to the point of concealing it from the Parliament, in the hope that the project would go ahead."*

2.2.3. Optimism Bias

Further developments of the strategic misrepresentation perspective by Flyvbjerg led to theories 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 (2008) draws on this concept and suggests that decision making in policy and infrastructure planning is flawed by the fact that we think we know, or at least are in control of all possible chain of events from project inception to completion. This only leads to unjustifiable confidence in the prospects of the project and unrealistic estimates. Unlike strategic misrepresentation, optimism bias might not be buoyed by deceptive intent, but also often leads to underestimating true cost, overestimation of benefits, and overlooking the potential effects of error and uncertainty. The potential gains of the

[1] Transcript of the Public Enquiry on 13th November 2003 [Available at www.holyroodinquiry.org]

project thus become overwhelmingly enticing and almost blinding to likely pitfalls.

According to the PsychoStrategists, deception (strategic misrepresentation) and delusion (optimism bias) are complementary explanations of the failure of large infrastructure projects leading to the underestimation of likely final cost of project. 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. This line of diagnosis of the problem of cost overrun might seem appealing, at least on cursory examination, especially in terms of large capital intensive public projects or those that are likely to make high political statements.

There is some evidence to support this supposition on the Scottish Parliament project. The unrealistic cost ceiling of £40million in the Government's devolution White Paper (Scotland's Parliament 1997) turned out to be a rather optimistic estimate, or better still, a guesstimate of final cost of the project by non-construction professionals. A member of the Scottish Parliament Corporate Body, Andrew Welsh MSP, stated that "*right from the very start, the budgets were totally unrealistic. The original budgets we inherited were for a fictional building*"[1]. Rusell Hillhouse, former Permanent Under-Secretary at the Scottish Office and a member of the team that estimated the cost of the project at £40million said:

[1] Transcript of Public Enquiry on 11th February 2004, [Available at www.holyroodinquiry.org]

> "*we couldn't possibly have done a thorough job, and this was very difficult because it was a time when people were working extremely hard on other aspects of the White Paper*"[1].

Sam Galbraith, former Under-Secretary of State at the Scottish Office also told the public enquiry,

> "*the figure of £40million in the white document, was never for Holyrood. That was for a bog-standard building on a greenfield site."*[2]

The 14,000 capacity entertainment and sports complex, Perth Arena in Australia also experienced considerable cost growth during its delivery. The project was approved at a cost of A$160 million but was eventually completed three years behind schedule at a reported A$550 million (Auditor General of Western Australia 2012). Citing optimism and poor project governance as causes of this apparent cost growth and delay, the Auditor General said in media statement, "*the initial estimates of the cost and opening date for the Arena were unrealistic and made before the project was well understood or defined*." (Auditor General 2010, pg. 1). Similarly, the Fiona Stanley Hospital project in the same country was approved at A$420 million, but experienced a staggering cost growth of A$1300 million. In a similar diagnosis, the Auditor General again stated that "*the original estimates were unrealistic*." (Auditor General of Western Australia 2012, pg. 50).

2.2.4. The Evolution Theorists

Another school of thought on cost overruns, referred to here as the Evolution Theorists[1], include Love *et al.* (2012, 2014), Osland and

[1] Transcript of Public Enquiry on 30th October 2003
[2] Transcript of Public Enquiry on 28th October 2003

Strand (2010) and Odeck (2004). Their thesis statement is straightforward - projects change, and when they do, they often come with increasing costs. 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.

Love *et al.* (2012) provide a rebuttal to Flyvbjerg's perspective on cost overruns in their paper "*Moving beyond optimism bias and strategic misrepresentation"* suggesting that industry should rather embrace a more holistic understanding of the escalation phenomenon that includes some level of the process and the social construct. Love *et al.* (2012) introduce the concept of 'pathogens', 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 drivers of cost growth. They further argue that Flyvbjerg's analyses are perhaps rather too simplistic and not generalisable to all projects undertaken within the industry. Love *et al's* (2012) case study of social infrastructure projects suggests that foul-play might not be the best explanation of cost overruns and that the fingers point at other events that occur before and during the project delivery stage, including rework and design changes.

Osland and Strand (2010) were also very critical of the strategic misrepresentation perspective of cost overruns. They questioned the theoretical and methodological validity of Flyvbjerg's work, claiming that the strategic misrepresentation framework "does not offer any

---

[1] The term Evolution Theorists is used here to describe the school of thought that is predominnantly concerned with *changes* that occur during project development, from inception to completion. The central argument of the supporters of this school is that projects change (i.e. evolve). The changes then lead to the differences betweeen expected and actual final cost of the project.

variation on the institutional variable nor when it comes to variation in planners (actors) motives and rationality."

They further argued that for

> *"Flyvbjerg and other proponents for the hermeneutics of suspicion, the actors actually admitting telling lies can be seen as the 'tip of the iceberg'. However, it is also a perspective that would not be falsified if no examples of actors admitting lying were found. On the contrary, it could easily be interpreted as a verification that they were lying also for the researchers."*

Osland and Strand's rebuttal is probably sustainable as it is almost impossible to draw valid distinctions, along a continuum of motivation, between reasonable optimism, over-enthusiasm to deliberate deceit or culpable error using statistical analysis, as adopted in Flyvbjerg's works. Furthermore, adopting a positivist perspective to understand a complex issue like construction project governance, which usually involves a complex interplay of construction professionals, planners, business strategy, institutional framework and politics, would merely be superficial at best and never actually provide substantial evidence to support the kind of conclusions reached by Flyvbjerg *et al*.

Odeck (2004) invested the size and causes of overruns for road projects in the Norwegian construction industry. The study reported an average cost overrun of 7.9% with a maximum cost overrun of 183%. The study however attributed these overruns to project specific factors such as project duration, location, estimated cost at contract award and size of projects. An interesting observation from this study is that cost overruns appear to be more predominant among smaller projects just as larger ones.

Love *et al.* (2005) previously conducted a questionnaire survey of 161 construction professionals in the Australian construction industry and found that rework was one of the main contributors to escalation of cost. The main sources of rework as found in their work are ineffective

use of information technology, staff turnover/allocation to other projects, incomplete design at the time of tender, insufficient time to prepare contract documentation and poor coordination between design team members. This conclusion is similar to that reached by Bordat *et al.* (2004) who found that the "dominant" source of cost overrun was change order due mainly to "errors and omissions" in design. In a more recent research, Love *et al.* (2014) challenged Flyvbjerg's strategic misrepresentation and optimism bias perspective as lacking in verifiable causality, and therefore limited in their application.

2.2.5. Relay Races and Project Governance

Research on leadership and governance of construction projects by Gil and Lundrigan (2012), perhaps offers a more holistic assessment of cost growth that aligns with the views of Love, *et al.* above. That projects evolve is essentially the core of their defence. Very often, construction projects change considerably in scope and design between conception, to inception and completion, often due to a client's proposed changes or technically imposed changes. This suggests that it might be erroneous to simply compare the cost of a project at inception, A, with the cost at completion, B, and wherever B>A, then overruns have occurred and estimators of A either lied or were incompetent. A and B are essentially very different. More robust explanations of overruns need to factor-in process and product, as well as sources of changes to scope. For Love and Gil *et al.* (*op. cit.*), project overruns are not really a case of projects not going according to plan (budget), but the other way round – plans not going according to project.

Gil and Lundrigan (2012) propose a "relay race" framework for understanding cost growth, particularly on mega projects such as the London Olympics Project, Scottish Parliament or Terminal 2 project at Heathrow Airport, all of which seemed to have suffered the curse of cost growth, at least on a perfunctory examination. In the relay race of construction delivery, the baton of project leadership is passed on from

one person(s) or organisation at the different stages of the project delivery. The aims and scope of the project, as well as skills and competencies of the project sponsors and promoters (project governors) at the conceptual stage, are often very different from their counterparts at the project design or delivery stage. Also, it is not unusual for most public projects to have long gestation periods, stretching over several years before final approval is reached, by which time project budget would also have changed a number of times. The Scottish Parliament Building is a paragon in this respect – the *circa* £40 million submitted by the Scottish Office as likely final cost did not take into consideration project location, VAT, fees, inflation or the building of a completely new parliament building. It is no wonder the final cost of the project was 10 times this initial proposed estimate.

2.3. RETHINKING COST OVERRUNS

The literature is clear: in the arena of cost escalation there are essentially two prevalent schools of thought, referred to in this thesis as the PsychoStrategists and the Evolution Theorists. Much debate in the existing literature concerns which is the correct view of the cost overrun phenomenon, which is the most practical, or most relevant. Could both actually be valid and complementary? Or maybe they are just different facets of the same problem? Some of these issues are explored in the papers "*Dealing with construction cost overruns using data mining*" (Ahiaga-Dagbui and Smith 2014b) and "*Rethinking construction cost overruns: Cognition, learning and estimating*" (Ahiaga-Dagbui and Smith 2014a) which were published in Construction Management Economics and the Journal of Financial Management of Property and Construction [attached as Appendix A1 and A2 respectively].

Hitherto, the term "overruns" has been used as though it conveyed an unequivocal meaning. Existing literature on cost overrun seems to conflate two related, but different issues – overruns and underestimation. This may largely be due to the fact that there is no

unanimity on the reference point used to measure what is loosely referred to as overruns in different studies. Much of the media hype on supposed cost overruns hardly makes this differentiation as well and thereby would often base their reportage on a simple comparison between cost at inception and cost at completion of a project, ignoring the mediating phases of project gestation and definition. It is therefore unsurprising that there is often a rather large disparity in the level of cost overrun reported in these researches.

For example, while Flyvbjerg *et al.* (2002) report an average cost overrun of 45% for tunnels and bridges, Hinze *et al.* (1992) and Love *et al.* (2012) report lower averages of 4.24% and 11.89% respectively. Furthermore, Odeck (2004) reports an average of 7.9% overrun for 420 road projects while Vidalis and Najafi (2002) report an average of 10.52% for 708 road projects. Flyvbjerg *et al.* (2002) however report a higher average of 20% for 167 road projects. This is because Flyvbjerg *et al.* use a reference point of cost at when project was approved while the other studies generally lean towards cost at contract award.

As already pointed out, most large publicly funded projects tend to go through a long definition period after project inception during which many changes to scope and accompanying costs occur. Sometimes the initial scheme bears little likeness to the defined project, as was the case of the New Children Hospital in Australia. The initially approved budget for the hospital was A$207 million. The scope at this stage was to relocate the Princess Margaret Hospital to the Royal Perth Hospital. However, this scope completely changed during project definition to the construction of a totally new Medical Center at A$962 million, a cost increase of A$755 if taken on cursory examination (Auditor General of Western Australia 2012). The Holyrood Project in Edinburgh also experienced a similar significant scope change, and thereby the astonishing cost growth recorded (see Audit Scotland 2000, 2004). It seems erroneous therefore to make direct comparisons between the initial “estimate” A and its final completion cost B – they are

comparisons between two very different projects. More robust explanations of cost growth would need to factor-in process and product, as well as sources of change to scope. Flyvbjerg's works make a direct comparison between costs A and B, and wherever B>A, overruns are reported.

In Figure 2, a conceptual model that divides the project delivery cycle into three sequential stages is presented. In practice, some projects are now procured with the definition and construction phases occurring concurrently. Suffice for now though that most projects follow Figure 2. As noted by Ahiaga-Dagbui and Smith (2014a), the scope of the project at the inception is often just rough ideas, schemes and concepts drawings, lacking any firm information for reliable cost estimation. This stage in some cases can be several years in advance of contract award and eventual construction. 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. The estimate at this stage often does accommodate a lot of details and information, largely because much of these are not yet available or uncertain.

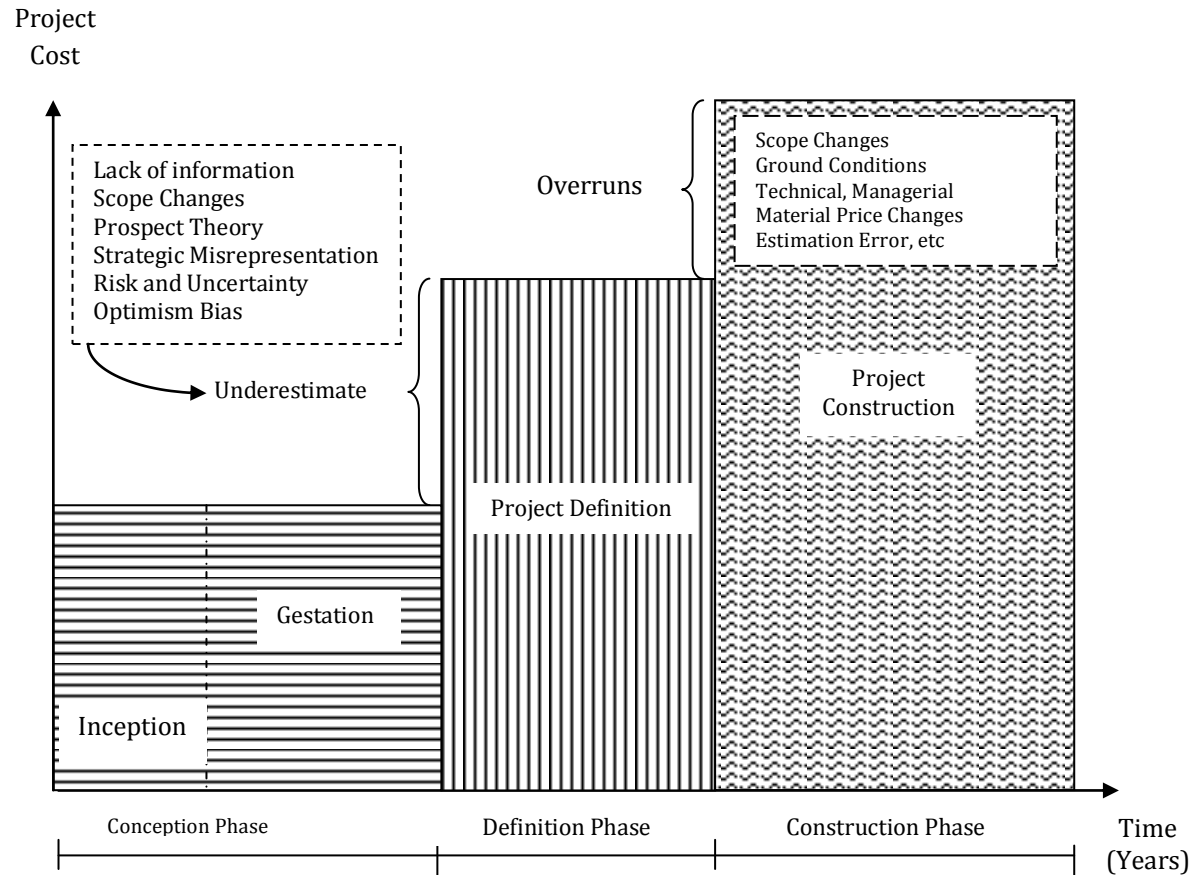


Figure 2: Conceptual model for understanding cost growth

However, it is often at the inception stage that project sponsors garner for green-lighting and funding. Arguably, it is perhaps at this stage the effects of uncertainty, lack of information, optimism bias and strategic misrepresentation might be particularly heightened, to keep cost at an attractive low to secure project approval. Significant cost growth usually occurs soon after project approval when the project business case is being developed and the project scope and costs were more accurately defined (Definition stage in Figure 2). In the case of the 20 capital intensive project analysed by the Auditor General of Western Australia, approximately 90% of the cost variance recorded, equivalent to A$2.95 billion, occurred before the construction phase (Auditor General of Western Australia 2012). In Figure 2, the difference between the cost estimated at project inception, where the decision-to-build is taken, and the estimated cost at the end of the project definition stage is referred to as cost underestimate. The main contributors to cost underestimation at this stage may be a lack of information, significant

scope changes, estimation error, strategic misrepresentation or optimism bias.

Overruns, however, are more appropriately described here as the difference in cost at project completion and project definition stage (see Figure 2). This is usually as a result of further scope changes, normally not as significant as those at project definition stage, rework, ground conditions, technical and managerial difficulties, material price changes or estimation error. These are the factors that Love *et al.* (2012) describe as "pathogens". So, whereas the PsychoStrategists (Flyvbjerg *et al.*) mainly deal with underestimation, the evolution theorists (Love *et al.*) focus on the latter phases of the construction project that contribute to overruns. Both perspectives are actually two complementing sides of the same coin.

Herein lies the dilemma then: which baseline should be used to measure project performance - the initial cost target at project approval and decision-to-build, or the cost at contract award after project definition stage? From the overwhelming statistics already presented in this thesis, it would seem that the cost estimate at the time of the decision-to-build is of little value to actual project delivery, at least from a theoretical point of view, as scope and design of projects seem to change very quickly after this milestone. From a practical perspective, it should also be noted that the cost at project approval is very crucial: it is upon this figure that funds are allocated, among competing opportunity costs. Therefore, even though there is good reason to justify using the cost estimates after project scope definition and contract award as the reference point for performance measurement, there could be a real risk to accountability if scope changes automatically led to resetting of the baselines. It is therefore crucial for the industry to find more effective ways of project approval and governance that better deals with underestimation of true cost and the setting of unrealistic cost targets.

Furthermore, focussing only on one side of the debate in dealing with overruns will do little to effectively tackle cost overruns in the management of construction projects. PsychoStrategic theorists neglect well documented prominent issues like design problems, unforeseen ground conditions, scope changes and rework that drive up cost during the actual project delivery. The unfortunate consequence of this perspective will be to brand planners, project promoters and estimators as unethical and suspicious without sufficient evidence to sustain the supposition. An evolutionary theorist perspective alone, on the other hand, would also rather naively not fully accommodate the strong influence and dynamics of business strategy, competition, power and organisation politics in setting unrealistic cost targets that will inevitably be unattainable.

2.4. CHAPTER CONCLUSION

This chapter has presented an overview of the scale of the problem of construction cost overruns as a global phenomenon that is indifferent to the size of the project under construction, the geographical location, duration or type of project. Two predominant schools of thought on the causes of construction cost overrun have been presented, referred to in this thesis as the PhychoStrategic and Evolution Theorists. The tension between these schools was discussed before presenting a conceptual model that holds the two perspectives as different sides of the same coin. It is unlikely that the construction industry will be able to adequately deal with the problem of cost overruns if only one of these perspectives is focussed upon. The way forward is to find ways of construction procurement and project governance that effectively circumvents intentional underestimation and unjustifiable optimism at the project definition stage which only ends up setting unrealistic cost targets. This must however be carried out in tandem with considerations of the adequacy of detailed project design before construction and thorough quality assurance regimes during project delivery to curb the cost of rework and unwarranted scope creep.

CHAPTER THREE

RESEARCH APPROACH AND EXPERIMENTAL DESIGN

"Research is formalized curiosity.

It is poking and prying with a purpose."

~ Zora Neale Hurston

(Folklorist, Anthropologist and Author, 1903-1960)

3.0 INTRODUCTION

No thesis that is concerned with the use of Artificial Neural Networks to model construction cost can be anything other than predominantly quantitative. So while the philosophical position of this research is post-positivisit and the approach primarily quantitative as it concerns numbers and models, it is important to place the study in the context of *all* research methodology. Indeed it may be that not all the work is quantitative, some elements needing a more qualitative approach, with an interpretive analysis. Even in the straightforward positivist world of numerical researchers there are numerous approaches and modelling techniques that can be utilised. Once the approach is established, the experimental design must be set out and the nature of the data modelling technique, in this cases Artificial Neural Networks, must be described.

This chapter will therefore present the considerations for adopting the research approach used as well as the experimental framework that will guide the modelling aspect of the research. The background, applications, strengths and weakness of artificial neural network will also evaluated in this chapter.

3.1. RESEARCH APPROACH

When choosing which approach to use, the critical considerations are the type of problem under study and how to maximise the chances of adequately answering the research question(s) of the study. There are essentially two traditional research approaches: quantitative and qualitative. According to Fellows and Liu (2008), "*quantitative approaches tend to relate to positivism and seek to gather factual data, to study relationships between facts and how such facts and relationships accord with theories and the findings of any research executed previously (literature).*" The quantitative approach thus adopts a scientific method where the study of theory and existing literature on the subject results in precise aims and objectives with propositions and hypotheses to be tested by examining the relationship between different variables.

Creswell (2009) prescribes three criteria for choosing a quantitative approach; if the problem calls for:

a) the identification of factors that influence an outcome; or

b) the measureable utility of an intervention; or

c) understanding the best predictors of an outcome.

The central tenets of the quantitative approach are measurement, causality, generalisation and replication, according to Bryman (2012). Measurement concerns the quantification of the degree of relationship between concepts or variables while causality refers to the connection between a set of variables (causes) and an observation or phenomenon (effect). Generalisation on the other hand deals with the extent to which the inferences and findings in a particular sample can be extended to other populations and settings while replication involves the ability to repeat a study using the same methods, different subjects and different experimenters. Replication is essential to lend credibility, or otherwise, to the results achieved in a study and to test the generalisability of the findings. This requires that the results of the initial research must be unaffected by the researchers special characteristics or expectations or biases.

In qualitative study however, there are usually no *a priori* propositions, so that the objective of the study is to gain understanding that may lead to formulation of theories. The qualitative study seeks to find out what people's perception are of different issues, or what meanings people might ascribe to different social or human phenomena (Creswell 2009). It might involve ethnographic strategies where the researcher studies an intact cultural group in their natural setting over a prolonged period by collecting observational or interview data. Alternatively, for example, it could take the form of a discourse analysis, where the researcher studies and analyses written or vocal work in an attempt to extract meaning and understanding for theory building. Fellows and Liu

(2008) thus note that in some way, a good qualitative study often forms a prelude to quantitative methods.

Somewhere between these two traditional classifications of research approach lies the mixed method. Creswell (2009) notes that the mixed method essentially combines elements of both the qualitative and quantitative approaches in order to adequately address a particular problem, and observes it is more than just simply collecting and analysing both kinds of data. It is used so that the overall strength of a particular enquiry is greater than if only the quantitative or qualitative approach is employed.

The research reported in this thesis largely adopts a quantitative approach with some elements of qualitative approach especially in the early stages of the research. A thorough exploration and critique of existing literature on construction cost overruns laid the basis on which to conduct a *quasi* experimental model development to estimate likely final cost of water infrastructure projects. The research framework adopted in this research has been mapped out in Figure 3. As seen in Chapter 1, the research began by defining the aim and objectives of the research before a thorough review and critique of existing literature and theories on the cost growth phenomenon. The literature review led to firming-up and clarifying the initial aim and objectives. An experimental approach was then designed to provide the framework for data collection, analyses and subsequent model validation. The results of the model development were then discussed before reaching conclusions and recommendations for achieving more reliable cost estimates at the tender stage of construction projects.

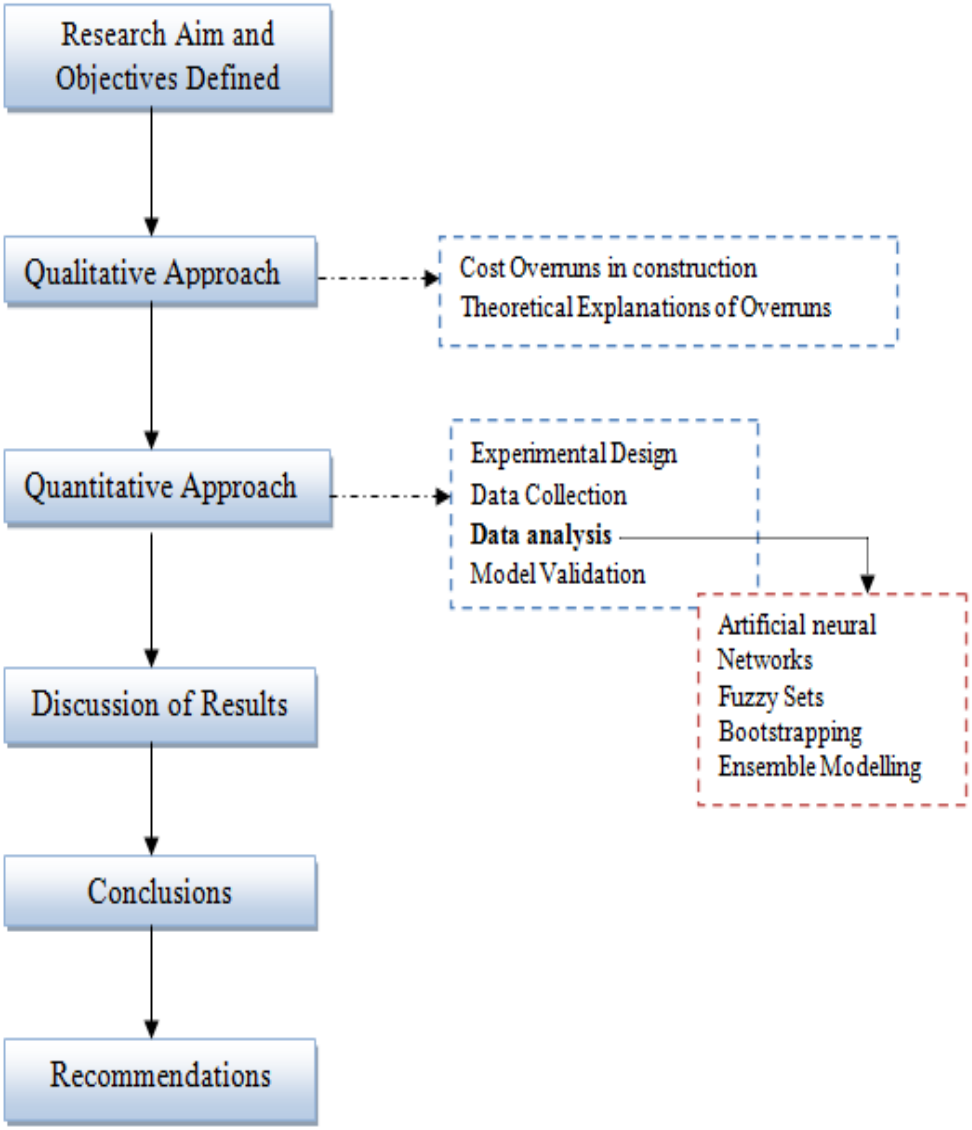


Figure 3: Research Approach adopted

3.1.1. RESEARCH DESIGN

Yin (2009) describes research design as the structure that guides the collection and subsequent analyses of data. It enables the researcher to connect empirical data and conclusions to the initial research question of the study in a logical sequence. The research reported in this thesis follows a *quasi* experimental approach with the collection of historical cost data on 1,600 water infrastructure projects followed by the development of cost models mainly using artificial neural networks.

3.1.2. EXPERIMENTAL DESIGN

After an initial exploration of the problem of cost overruns and clearly defining the aim and objectives of the research, a data mining approach was adopted for the actual cost modelling stage. Data mining allowed for the conversion of information embedded in historical cost data into decision support tools for final cost estimation at the tender stage of a project.

StatSoft Inc. (2008) describe data mining as 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. Data mining attempts to scour databases to discover hidden patterns and relationships in order to find predictive information for business improvement.

As already identified in earlier chapters in this thesis, early cost estimation is often hampered by the paucity of reliable information of accurate estimation. Project approval also tends to precede the availability of detailed designs and contract award. Using data mining techniques, it is possible to use the information that is already available in a construction firm's database to build cost models to support the estimation process in the early stages of a project. Producing reliable estimates at this stage is crucial because project feasibility, approval, budget allocation and contract award usually occurs at this stage of the project.

Figure 4 details the various stages followed in the experimental phase of this research to develop the final cost models.

3.1.2.1. Data and Business Understanding

It is always important to understand the application domain of the problem under study. If the data mining is being carried out within the context of a particular firm and its business structure, it is crucial to

understand the data within the framework of that firm. For example, one of the collaborating firms in this current research uses a stage case system called CAPEX. The CAPEX codes range from 1 to 6, corresponding to rough estimates and final account costs in more traditional terms.

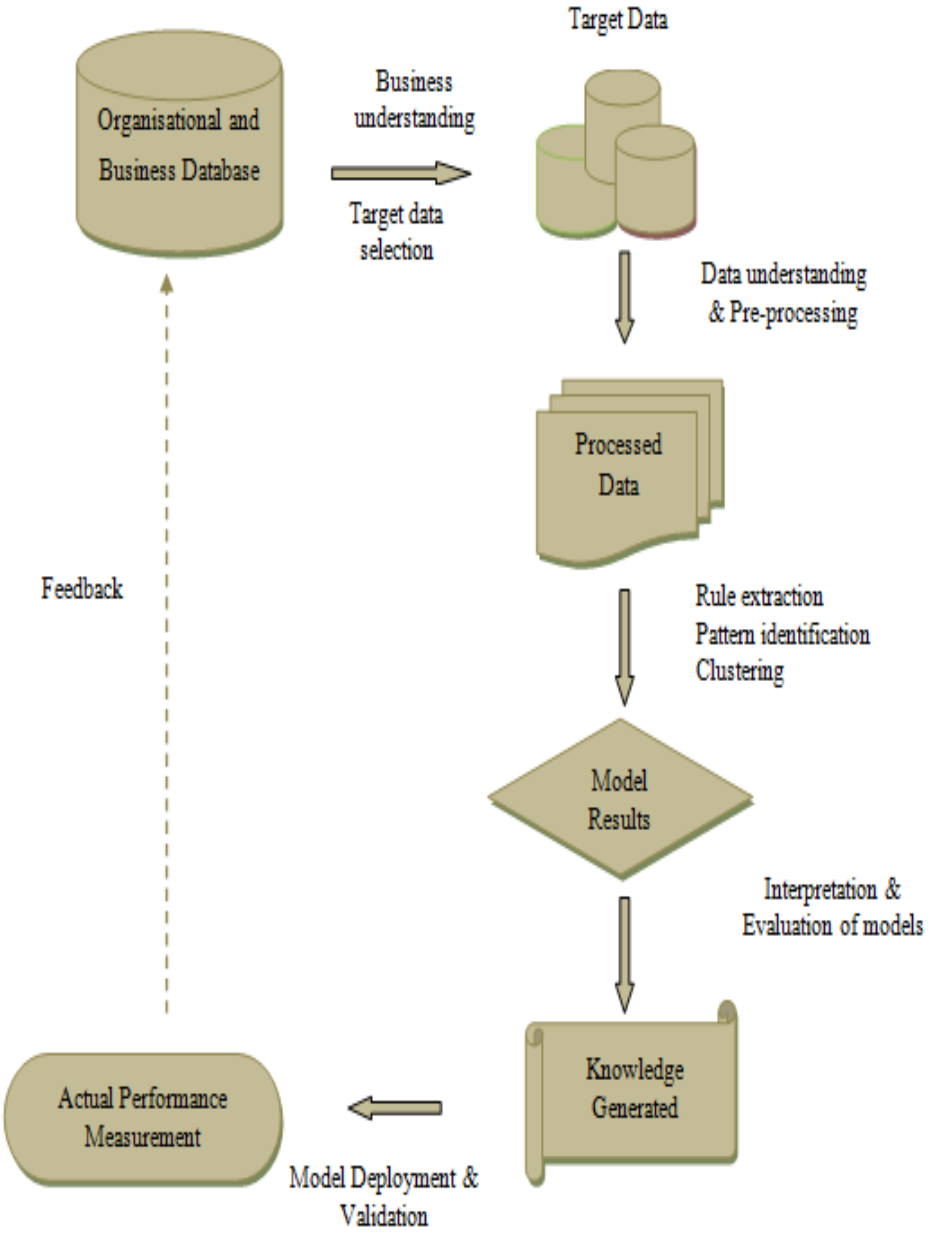


Figure 4: Experimental Design Procedure

3.1.2.2. Selection of Target Data

After understanding the application domain and database for the research, the next stage usually involves the selection of target data from the main database. Some level of experience, intuition and expertise is usually required to guide the initial decision of variables that are likely to be influential in predicting final cost. As a rule of thumb though, it is important that any factor with even the slightest likelihood of explaining the variability in the data is included in the model at the start. Part of the modelling process will involve pruning these variables down to an optimum number.

3.1.2.3. Data Pre-processing

The target data is then pre-processed before the modelling proper. The aim of the pre-processing is to structure and present the data to the model in the most suitable way as well as to offer the modeller the chance to get to know the data more thoroughly. Pre-processing might involve simple steps such as removing of duplicate entries and missing data treatment to more advanced techniques like clustering, transformation and de-noising[1]. It might be important at this stage to evaluate basic statistics such as means, modes, cross-tabulations and standard deviations, or use plots such as histograms, bi-variate and scatter plots to provide an initial understanding of the nature of data that is being used for the modelling. The importance of the pre-processing stage to the success of the final model cannot be overemphasised as it offers the modeller the opportunity to have a

[1] Noise is unexplained randomness or variations in data. Noise results in the loss of generalisation as data patterns are not constant or replicable (Statsoft, 2008). Construction cost data, for example, is particularly noisy as even the same building built at different locations might have appreciably different costs.

good understanding of the problem under investigation, explore the kind and type of data that is available for the modelling, and identify problems within the data such as missing data or outliers.

3.1.2.4. Actual Data Modelling

The next stage of the data mining approach involves the actual modelling for the discovery of patterns, clusters or relationships within the dataset. 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, neural networks, decision trees, genetic algorithm and fuzzy logic. Table 2 provides a guidance for some issues and options to consider when selecting a particular modelling technique. The main issues to consider are usually around the aim of the modelling exercise, the predictive performance required or 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, neural networks on the other hand, do not produce any equation. Neural networks have thus been derided as 'black-boxes' by some researchers – for instance by Sarle (1994) who is a statistician. However their ease of use, 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).

Table 2: Some issues to consider when selecting a data modelling 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 Self-Organising maps Genetic algorithm, etc	Flexibility Accuracy (Precision) Power "Interpretability" Ease of deployment

Artificial Neural Network, discussed in more detail in the next section, was the main modelling technique used in this thesis. This was combined with fuzzy sets theory to develop Neuro-Fuzzy Hybrid models in further iterations of the model development. Data bootstrapping, sub-sampling and ensemble modelling were also experimented with in this research. Details of all these techniques with their accompanying results are provided in later sections of the thesis.

3.1.2.5. Result Evaluation and Presentation

After the actual data modelling stage, the results achieved are then evaluated and presented in a meaningful form to aid business decision-making. This step might involve graphical representation or visualisation of the model for easy communication.

3.1.2.6. Model Validation and Feedback

Reliability and confidence in the use of the model can be improved if it can be shown that the model performs satisfactorily when new queries

or situations are presented to it. This can be achieved through testing and validation of the models using unseen[1] data before deployment in real life situations. For continuous improvement purposes, model performance is usually fed back to the database so that incremental learning can be achieved in the model.

3.2. ARTIFICIAL NEURAL NETWORKS

Artificial Neural Networks (ANN) is a simplistic abstraction of the biological neural networks of the brain with the capability for information processing. The cost prediction models developed in this thesis are based on neural network techniques. This section of the thesis provides a brief overview of ANN, its application in civil engineering and construction management, as well as some advantages and limitations.

3.2.1. Brief Background

The development of neural networks was motivated by the desire to both understand and emulate how the brain functions. Often just referred to as Neural Networks (NN), with artificial implied, NN importantly retains two important features of the biological neural network, i.e. the ability to learn from experience (Hinton 1992) and make generalisations based on this acquired knowledge (Haykin 1994).

McCulloch and Pitts (1943) are generally credited for outlining the first formal model of a basic computing neuron after which Hebb (1949) identified how information can be stored in a NN. Hebb also proposed how the neuron's connection weights can be updated through different learning schemes. Rosenblatt (1958) thereafter developed the

[1] Unseen data is data that has not been used in the model development and therefore can be used as independent assessment of the models predictive performance.

perceptron, a hypothetical nervous system that is capable of learning so as to be able to store, recognise and influence behaviour or decisions.

3.2.2. Neural Network Structure

Figure 5 shows a basic neural network structure with 1 input layer with 3 units, a hidden layer and an output layer with 2 units. The input layers accept data presented to the network and assigns weights (w_1, w_2, w_3) according to the relative importance of the information. In the case of cost data, weights are apportioned according to the sensitivity of each factor to the overall cost estimate. Neural network's computations happen within the hidden layer, which also becomes the permanent memory of the model after training for predicting new cases. The predictive performance of the neural network can be increased by increasing the number of hidden layers or nodes in the hidden layer, although this must be kept as low as possible to ensure the network does not just *memorise* the data instead of *learning* the underlying patterns and correlations within it. Memorising in model development is called overfitting (Haykin 1994). An overfitted model generally tends to perform very well during training, but performance drastically deteriorates when new data is presented to the network to validate it.

The output layer receives and stores or transfer processed data from the network (Fausett 1994). The output of the neural network is a function of the weighted sum of all neurons in the network, a completely deterministic result.

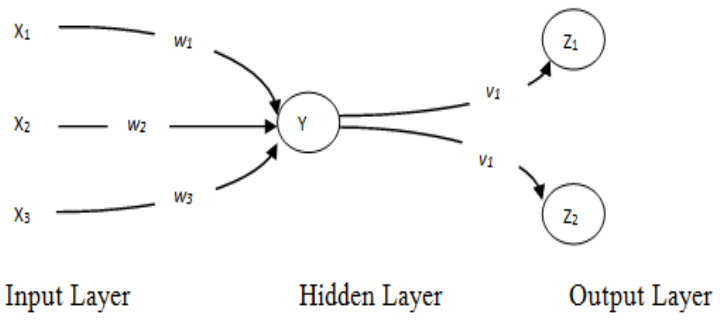


Figure 5: Basic neural network architecture.

3.2.3. Training

Similar to the brain, NN learn by examples. *Learning* in NN terminology has been defined by Haykin (1994) as "*a process by which the free parameters of a neural network are adapted through a continuing process of stimulating by the environment in which the network is embedded*." Training the neural network is thus a process of iteratively adjusting the network weights and threshold until it is able to learn, i.e. approximate the underlying functional relationship between inputs and target(s). The learning period of a NN is bounded by rules that govern how weights adapt in response to a learning example, how many times the learning process is carried out, etc.

A training dataset, consisting of input-target pairs are presented one after another to the network during training for learning. It might be easier to view training as a question and answer session, with inputs as 'questions' and targets as 'answers'. During learning, the network is effectively being asked questions (inputs). Given the current question, the network error is measured by comparing the answer produced by the network (output) with the actual expected answer (target). Should the performance not be satisfactory, the network weights are adjusted to produce a more correct answer on another attempt, a process called back-propagation in NN terminology (Fausett 1994).

The type of learning paradigm adopted is determined by the manner in which the network weights and threshold changes take place during training. There are essentially two main training techniques for learning in neural networks, i.e. supervised and unsupervised. In a supervised learning technique, the NN is supplied with both inputs and desired response (target). As illustrated in Figure 6, the output of the network is measured against the desired response (target) using a predefined performance criterion, e.g. Mean Absolute Percentage Error (MAPE), and the connecting weights are modified to minimise the model error, $\sum$. Learning is completed when an optimal solution has been found and $\sum$ is no longer modified significantly, or when a

previously specified number of iterations have been run. This is the learning paradigm used in this research.

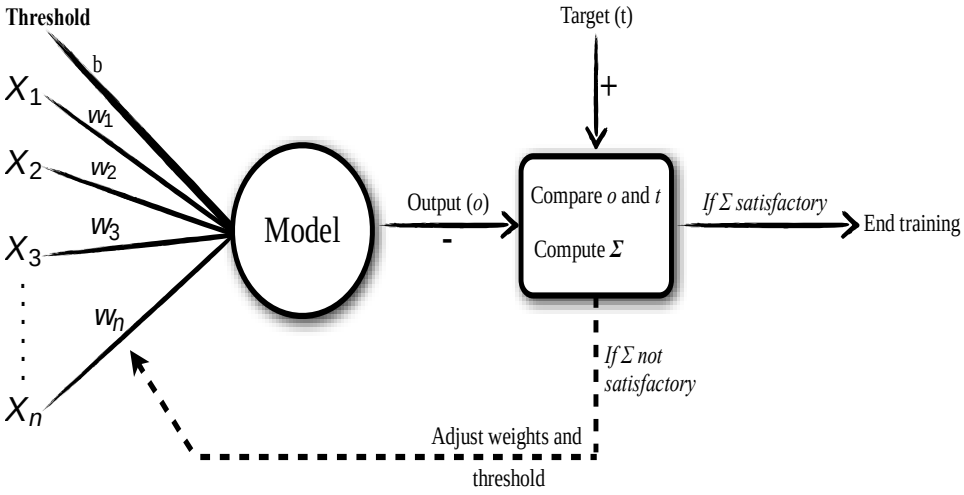


Figure 6: Supervised Learning Process

Anderson (1995) notes that no desired output (target) is given to the network in the unsupervised training regime. Instead, the network has to make sense of the data by itself without external assistance. The neural network adjusts its own weights so that similar inputs cause similar outputs. The Self-Organising Maps (SOM) for clustering and pattern recognition are some of the main application for this type of training (Haykin 1994).

Neural networks do exact their own demands however. Bode (2000), warned that NN would only be able to produce superior results in comparison with other methods of cost estimation if there is sufficient and reliable data available for both the training and validation of the network. Anderson and McNeill (1992) note that NN are desperately dependent on plenteous, representative and reliable data to be able to make accurate predictions and generalisations. Furthermore, there is no consensus within the literature on what represents a satisfactory amount of data to undertake modelling using neural networks.

3.2.4. Application of Neural Networks

Neural network lends itself to application in different problem domains largely because of its power and ease of use (StatSoft Inc 2008). It has been applied to a wide breadth of problems that involve pattern association, classification, recognition, clustering, reasoning with imprecise and incomplete data, forecasting and control. In Finance, it has been widely used for loan applicant assessment (Handzic *et al.* 2003), foreign exchange prediction (Wang *et al.* 2010), stock price prediction (Cao *et al.* 2011) among others. It has also been used for medical diagnosis research (Kodogiannis *et al.* 2008, Dreiseitl *et al.* 2009) as well as for speech and character recognition (Dahl *et al.* 2010, Pradeep *et al.* 2011).

The earliest civil engineering application of neural networks can be traced back to Adeli and Yeh (1989) on engineering design and machine learning. It has since been applied in civil engineering to estimate the elastic modulus of normal and high strength concrete (Demir 2008); estimating the compressive strength of concrete (Topcu and Saridemir 2008, Saridemir 2009); forecasting the cost of rail transit and metro track works (Gunduz et al. 2011) and passenger flow forecast (Zandieh *et al.* 2009, Wei and Chen 2012). See Adeli (2001) for a detailed review of neural network applications in civil engineering research.

3.2.4.1. Neural Networks in Construction Management

In the field of construction management, Portas and AbouRizk (1997), as well as Sonmez and Rowings (1998) conducted experimental studies to estimate construction productivity using neural networks. Portas and AbouRizk estimate labour productivity using input factors collected from the literature, primary data from formwork activities, in addition to factors from a survey involving superintendents and project managers. They compared their model results to those used by the participating firm and reported improved quality of the estimates attained. Sonmez and Rowings (1998) developed neural network

models for quantitative evaluation of labour productivity for concrete pouring, formwork, and concrete finishing tasks, using data compiled from eight building projects. Among others, they used factors such as gang size, temperature, humidity and precipitation at the time of the work activity to develop their models. They compared the neural network models to regression models and found that the neural network models were superior to the regression models in all except the formwork task.

In a similar research to that reported in this thesis, Bousabaine and Elhag (1997) explored the use a hybrid of neural networks and fuzzy logic to predict the duration and cost of construction projects. They attempted to exploit the ability of neural networks to generalise solutions from past events and combine that with fuzzy logic's ability to deal with impression and uncertainty of future events. They used a rather small dataset of only 12 projects from the BCIS database for the training regime and tested the model with 7 unseen cases. Although they demonstrated a robust approach to modelling project duration and cost, their models did not achieve satisfactory results, possibly due to the small dataset used in training the model.

Al-Tabtabai and Alex (1999) present a neural network approach to estimating the preliminary cost of highway projects using 9 project factors including type of road, soil nature, location and hauling distance. They first used case-based reasoning to collect relevant project factors on 40 different projects before modelling these factors using NN. They reported a mean square error of 9% when the models were validated with new data. In a similar research, Wilmot and Mei (2005) present a neural network approach that estimates the escalation of highway construction costs over time. The model relates overall highway construction costs to the cost of construction material, labour and equipment as well as the characteristics of the contract and the contracting environment prevailing at the time the contract was let.

They report that their model was able to replicate past highway construction cost trends with "reasonable accuracy."

Emsley *et al* (2002) developed neural network models to estimate the tender cost of building projects using about 288 project cases in the UK. They collected primary data from project files, supplementing this with details from the Building Cost Information Service (BCIS) and responses from a questionnaire survey. Their models include such project factors as number of floors below/above ground, type of stairs, gross internal floor area, project duration and number of lifts within the building. They report that their NN models were able to model the non-linear relationships within their data with a mean absolutely percentage error of 16.6%.

It is generally accepted that the success of the early stage planning of a construction project plays a crucial role in determining the success, or otherwise, of a project during its delivery. Wang *et al* (2012) developed neural network and Support Vector Machine (SVM) classification models to predict project cost and schedule success, using status of early planning as the model inputs. They adopted a simplistic measure of success of projects that report a lower actual final cost than was awarded. The same measure is used in terms of project durations. They collected early planning and project performance information from 92 building projects in Taiwan using questionnaire survey. Their NN models produced 76% and 68% accuracy for cost and duration prediction respectively. The SVM models achieved 76% and 72% accuracy at classifying successful projects.

"*My Cost runneth over: Data mining to reduce construction cost overruns*" is one of the publications from this current research which also uses artificial neural networks to estimating the likely final cost of construction projects (Ahiaga-Dagbui and Smith 2013). The modelling is based on information from almost 1,600 water infrastructure projects completed between 2004 and 2012 within the UK. The models were then developed using a combination of data mining techniques

such as factor analysis, optimal binning, and scree tests with neural networks. 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%.

3.2.5. Neural Network Criticism

One major criticism of the NN approach to data modelling is that it offers little explanation on the complex relationships between the variables it is modelling (Ripley 1993, Sarle 1994). As previously mentioned, it is thus often derided as a 'black box' technique because the network parameters (i.e. transfer functions, learning rules, network architecture, weights, etc.) do not show casual explanations, making it difficult to elucidate what is learnt from the neural network model. In regression analysis, for example, an equation whose physical properties that can be easily interpreted is produced. There is no such equation or coefficients in neural networks - the model in essence is the equation.

In an attempt to illuminate the 'black box', Olden and Jackson (2002) demonstrated the possibility of understanding variable contributions by using a randomisation test called connection weight method, a process somewhat similar to statistical pruning techniques to eliminating connection weights that do not contribute significantly to the network output. In a breast cancer related research, Ravdin and Clark (1992) excluded one variable at a time from their model to find out that particular input's contribution to the final model. Another approach to reducing the vagueness of the prediction process of neural networks is to combine it with qualitative causal descriptors of fuzzy-logic theory to create neuro-fuzzy hybrid systems. This was the approach adopted in the by Ahiaga-Dagbui *et al* (2013) for predicting the final cost of water infrastructure projects. Neuro-fuzzy hydrid models were also by Boussabaine and Elhag (1997) for tender price forecasting. Overall, as there are several modelling techniques, all with different capabilities and weaknesses, it could be argued that neural

networks should be selected for modelling when the goal is 'how well?' a model performs and not just 'how?' it actually reaches those results.

Another neural network criticism is the unique models, best models, good models argument (StatSoft Inc 2008). Models that are generated may not necessarily be the best models that could be found, nor is it necessarily true that there is a single best model. After training, several models with similar performance quality may result. Each model can be regarded, in this case, as a unique solution. It is not unusual that even models with the same number of hidden units, activation function, etc., may actually have different performance when validated. This is due to the nature of neural networks as highly non-linear models capable of producing multiple solutions for the same problem. Even though this may not necessarily be a problem, it always leaves the modeller wondering if the model could get any better, even if very good generalisation is being achieved with the present model.

3.2.6. Why Neural Networks for this research?

Neural networks was chosen for this modelling based on the following key considerations:

- *high number and type of variables*

  There were at least 20 variables to model during the prototyping stage of this research. Most of these variables had at least 3 options to choose from. Site access, for example, had options of unrestricted, restricted and highly restricted whiles tendering method had options of selective competitive, open competitive, negotiated and serial tendering. Furthermore, there was a wide mix of types of variables and their scales of measurement. In particular, most of the variables were categorical, instead of the more usual continuous type. Anderson (1995) suggests that neural networks cope better with categorical variables, the curse of dimensionality and multicollinearity, statistical conditions 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 (Hair *et al.* 1998).

- *non-linear relationships*

  Neural networks is a sophisticated modelling technique that is capable of modelling extremely complex functions. Linear modelling, typically regression, has been the commonly used technique in most modelling domains since linear models have well-known optimization strategies. These linear functions usually take the form of y= m*x* + c and illustrated in Figure 7. However, where the linear approximation is not valid (which is frequently the case) the models suffer accordingly. Instead of trying to model all data to a "best-fit" line in regression equations, neural networks attempts to model the non-linear relationship between the variables, in a manner shown in the illustration in Figure 8 . As will be demonstrated in the data exploration stage in the next chapter, non-linear relationships were identified between most of the variables and final cost of the project.

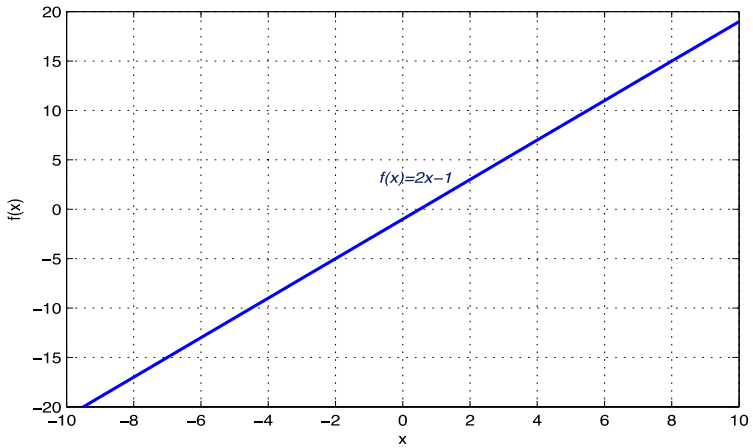


Figure 7: Linear Relationship

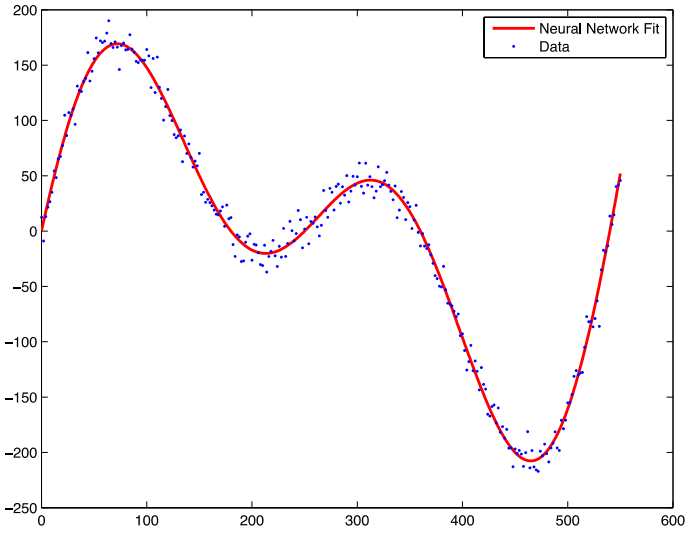


Figure 8: Non-linear relationship modelling

- *relationships between variables are vaguely understood or difficult to describe by conventional approaches*

  It may be straightforward to guess the likely relationship between type of cost and risk level on a project for example. Normally, low risk begets lower cost. However, it is more difficult to estimate the final cost of a project should the size of the project, duration, type of contractor, ground condition, operating region or purpose of project be added to the list of variables. Combine this with the cost of risk and inflation and the possible variable relationships get murkier. Unlike regression for example, Elhag and Boussabaine (1998) observe that neural networks do not require any prerequisite establishment of rules about how variables combine or the relationships between them. They essentially seek underlying relationships between variables and are particularly suited for complex, hard-to-learn problems where no formal underlying theories or classical mathematical and traditional procedures exist (Adeli 2001).

3.3. CHAPTER SUMMARY

This chapter has provided details of the research approach adopted in this thesis. A rationale for this approach has been detailed, along with a framework to guide the experimental exercise in this research. An overview of artificial neural networks, with some applications in construction management research has also been provided. Some of the strengths and potential weakness of ANN has also been discussed.

In the next chapter, the experimental design framework described in this chapter will be applied to construction cost estimation using a database of about 1,600 water infrastructure projects completed in Scotland. Neural networks will be used as the main modelling technique, combining it with data bootstrapping, ensemble modelling and fuzzy set theory. All this is carried out as an attempt to extract information embedded in historical project data to build cost estimation models that can help circumvent the problem of lack of information at the early stages of a construction project for accurate estimation.

CHAPTER FOUR

PUTTING CONSTRUCTION DATA TO WORK

"We need to think harder and smarter.

What we really need is holistic analysis, not holistic data.

We need to make better use of what we have.

We need to dig deeper..."

~ Jennie Beck, Global Director, TNS Media

4.0 INTRODUCTION

"...We need to dig deeper" aptly sums up the purpose of this chapter. Quite simply, it is making data work for an organisation by exploring available data in search of consistent patterns, correlations or systematic relationships between variables in a process called data mining. The discovered information is then used to improve business performance.

As already identified in earlier chapters, early cost estimation is often hampered by the unavailability of reliable information for estimation. In the absence of the detailed design and project information, construction clients still require reliable estimates upon which they can base their feasibility studies, budget allocation, tender evaluation, eventual contract award and project control.

This chapter is structured to first introduce the concept of cost modelling and the modelling philosophy adopted in this research. A description of the data to be used for the modelling, as well as the the data processing techniques used for training and validating the developed cost models are then presented. The aim in this chapter is to extract information embedded in historical project data to build cost estimation models that can help circumvent the problem of lack of information for reliable early cost estimation.

4.1. COST MODELLING

Seeley (1999) describes cost modelling as a procedure developed to reflect, by means of derived processes, a procedure developed to reflect, by means of derived processes, adequately acceptable output for an established series of input adequately acceptable output for an established series of input data". Similarly, Ferry *et al.* (1997) define cost modelling as a symbolic representation a system, expressing the content of that system in terms of the factors which influence its costs system in terms of the factors which influence its costs. For the

purposes of analysis and forecasting, Raftery (1998) adds that the symbolic representation must be "manipulable".

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

Skitmore's (1986) early research broadly classified cost models into high and low level models. Low level models either based their estimates on the elements of a completed building (i.e. functional unit/storey enclosure methods) or a *work-in-place* model such as Bills of Quantities. The high level models include statistical methods such multiple as regression functions that are based on a number of priceable variables that contribute to final cost of the built asset.

Ashworth (1999) classified the development of cost models over that time:

- 1960-1970: traditional or single point deterministic models such superficial (costs per m^2), elemental analysis and approximate quantities.
- 1970-1980: mathematical models such as expert judgement, parametric modelling and process models.
- 1980-1990: value related models such as life-cycle models, Monte-Carlo simulation and risk analysis.

- 1990-2000: integrated knowledge-based models such as in-house expert systems.

Fortune and Cox (2005) however, classified cost models into four, namely:

1. Traditional Methods

These models are based on comparable cost of projects in the past based on either similarities or differences in their function or geometric and spatial arrangements. It also includes the traditional bottom-up estimating methods which are usually based on some of standard form of measurement. Examples include:

a. Functional unit method
b. Superficial method
c. Bill of quantities.

These methods are perhaps the most popular in the construction industry, probably because they are usually straight-forward, well established and familiarity. For a long period, bills of quantities based on the Standard Rules of Measurement, published by the Royal Institute of Chartered Surveyors (RICS) remained the most popular method of estimation. First published in the 1922, it is now in its Seventh Edition (SMM7) and provides detailed information, classification tables, and most importantly, a uniform basis of quantifying building works in an attempt to facilitate consistency and best practice in the construction industry. Its civil engineering counterpart is the Civil Engineering Standard Method of Measurement (CESMM), published by the Institution of Civil Engineers.

The SMM7 was recently superseded by the New Rules of Measurement (NRM2) in January 2013, which perhaps better reflects how the construction industry works now. It allows for better consideration for costs centers such as cost of acquiring land, planning costs, contingency, cost of finance, fees, marketing costs and risk.

The Building Cost Information Service (BCIS) has recently proposed the of a functional approach to costing of civil engineering works called

Standard Form of Civil Engineering Cost Analysis (BCIS 2012). The aim of this approach is to support cost estimation based on equivalent functions in other projects, such that, information from existing projects can inform the budgeting and benchmarking of current and future projects. It is the analysis of the cost of a project in terms of its functions. The Standard is organised into Elements of a project , defined as a major physical part of an entity that fulfils a specific function irrespective of its design, specification or construction. Some specific elements include pavements, retaining structures, pipelines and ducts.

2. Mathematical Models

Examples of the mathematical traditional methods include:

a. Statistical modelling (e.g. Regression analysis),
b. Life-cycle costing models
c. Parametric models

Regression models are perhaps the most popular in the literature (Williams 2003, Lowe *et al.* 2006). This modelling technique establishes a general relationship between dependant and independent variables. A linear model, line of best fit, is produced which either expresses the exact functional relationship between predictor and dependant variables or an acceptable approximation of a more complex relationship.

Lowe *et al*. (2006) developed linear regression models to predict the construction cost of buildings, based on 286 data cases collected in the United Kingdom. The performed both forward and backward stepwise analyses, producing a total of six models. Forty-one initial independent variables were used with five significant influencing variables in each of the 6 models. These factors are gross internal floor area (GIFA), function, duration, mechanical installations and piling. They reported the best model performance of 19.3% Mean Absolute Percentage Error (MAPE) with $R2$ of 0.661.

3. Knowledge-based models

Knowledge-based models include

a. Expert systems
b. Case-based reasoning
c. Price

Expert systems use domain specific knowledge and heuristics to simulate the reasoning of an expert in that field in order to perform an intelligent task (Adeli 2003). Skitmore (1986) is probably one of the earliest to introduce the use of expect systems in construction management research. The paper examines who construction experts produce their cost forecasts for new projects and further developed an expertise scale to identify common methodologies and practices amongst experts. The research interestingly suggested that the reliability of the estimators 'first guess' of likely cost based on project size and type could be the most distinguishing quality between different levels of experts.

In the case-based reasoning (CBR) approach, a knowledge-base containing past cases is created from which a case similar to a proposed project is retrieved and revised in order to estimate the cost of that proposed project [see Figure 9].

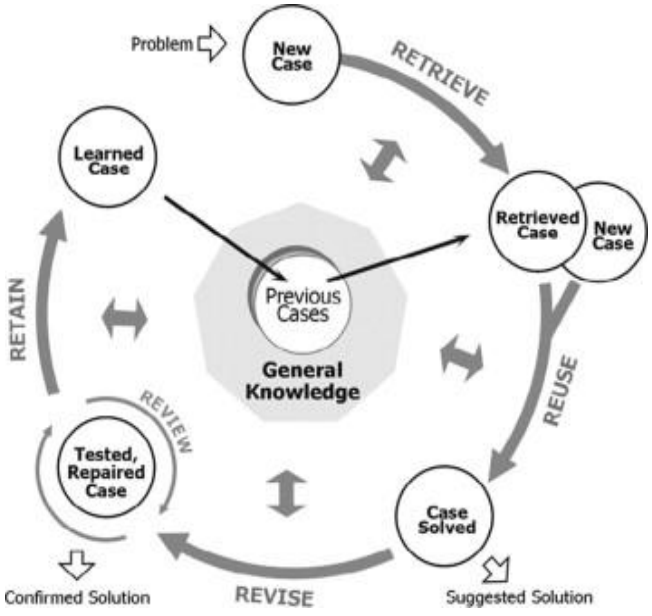


Figure 9: Case-based reasoning approach (Aamodt and Plaza (1994)

Marzouk and Ahmed (2011) developed CBR models to estimate the cost of pump station works, based on 14 cost-influencing factors from 44 completed projects. Some of the factors used in the their model include the size and length of pipes, the distance between pump station and water source, number of pumps, capacity of the pumping station and population size to be served by the pump station.

4. New-wave models:

These models are based on developments in the field of artificial intelligence and include neural networks, genetic algorithm and fuzzy logic. They also include new modelling paradigms that encompass sustainability themes and Computer-Aided Design (CAD). Fuzzy logic models attempt to address the imprecision and uncertainty in decision making and the boundaries of different classes or rankings used in costing. First developed by Zadeh (1965), variables in fuzzy logic have set of values, which are characterised by linguistic expression, such as very high £/m^2, average £/m^2, low £/m^2 etc. These linguistic expressions are represented numerically by fuzzy sets, more appropriately termed membership functions.

Genetic algorithms (GA) are general algorithms based on an evolutionary mechanism, where natural evolution and survival-of-the-fittest are simulated to perform a random search for the optimal solution to a problem. Kim *et al* (2004) combined GAs optimisation qualities with the learning abilities of neural networks to develop cost estimation models. GA was adopted in their research to determine the optimal parameters for the back-propagation neural network architectures based on 530 residential building projects completed in Korea between 1997 and 2000. They report improved efficiency and accuracy in the final models when GA was incorporated into the neural network learning process.

In a similar research, Kim *et al* (2004) also compared the performance of multiple regression, neural networks and case-based reasoning models for construction cost estimating based on the results of 530 completed projects. They found that although neural networks generally out-performed the other modelling approaches, the case-based reasoning models were preferable for long-term use, quantity of data requirement as well as time versus accuracy trade-of.

Artificial neural network has been adopted in this thesis for the cost model development. As seen in Figure 10, this is a new wave type of cost model. Further details of artificial neural networks, along with the rationale for using this approach is provided in Section 3.2.]

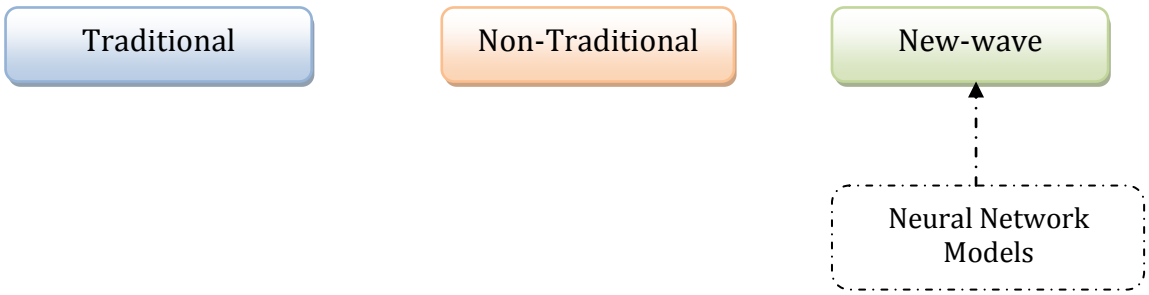


Figure 10: Classification of Cost Models

4.2. MODELLING PHILOSOPHY

Models are reductions of reality. They represent the critical aspects of a complex system in simple forms using variables within those systems. They might take mathematical forms like statistical models and differential equations or be in the form of neural network models and regression trees.

The use of models for construction cost estimating is appealing for a number of reasons including potential savings in time, resources and effort. For example, consider using only four different parameters that might influence the final cost of a project, each with three alternative values. Varying one parameter at a time in a what-if analysis could generate up to 81 different project solutions or alternatives (i.e. 3^4). This can be done rather rapidly using a computer-based model but will

undoubted be a laborious task using traditional cost estimation. The time, effort and resource level required for this task would mostly be unjustifiable at the planning stages of a project, perhaps a strong suggestion that detailed cost estimates at strategic level are often far from optimal solutions because of time and resource constrains.

The modelling in this research is carried out as structured in Figure 11. The concept and approach of using artificial neural networks for construction cost modelling is initially piloted using a small dataset collected from a civil engineering company in the UK (Dataset 1). Having achieved successful results with this dataset, the lessons, experience and approach used were then expanded to a larger database of 1,600 project cases with a major client organisation in Scotland (Dataset 2).

During the actual modelling using each of the datasets, three sub-samples of each dataset was created for training, testing and validation of the developed models. Further details of the data splits will be provided later in the thesis.

Three different modelling strategies, vis, standard neural networks, data boostrapping and ensemble modelling were experimented. Each new strategy was designed in an attempt to deal with the possible weaknesses of the previous. Judgement was then made on each strategy's predictive performance, the complexity of the model, the effort and data requirements for each strategy and the ease of deployment of that strategy in practice. The overall aim of the modelling is to help "dig deeper, and smarter" into existing construction data to produce cost models that could aid the estimation process in early stages of the project.

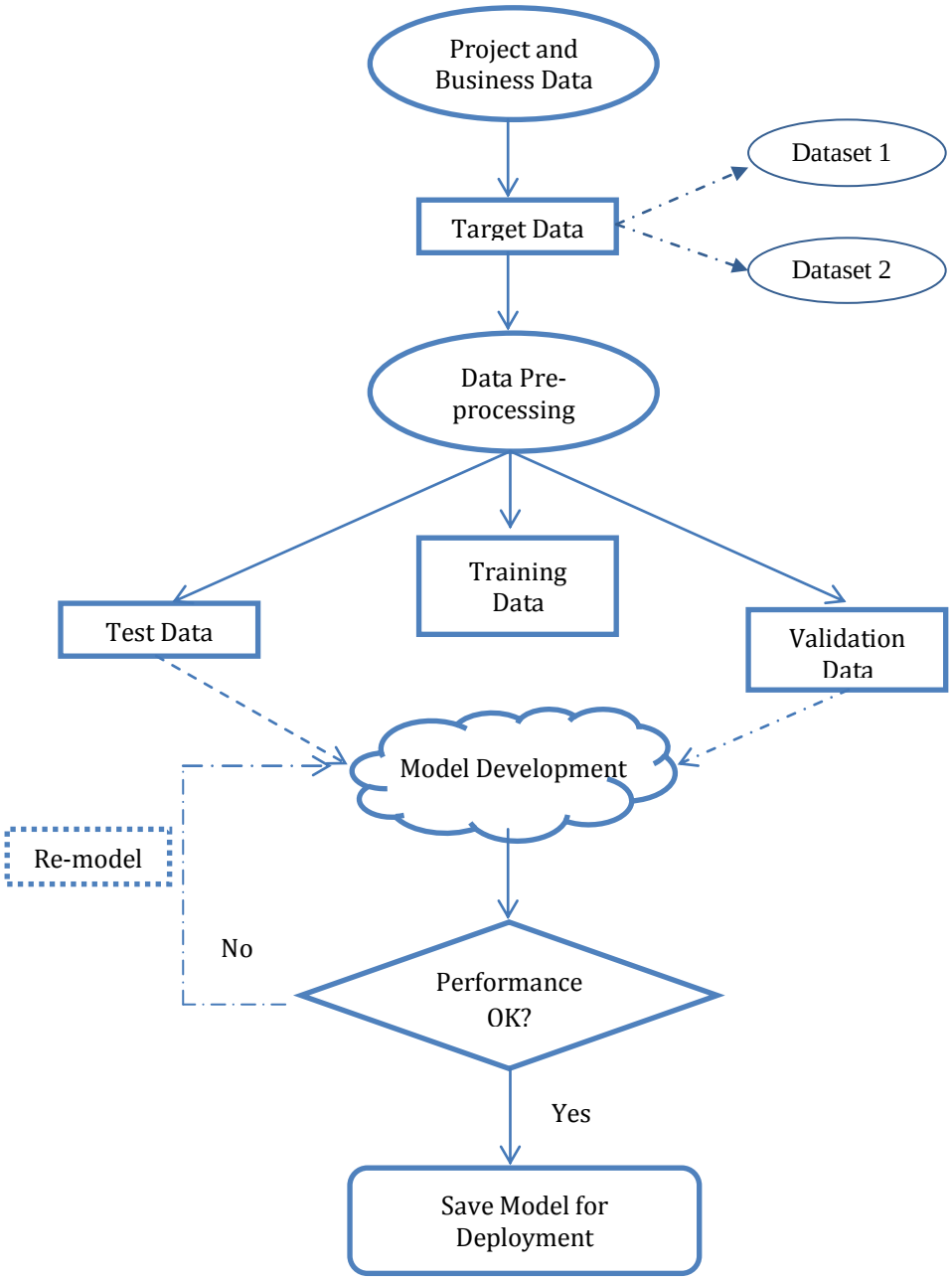


Figure 11: Model training procedure

4.3. THE DATA

The data used in this research was collected from two sources:

Dataset 1: 98 project cases, with a total value of about £99 million, completed in Scotland between 2007 and 2011 by Morrison Construction, a UK Civil Engineering contractor;

Dataset 2: Approximately 1,600 projects from a major public utility company in Scotland (total value over £800 million).

4.3.1. Dataset 1

Dataset 1 was used to develop trial models to experiment with using neural networks for cost modelling. This allowed for several trials with different neural network architectures, algorithms, transfer functions and data transformation.

The data collection process with Morrison involved an initial shadowing of the tendering and estimation procedure as a quasi member of the tendering team. This provided the opportunity to gain a first-hand understanding of how the data to be used for the modelling was generated and what different variables meant. It was followed by completing the datasheet in Appendix B for each project with details on the estimated and final costs, compensation events and duration, as well as qualitative information such as tendering method, location, type of project, fluctuation measure and type of deadline.

The nature of these projects were rather varied, ranging from construction of water mains, water treatment plants, combined sewer overflows, installation of manholes or water pumps and upgrades and repairs to sewers. All the projects were target cost contracts with values between £1,000-£14 million and durations from 1-22 months. Full details of the analysis and results from this dataset has been reported in the paper "*Neural networks for modelling the final target cost of water projects*" (Ahiaga-Dagbui and Smith 2012) and attached as Appendix A5. An overview is provided below to give an indication of some of the analysis carried out and the lessons to carry forward to the main analysis with dataset two.

First, after normalising the cost values across different years using cost indices for infrastructure resources from the Building Cost Information Services (http://www.rics.org/uk/knowledge/bcis), separate cost models were developed for the untransformed normalised target cost and the common log of target costs. However, the results from using the untransformed normalised target cost were mostly inconsistent and unreliable. This was possibly due to the wide range in the cost values

that the models had to learn, from £1,000 to £14million. It was found that neural networks generally require that numerical inputs be transformed into a small range of variability before training the models. If one input has a range of 0 to 1, while another has a range of 1,000 to 14,000,000, the model will expend most of its effort learning the second input to the possible exclusion of the first. Log transformations were therefore used in further iterations for the model development as this reduces the range of the cost and duration input.

The common log models showed significant improvement in the error values but slightly deteriorated in correlation. It was hypothesised at this stage that even though the log transformations reduced the cost inputs to a smaller range, making them more sensitive to the training algorithms of neural networks, they possibly imposed a logarithmic scale on the data, which might not actually be true of the real data. Standard values (zScores), explained later, were to be used to overcome this problem in dataset 2.

Furthermore, 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 2 phases: it first performs a probability distribution of the inputs before searching for relationships between the input and output space in the next stage (StatSoft Inc. 2011b). MLPs on the other hand model using just the second stage of the RBF. They usually thus complete their learning a lot quicker and tend to be used for most regression type problems. RBFs are usually applied to classification problems. The MLP models were superior to the RBF networks for the analysis in dataset 1 and so the rest of the modelling was carried out using just MLPs.

Additionally, the effect of using weight decay regularisation in the *hidden*[1] and *output layers* of the neural networks was also investigated. This was an attempt to encourage the network to develop smaller weights to reduce the problem of *over-fitting*[2], thereby potentially improving generalization performance of the network. Weight decay modifies the network's error function to penalise large weights. The result thereof is an error function that compromises between performance and weight size (StatSoft Inc. 2011a). The models showed improvement in both the error and correlation coefficient for the validation samples. This was to be carried out on the main modelling as well.

The final stage of the modelling with dataset 1 involved a test for parsimony, termed here 'survival of the fittest' test. Ockham's Razor principle, attributed to 14th century logician, William of Ockham, posits that one should not increase, beyond what is necessary, the number of entities required to explain anything and that all things being equal, preference should always be given to the simplest hypothesis (Chase *et al.* 1996). This principle of simplicity is used to prune down the number of variables required in the model to predict the final cost, thus reducing possible inconsistencies, ambiguities and potential redundancies in the model.

To implement this strategy, a relative importance list of the variables used in the modelling was developed to indicate each factor's contribution to predicting final cost of the project (see Table 3). Then, model performance was measured, while deleting one variable at a time, starting from the least important until the model showed no

[1] Explained later in the actual model development section of this chapter.

[2] Over-fitted models tend to just memorise the data without actually learning the underlying patterns and correlations. They perform very well during training but fail to generalise satisfactory when new data is used to validate them.

improvement or begun to decay. The final model's prediction did significantly improve after the bottom three factors were removed from the input space. The similar analysis was later carried out on the larger dataset as well.

Table 3: Relative Importance of Variables in Dataset 1

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

Figure 12 shows the performance of the final model from dataset 1, validated over 9 project cases. The error range of the model was between -2% (underestimation) to 7.9% (overestimation) with an average error of -1.8% underestimation. This compares favourably with the -10% to +15% estimation error commonly found and accepted in practice (Potts 2008). This result demonstrates the potential of using neural network and data mining to increase the reliability of early cost estimates using historical cost data.

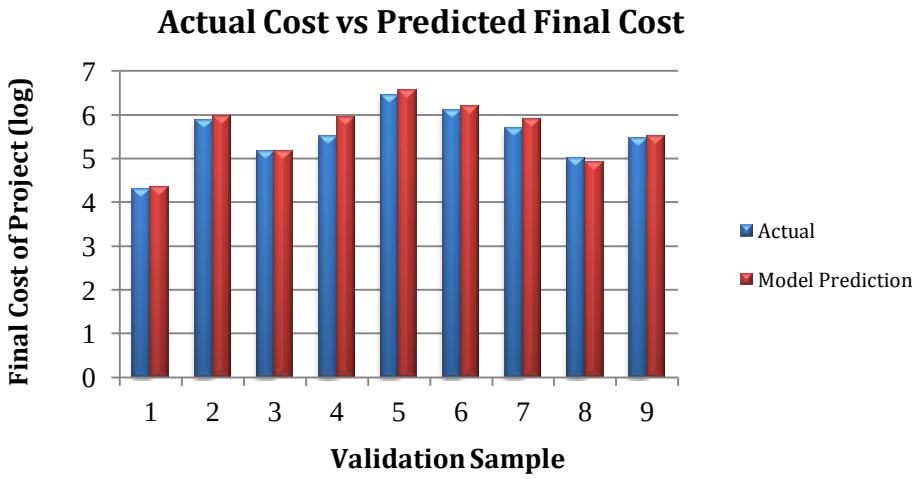


Figure 12: Performance of the final model from Dataset 1

Although not extended to dataset 2 because of time constraints, a neuro-fuzzy modelling approach was experimented with dataset 1 as well. This approach allows the learning and generalisation capabilities of neural networks to be combined with the capacity for tolerance and imprecise knowledge representation of fuzzy set theory. It has the possibility of increasing the reliability and flexibility of the models. A three-point fuzzy lower, upper and mean estimate of likely final cost was generated to provide a tolerance range for final cost rather than the traditional single point estimate. The performance of the final models ranged from -3.3% underestimation to +1.6 % overestimation. The conference papers "*A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects*" and "*Mapping Relational Efficiency in Neuro-Fuzzy Hybrid Cost Models*" [Appendix A6 and A7] were published based on the results of the neuro-fuzzy modelling strategy.

4.3.2. Dataset 2

The lessons learnt from developing the trial models using Dataset 1 were then extended to the larger database collected from the utility company. This company oversees the construction, operation and maintenance of water infrastructure in Scotland. Their asset base includes over 47,000km of water pipes, 50,000km of sewer pipes, 1,837 waste water treatment works and 297 water treatment works plus pumping stations, sludge treatment centres, and reservoirs.

The company has an in-house costing team that is directly in-charge of producing estimates and letting projects out for tender. It develops a series of Capital Expenditure (CAPEX) estimates at various stages before and after contract award. CAPEX1 estimate is usually just a rough estimate based on previous similar works. It is usually not based on any design drawings. CAPEX3, however, is based on about 50-60% of scope design and is used as benchmark for evaluating tenders after which detailed design is carried out by the selected contractor in, most commonly, a variant of a design-and-build contract framework. To ensure that they are getting value-for-money and awarding contracts at the most economically advantageous prices, as well as estimate their likely final financial commitment, Dataset 2 was supplied by the collaborating firm in this research for the development of cost models.

4.3.2.1. Project Cost and Duration

As shown in Table 4, Dataset 2 contained nearly 1,600 project cases with a total value of over £800 million. These projects were completed fairly recently between 2009 and 2012. The project costs range from a mere £1,000 on typical replacement projects to £30 million on large water treatment plants.

About 99% of the total number of project cases cost less than £25 million with only 3 projects costing more than this figure. On a cursory

level, this might suggest that the models developed from the database will be more sensitive to projects costing up to about £25 million.

Furthermore, 80% of the projects were completed within 3 years, with only 64 projects completed after 5 years. The average duration of the projects was about 24 months.

Table 4: Frequency Table of Final Cost of Projects

Project Final Cost (in millions, £)	*Count*	*Cumulative Count*	*Cumulative %*
0<x<=5m	1,535	1,535	97.77
5m<x<=10m	24	1,559	99.30
10 <x<=15m	7	1,566	99.75
15m <x<=20m	1	1,567	99.80
20m <x<=25m	1	1,568	99.87
25m <x<=30	2	1,570	100.00

Table 5: Frequency Table of Duration of Projects

Duration (Months)	*Count*	*Cumulative Count*	*Percent %*
0<x<=20	557	557	35.48
20<x<=40	700	1,257	80.06
40<x<=60	249	1,506	95.92
60<x<=80	46	1,552	98.85
80<x<=100	18	1,570	100.00

A plot of project duration and final outturn cost in Figure 13 does not show a linear relationship between the two factors. This is perhaps an indication that linear modelling techniques like multiple linear regression might not be appropriate modelling this data.

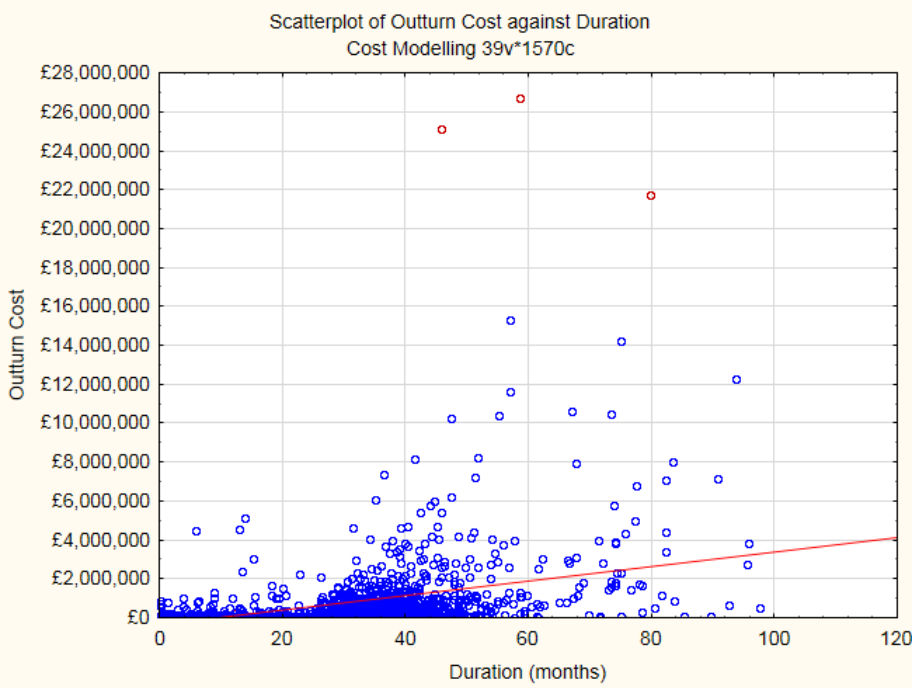


Figure 13: Scatter plot of final cost versus project duration

Unlike Dataset 1 that had details up to project level of each individual case, Dataset 2 contained only strategic and management level details. These variables are shown in Table 6.

Table 6: Input factors for modelling exercise

	Input Factor	*Input Options*			
1	Primary Purpose	Wastewater	Water	General	-
2	Project Scope	Upgrade	Replace	Refurbishment	New-build
3	Delivery Partner	X	Y	Z	-
4	Operating Region	North	South	East	West
5	Project Duration (months)				
6	Estimated Cost (CAPEX 3), £				

4.3.2.2. Purpose of the project

The projects can be categorised into wastewater, water and general purposes. Wastewater projects refer to projects that concern the construction or maintenance of pipework systems and other physical infrastructure required for the transport or treatment of waste effluent

from homes and industries through combined or sanitary sewer. These represent 44% of the total number of projects in Dataset 2 and might take the form of combined sewer overflows, wastewater treatment plants or installation of pipes. Water projects on the other hand, are concerned with infrastructure works relating to collection, treatment, storage and distribution of drinkable water. These might be pumping stations, storage tanks or pipes which represent 55% of the project cases used in this research. General projects are ancillary works like upgrades for health and safety or environmental compliance or minor repair works that would not merit the classification of major water or wastewater projects.

Figure 14 shows the sensitivity of final outturn cost to the different project purposes. Wastewater projects are generally the more expensive projects with general projects averaging about £350,000.

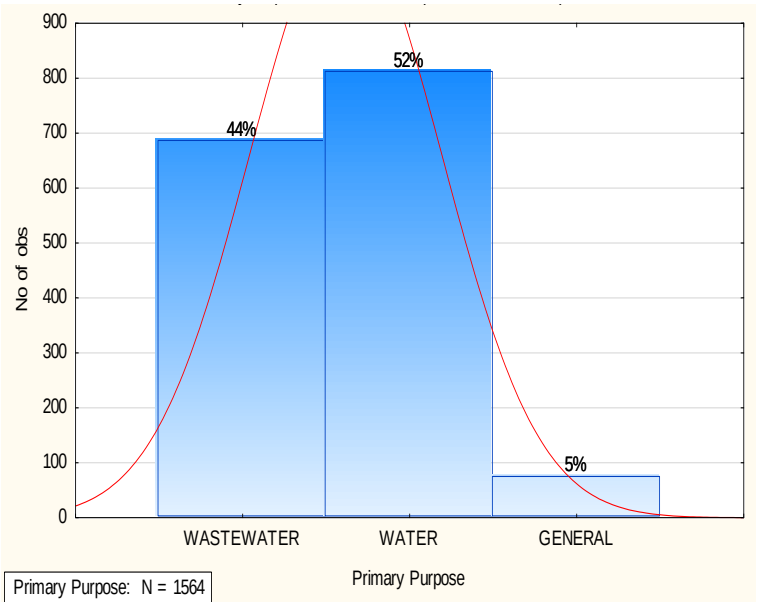


Figure 14: Histogram showing distribution of the purpose of projects

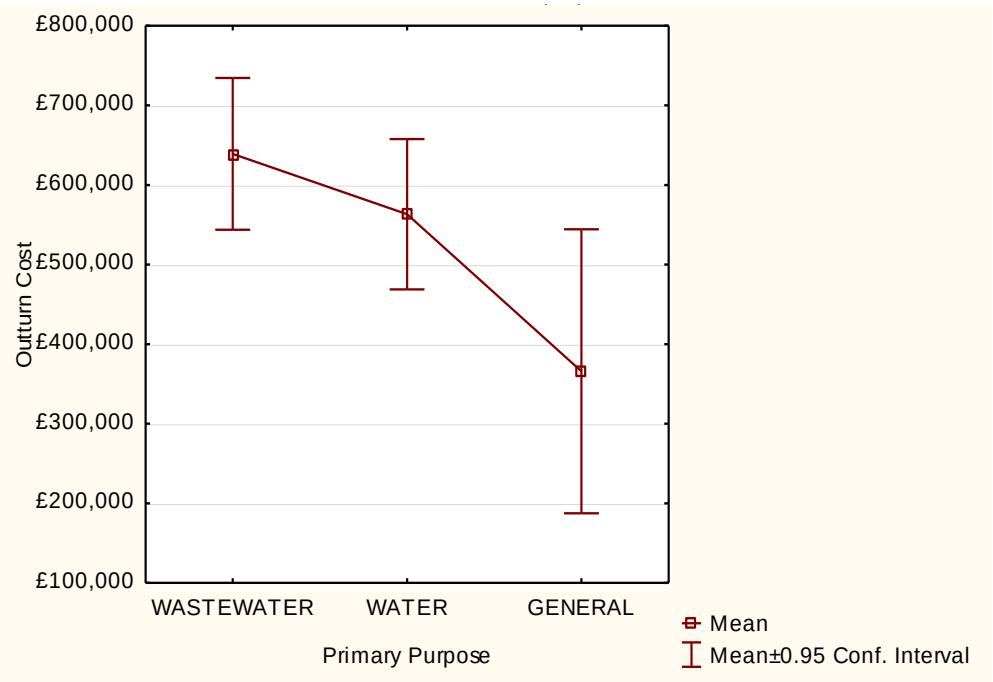


Figure 15: Mean Plot of Cost Variation with Primary Purpose

4.3.2.3. Delivery Partner

The construction division of the collaborating firm is divided into three sections, referred to here as X, Y, Z for confidentiality. These delivery partners directly oversee the procurement and administration of the contract with the eventual construction company that will undertake the project. The mean plot in Figure 17 shows the change in average cost of projects depending on the different delivery partners. It can be observed that even though Delivery Partner Y was in charge of most of the projects, Z mostly carried out the more expensive projects.

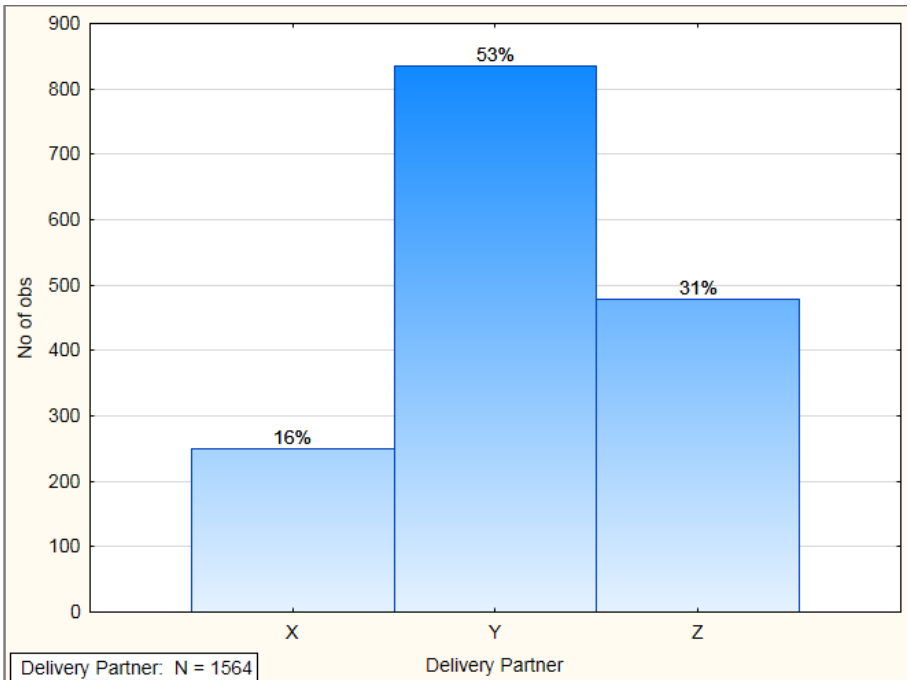


Figure 16: Histogram showing distribution of delivery partners

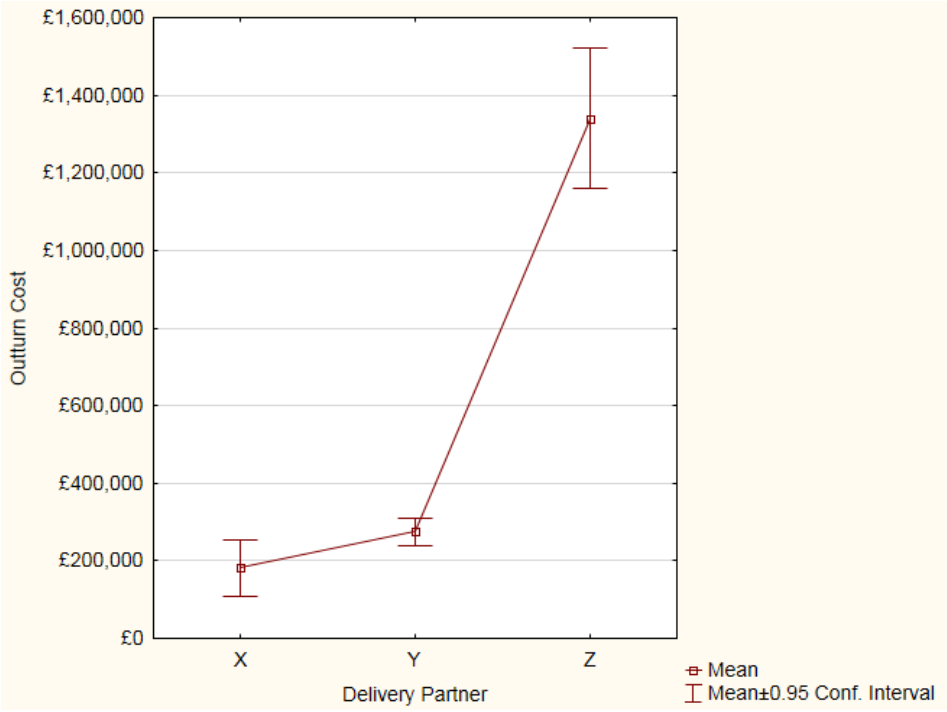


Figure 17: Mean Plot of Cost Variation with Delivery Partner

4.3.2.4. Scope of Project

The projects are further classified according to their scopes, as seen from Figure 18. These are upgrade, replacement or refurbishment projects. As Scotland already has an extensive existing water and

wastewater system, there are no 'new built' projects in the database used. Most of the projects are instead appropriately classified as upgrades to this existing network as seen in Figure 18. Replacement projects are usually simple component replacement at, maybe a pumping station or treatment works, whereas refurbishments are usually more extensive, consisting of a number of different job centres. As shown in Figure 19, the upgrade type of works were on average more expensive than the other classes under project scope category.

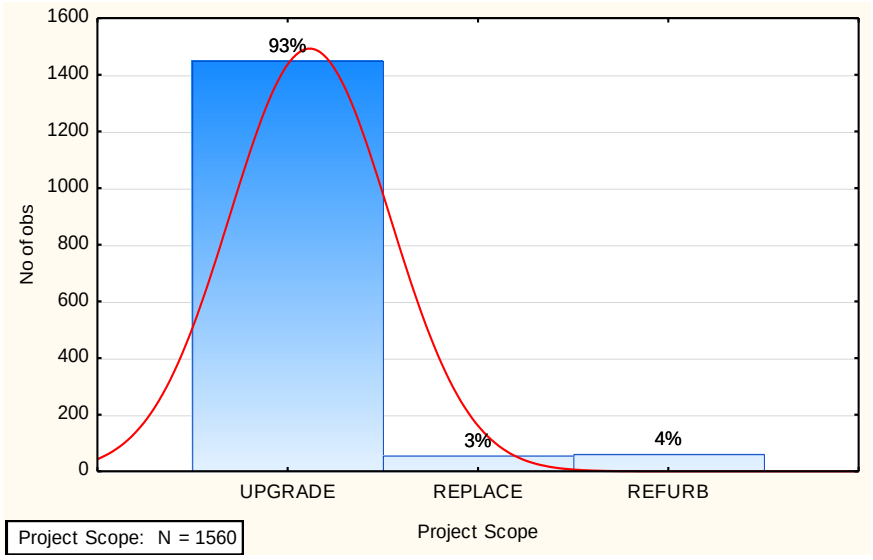


Figure 18: Histogram showing distribution of scope of project

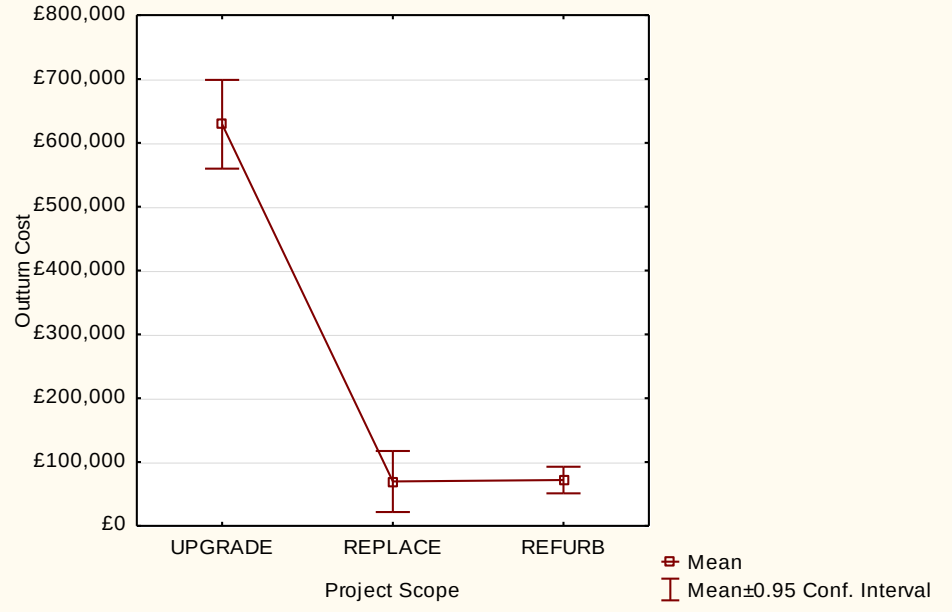


Figure 19: Mean Plot of Cost Variation with Project Scope

4.3.2.5. Operating Region

Previous research suggests that location of the project tends to affect its eventual cost. This was evident during the trial model development in early stages of this research reported in Ahiaga-Dagbui and Smith (2012). The generic location categorisation of the projects in this database as West (W), East (E), South (S) and North (N) of Scotland allowed for the testing of the location hypothesis. In Figure 20, the highest number of projects (i.e. 36%) were completed in the South of Scotland around the Edinburgh region, although the projects in the Eastern parts of Scotland (Aberdeen and Aberdeenshire) appeared to be more expensive in Figure 21.

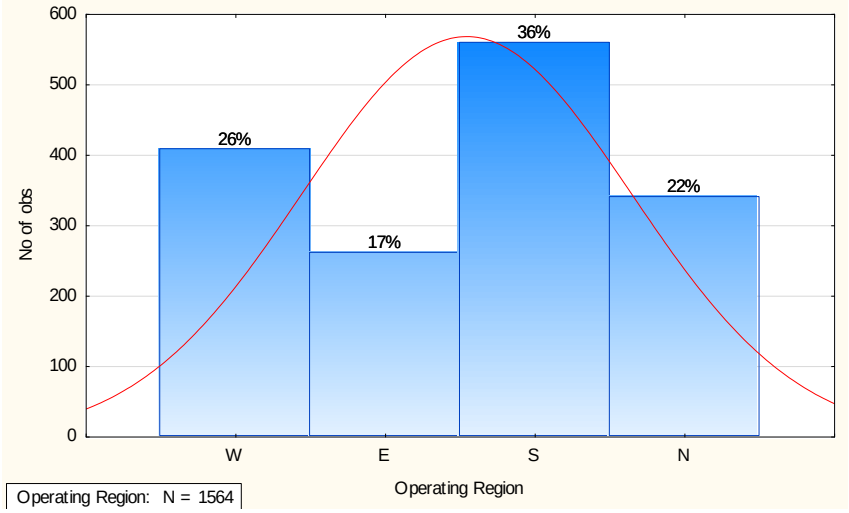


Figure 20: Histogram showing location of project in Scotland

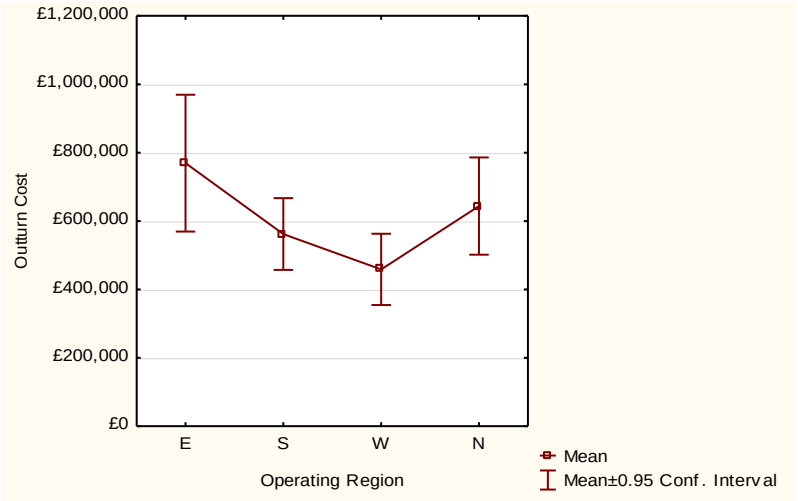


Figure 21: Mean Plot of Cost Variation with Operating Region of the Project

4.3.3. Data pre-processing

Real data is 'dirty', inconsistent, and incomplete. They often, according to Pyle (1999), contain errors, outliers, wrong measurements or aggregate data. Pyle further writes that "correct data preparation prepares both the [data] miner and the data. Preparing the data means the model is built right. Preparing the miner means the right model is built." Data pre-processing therefore allows the data to be 'cleaned', structured and presented to the model in the most suitable way in order to develop reliable models as well as offer the modeller the chance to get to understand the data thoroughly. The importance of data pre-processing is thus very crucial to the success and reliability of the models generated. Pre-processing might involve simple steps such as removing duplicate entries and missing data treatment to more advanced techniques like clustering, data transformation and de-noising (treatment of unexplainable randomness and variability in the data). The data used in this research was pre-processed as follows:

4.3.3.1. Data Integration

Data integration usually involves the merger of multiple databases and the removal of conflicting information from these different sources. The data used for this research was extracted from two different internal databases of the partnering organisation. While both databases contained the same number of project cases, one recorded project information such as delivery partner, project location and scope while the other was used for the cost control of projects and thus contained the cost targets and CAPEX estimates. These two sources were merged-up while removing duplicated or aggregated information.

4.3.3.2. Data Cleaning

Data cleaning unusually involves the removal of duplicate cases, identifying or removal of extreme cases, treatment of missing or

incomplete data and resolving inconsistencies within the data. Instead of removing incomplete cases from the data modelling, missing fields were replaced with the mode or mean of the distribution, depending on whether the entry type was categorical or continuous in nature. Rare values can create biases in data analysis as they often might appear as more important than they really are. There were three extreme cases of projects costing more than £25 million as shown in Figure 13. These have the potential of resulting in inconsistent predictions if included in the model as 99% of the data cases cost less than £25 million. On the other hand, if they were excluded, chances are the model would not adequately capture all possible range of cases that would be encountered in practice. At this stage of the research, the extreme values were not removed. The model's sensitivity to the extreme values would be tested at the modelling stage before deciding whether to include them in the final model.

4.3.3.3. Data Transformation

Another important step in data pre-processing is to transform the data into a small specified range. Some modelling techniques like neural networks require that numerical inputs are normalized into a small range of variability before training the models. Using raw values or log transformations for the trial models did not prove very effective as already explained in the analysis using dataset 1. The log transformations imposed an s-shape curve on the cost values, which upon scrutiny, does not truly fit the data (see Figure 13). Numerical predictors were thus standardized using the linear transformation of *scores*. This is defined as

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

Where: *score* 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

Apart from maintaining the shape of the numerical predictors because of the linear transformation, scores also lend themselves to easy interpretation as they measure how much a score deviates from the mean value of the distribution (Hair *et al.* 1998). A zScore of 2.0 for the final cost of any project would mean that the project is twice more expensive than the average project in the database.

Furthermore, all 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.

4.3.3.4. Data Coding

Data coding refers to the nature of the data presented to the models. Each type of data requires a different representation. During the trial model development, categorical variables such as type of soil, type of project and contractor's need for project were coded using the one-of-N coding, resulting in 4 sub-variables for type of soil for example (Good, Moderate, Poor, Not Applicable). On hindsight, this was possibly not the most appropriate coding to use as the one-of-N approach suggested that the categories could be calibrated on a nominal scale, with a sort of implicit degree of equal step variation between factors.

It was thus decided to use binary coding (0, 1) for all categorical variables in Dataset 2. This allowed for the creation of "dummy" variables with the value zero, or one where the input corresponds to the correct category. If the variable was Soil Type with categories Good, Moderate, Poor or Not Applicable for example, then the data would be presented to the model as shown in Table 7. This coding allowed the model to infer importance on its own without the modeller imposing weightings or subjective ratings to the variables.

Table 7: Example of Binary Coding of Categorical Variables

Category	*Dummy Variables*			
Good	1	0	0	0
Moderate	0	1	0	0
Poor	0	0	1	0
Not Applicable	0	0	0	1

4.3.3.5. Data Partitioning

The actual data mining process in neural networks is usually done in three different steps: training, testing and validation. The performance of the neural net is measured by how well it generalises unseen data (i.e. data that was not used in training the model). To avoid model over-fitting, it is imperative that separate data samples are used for training, testing and validating the neural network models. The three samples are used as follows:

1. **Training Set:** to train the network to identify patterns, correlations, etc
2. **Testing Set:** to assess how well the model is learning while it is still under training
3. **Validation Set:** to verify the performance of the model to determine how well it predicts unseen data which has neither been used for training or testing during model development. In this research, two validation sets were created: one to be used for an automated verification of model performance after training and the other for a manual validation of the models. Further details of how these two different validation sets are used will be provided at the actual modelling stage.

Out of the 1,570 project cases, 100 were selected using stratified random sampling with cost as the strata variable to be used as the second stage manual validation set. 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, using simple random sampling. Further details on the datasets used for the modelling is found in Table 8.

Table 8: Data Partitioning Details

Dataset	*Percentage split*	*Number of cases Total Size (1570)*
Training	70	1029
Testing	15	c. 220
Validation 1	15	c. 220
Validation 2	-	100

4.4. DEVELOPING THE MODELS

Three major modelling strategies were used to develop a series of final cost models in the research in an attempt to extract useful information embedded in construction data to support the estimation process within two industry collaborators. These strategies are the standard neural network modelling, data bootstrapping and ensemble modelling.

4.4.1. Standard Neural Networks

These models were developed using only artificial neural networks in Statistica® 10 software [See Appendix C for details on how different softwares were evaluated before choosing Statistica 10]. The models were developed in a trial and error manner to identify optimum network parameters and network performance. Several networks were trained using the input factors project scope, delivery partner, operating region, project duration, estimated cost at CAPEX 3 (see Table 6). The model output was cost at final account (CAPEX 6).

One of the challenges of using neural networks is that there are no set rules on the nature of the network architecture or number of neurons or layers to use. Each problem must thus be tackled using a trial and error approach until optimum network performance is reached. The automatic network search function of Statistica® 10 was thus used to optimise the search for the best network parameters, i.e. type of network architecture, number of hidden layers, activation functions, number of nodes in the hidden layer, etc. after which customized networks were developed using the optimal parameters identified.

4.4.1.1. Type of Neural Network Architecture

Initially, two different network architectures, the Multilayer Perceptron (MLP) and the Radial Basis Function (RBF), were experimented. According to Santos *et al.* (2013), the RBF and MLP networks are usually applied to the same kind of problem domains, i.e.

approximation and pattern recognition. They mainly differ only in their internal calculation structures. RBFs are linear while MLPs are non-linear functions. Also, while RBF maps the relationship between input and target in a 2 phases, first performing a probability distribution of the inputs before the searching for relationships between the input and output space in the next stage, MLPs on the other hand go through just the second stage of the RBF (StatSoft Inc. 2011b).

At the prototyping stage using Dataset 1, it was found that even though the MLP models trained a lot slower than the RBFs, the MLP models were always superior to the RBF networks. With preference to accuracy rather than speed, Dataset 2 modelling was thus carried out using only MLPs.

4.4.1.2. Hidden Layers and Hidden Nodes

One of the often asked questions about neural network modelling is, "how large does the network have to be to be able to adequately perform the task at hand?" The size of the network here refers to the number of hidden layers and nodes. Answers to this in the literature can be reduced to, "it depends". It depends on the complexity of the problem being studied, the quantity and quality of data available and perhaps more importantly, the level of accuracy required for the models (Anderson 1995).

On first principles though, the number of hidden nodes must be kept as low as possible to avoid model over-fitting or memorising. During prototyping, 2,000 networks were trained, iterating between 1-100 hidden nodes in one hidden layer. The 5 best networks were retained and examined for performance improvement. Repeatedly, all the retained networks were found to have between 3-10 hidden nodes. This bound was thus used to custom build the models using Dataset 2.

4.4.1.3. Training Algorithm

There are quite a number of training algorithms that can be used in neural networks, with the most popular being the back-propagation (Fausett 1994). Training algorithms are mathematical procedures used to automatically adjust the network's weights and biases during training to minimise prediction error. Without going into much details about all of these training algorithms, a brief summary of the ones deployed within Statistica® 10, the software used in this research (StatSoft Inc. 2011a) are presented below:

1. *Gradient descent*: This is a first order optimization algorithm that moves incrementally to successively lower points in search space in order to locate a minimum.
2. *Broyden-Fletcher-Goldfarb-Shanno (BFGS):* This is a more powerful second order training algorithm with very fast convergence but requires high processing requirements. It is also called Quasi-Newton algorithm.
3. *Conjugate descent:* This is fast converging generic learning algorithm. The method iterates a series of line searches for global minimum in the error space. Succeeding search directions are selected to be conjugate.

4.4.1.4. Activation Functions

The behaviour of a neural network during training is further controlled by the activation function used. These are mathematical functions that determine the nature of the network weights transferred from one neuron to the other. According to Haykin (1994), activation functions 'squash' or limit the output of a neuron into a permissible range, usually between the closed units [0,1] or [-1,1]. Five different activation functions are iterated in this research and are detailed in Table 9 below.

Table 9: Activation functions used in this research

	Function	*Definition*	*Description*	*Range*
1	Identity	-	The activation of the neuron is passed on directly as the output.	(-∞, +∞)
2	Logistic Sigmoid	$\frac{1}{1-e^{-\alpha}}$	An S-shaped curve. Output varies continuously, but not linearly.	(0,1)
3	Hyperbolic Tangent (tanH)	$\frac{e^{\alpha}-e^{-\alpha}}{e^{\alpha}+e^{-\alpha}}$	A sigmoid curve similar to the logistic function. Often performs better than the logistic function because of its symmetry. Ideal for multilayer perceptrons, particularly the hidden layers.	(-1, +1)
4	Exponential (Exp)	$e^{-\alpha}$	The negative exponential function.	(0,+∞)
5	Sine	$\sin\alpha$	Useful if recognizing radially distributed data.	(0,1)

Source: StatsSoft Inc., 2011

4.4.1.5. Performance Measurement

Model performance was measured over the training, testing and validation datasets using the correlation coefficient between predicted and output values as well as the Mean Squares Errors (MSE). 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. *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.

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. Each model was repeatedly trained as long as testing error was on the descent. Figure 22 shows an illustration of a training regime with early stopping. In this case, the model training is halted when testing error begins to increase.

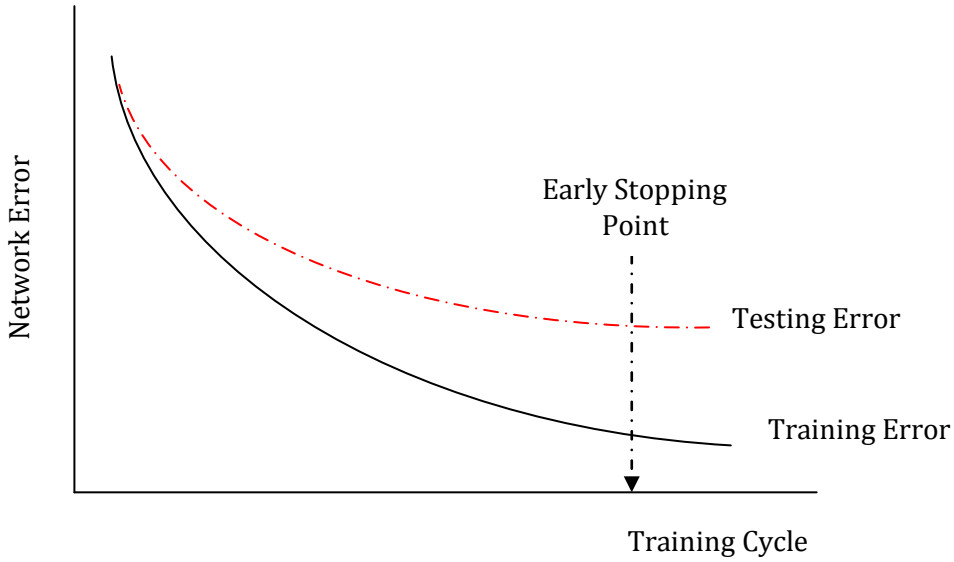


Figure 22: Neural network training with early stopping.

4.4.1.6. Training the standard models

After using the automatic network search in Statistica® 10 to experiment with possible number of hidden layers and nodes as already described, 2000 different cost models were custom trained using a hidden node range of 3-10 and a data split of 70:15:15% for training, testing and first stage validation respectively. All five activation functions in Table 9 were used with the three training algorithms already described (i.e. BFGS, gradient descent and conjugate descent). All 6 input factors in Table 6 were used initially with Final Project Cost as model output.

Early stopping was used to avoid model over-fitting and model performance was measured using the Mean Squared Error (MSE) over the training, testing and validation datasets. The best 10 models, out of

the 2,000, were retained for further validation with the 100 data cases sampled using the stratified sampling at the data pre-processing stage.

Figure 23 and Figure 24 show a sample of the plot of target versus model prediction of final cost for one of the retained models at this stage (MLP 16-7-1) while Figure 25 and Figure 26 plot the spread of residuals for the test and validation samples of the same model. These plots were generated for each model and inspected to give a quick visual indication of possible model performance.

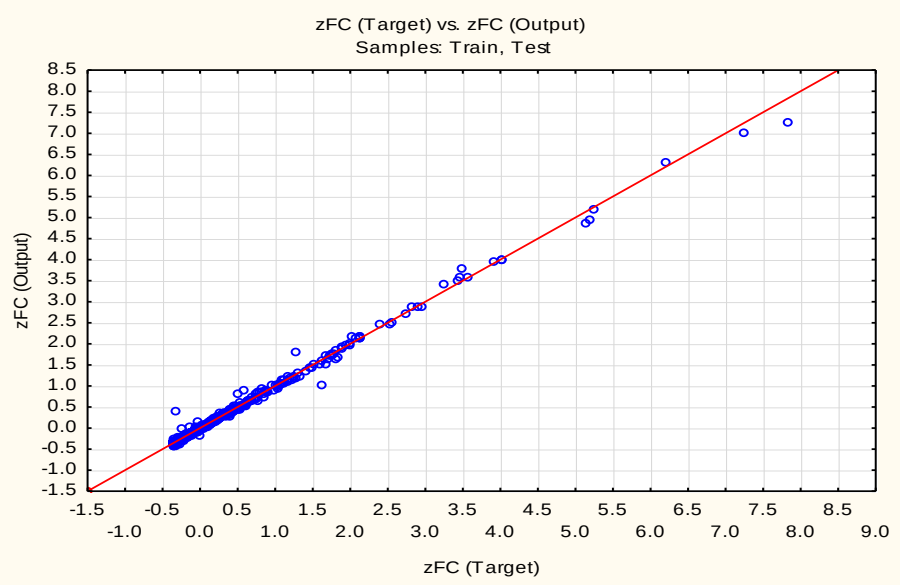


Figure 23: Plot of Target vs Output of MLP 16-7-1 (Training and Test Datasets)

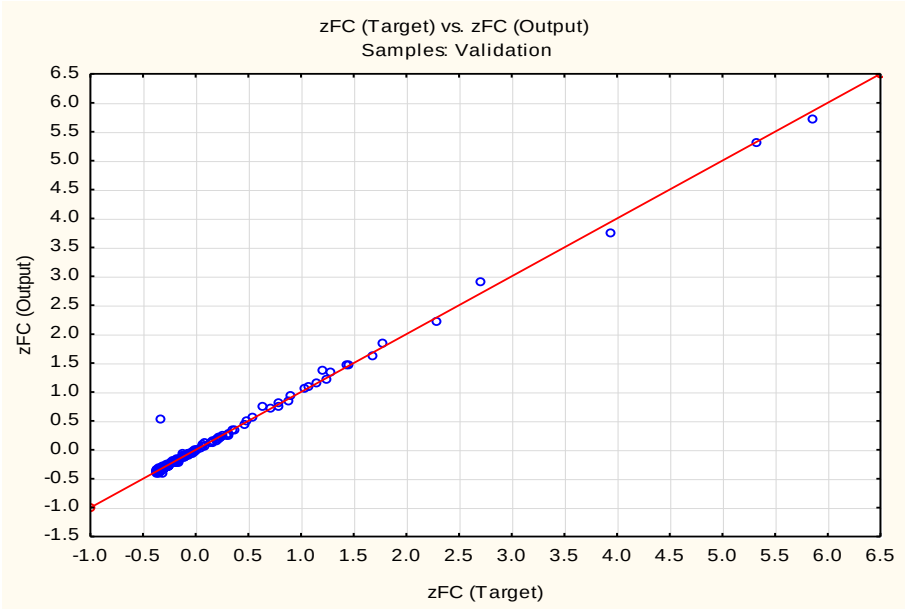


Figure 24: Plot of Target vs Output of MLP 16-7-1 (Validation Dataset)

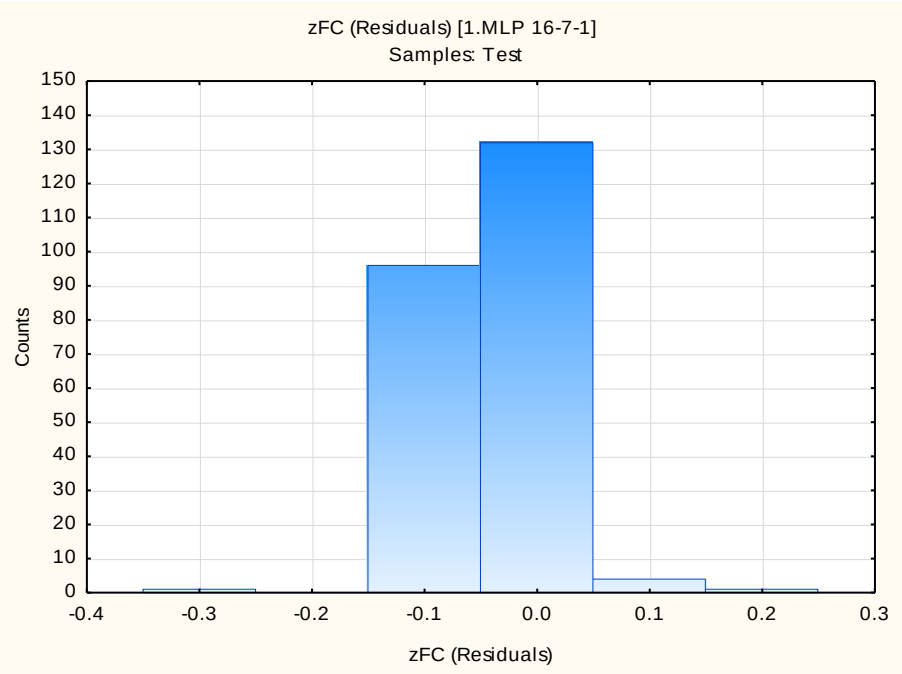


Figure 25: Plot of Residuals for MLP 16-7-1 (Test Dataset)

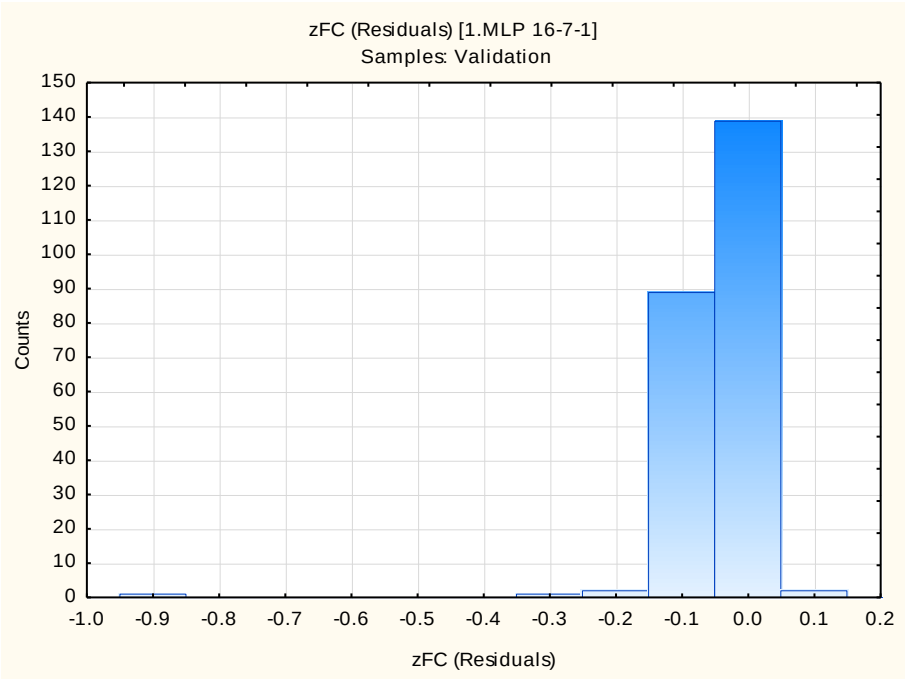


Figure 26: Plot of Residuals for MLP 16-7-1 (Validation Dataset)

A better and more thorough performance comparison was carried out using the correlation coefficients and MSE shown in Table 10. The table shows details of the 10 best retained models, in no particular order of superiority. Of most importance is the mean squared error over the validation dataset of these networks, as it shows the performance of each model when new data is presented to it. Curiously, all the top 10

models have virtually the same MSE of about 0.0021. This might be because of the rather exhaustive number of models trained (2000), increasing the likelihood of finding several models with similar predictive capabilities. These models will be further examined later. Before then, the sensitivity of the networks to the different input factors will be evaluated in the next section.

Table 10: Summary of results for the best 10 standard models

Model Architecture		Correlation Coefficient			Error (MSE)			Hidden activation	Output activation
		Train	Test	Validate	Train	Test	Validate		
1	MLP 16-7-1	0.9976	0.9986	0.9963	0.0011	0.0008	0.0021	Tanh	Sine
2	MLP 16-8-1	0.9982	0.9988	0.9963	0.0008	0.0006	0.0020	Exp	Sine
3	MLP 16-4-1	0.9976	0.9989	0.9964	0.0011	0.0005	0.0021	Logistic	Sine
4	MLP 16-5-1	0.9982	0.9987	0.9964	0.0009	0.0007	0.0022	Exp	Sine
5	MLP 16-6-1	0.9977	0.9978	0.9963	0.0011	0.0016	0.0021	Tanh	Identity
6	MLP 16-3-1	0.9958	0.9957	0.9964	0.002	0.003	0.0020	Exp	Logistic
7	MLP 16-3-1	0.9976	0.9989	0.9963	0.0011	0.0006	0.0021	Logistic	Sine
8	MLP 16-4-1	0.9976	0.9985	0.9964	0.0011	0.0007	0.0024	Tanh	Identity
9	MLP 16-6-1	0.9981	0.9987	0.9963	0.0009	0.0007	0.0021	Exp	Identity
10	MLP 16-7-1	0.997	0.9982	0.9964	0.0014	0.0014	0.0019	Exp	Tanh

4.4.1.7. Sensitivity analysis

A sensitivity analysis was performed on the input factors used in each of the 10 best retained models from the previous stage. This was an attempt to evaluate the contribution of each factor to the model's performance and also help to prune the number of variables used to an optimum. Table 11 shows the relative influence of the various inputs on the predictive performance of the models. The table was generated by comparing the predictive error of the 'full model' to that of a 'reduced model' when each factor is removed from the neural network in terms. The variables were then arranged in order of importance according to the change in performance noticed when they were removed.

The least significant input factor in this case is the Operating Region of the project. The in-house CAPEX 3 estimate of final cost was the most

important contributor to the model's ability to predict final cost as expected. This probably suggests the importance that ought to be given to the estimates generated by the collaborating firm before it invites tenders. The choice of the project's delivery partner seems to have a strong influence on the ultimate cost of the project as well.

Table 11: Sensitivity analysis

Sensitivity analysis Validation							Samples: Test,
		CAPEX 3 Estimate	Delivery Partner	Project Scope	Project Duration	Primary Purpose	Operating Region
1	MLP 16-7-1	9.24	3.16	1.09	1.58	1.04	0.06
2	MLP 16-8-1	10.25	6.89	2.98	2.97	1.13	0.05
3	MLP 16-4-1	10.29	1.28	2.48	1.68	1.05	0.26
4	MLP 16-5-1	7.56	2.89	5.73	5.94	2.01	0.11
5	MLP 16-6-1	6.95	4.56	0.99	1.34	1.07	0.01
6	MLP 16-3-1	10.65	2.34	5.91	3.2	4.9	0.15
7	MLP 16-3-1	9.56	6.37	4.31	2.42	1.52	0.19
8	MLP 16-4-1	11.82	4.53	5.01	3.15	2.61	0.05
9	MLP 16-6-1	8.11	6.21	1.52	4.13	1.01	0.91
10	MLP 16-7-1	10.04	2.01	0.95	1.69	0.96	0.03
	Average	**9.45**	**4.02**	**3.10**	**2.81**	**1.73**	**0.18**

As it is usually unhelpful to increase the number of modelling parameters beyond what is objectively necessary to explain the variance in a dataset as that has the likelihood of introducing potential redundancies, ambiguities and inconsistencies in the model. Thus, using the relative importance list in Table 11 from the sensitivity analysis, 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.

The model's performance significantly improved with the exclusion of project operating region, but reduced significantly when Primary Purpose was also excluded from the input space. Table 12 shows improved coefficients of correlation and reduced mean square errors when Operating Region was removed from the input space. Unlike the previous stage where the activation functions were random in both the

hidden and output layers all the new models, without Operating Region, had identity functions and mostly logistic functions in their output and hidden layers respectively, suggesting a better model consistency when the operating region was excluded.

Table 12: Summary of best models and performance (Without Operation Region)

Model		Correlation Coefficient			MSE			Hidden activation	Output activation
		Training	Test	Validation	Training	Test	Validation		
1	MLP 12-7-1	0.9951	0.9974	0.9993	0.0039	0.0052	0.0013	Logistic	Identity
2	MLP 12-3-1	0.9983	0.9984	0.9995	0.0013	0.0022	0.0012	Tanh	Identity
3	MLP 12-9-1	0.995	0.9975	0.9993	0.0039	0.0051	0.0013	Logistic	Identity
4	MLP 12-6-1	0.9985	0.998	0.9994	0.0012	0.0026	0.0018	Logistic	Identity
5	MLP 12-3-1	0.9951	0.9974	0.9993	0.0038	0.0053	0.0014	Identity	Identity
6	MLP 12-3-1	0.995	0.9976	0.9994	0.004	0.004	0.0012	Logistic	Identity
7	MLP 12-8-1	0.9985	0.9981	0.9995	0.0012	0.0025	0.0019	Logistic	Identity
8	MLP 12-3-1	0.9984	0.9982	0.9998	0.0013	0.0027	0.0005	Logistic	Identity
9	MLP 12-3-1	0.9983	0.9979	0.9996	0.0014	0.0028	0.0015	Logistic	Identity
10	MLP 12-3-1	0.9951	0.9974	0.9993	0.0038	0.0053	0.0014	Logistic	Identity
Average		**0.9967**	**0.9978**	**0.9994**	**0.0026**	**0.0038**	**0.0013**	-	-

Recall that 100 project cases were selected using stratified sampling during data partition of the pre-processing stage. The ten retained models in Table 10 were then further evaluated using these 100 validation cases to manually test their performance over the 100 data cases. Model performance was evaluated using the equation:

$$\text{Model Performance}(\%) = \frac{Actual\ Final\ Cost - Predicted\ Final\ Cost}{Actual\ Final\ Cost} * 100\% \qquad \text{Equation 3}$$

Table 13 shows a summary of the performance of the 10 standard models when validated with the 100 data cases. It can be seen that the model performance is within a range of 9.6% average underestimation to 8.35% average overestimation of actual final cost. Although the results indicate a similar range of performance across all the 10 models, Model 4 shows the smallest range of error between -7.69% and 7.09%. The architecture of this particular model is an MLP 12-6-1 (i.e. 12 input

nodes from 5 inputs, 6 hidden nodes and 1 output). It was trained with a Quasi-Newton algorithm using a logistic function in its hidden layer and identity function in the output layer. Figure 27 shows the error plot of Model 4, with the model achieving minimum error after 49 cycles of training. The training was set to continue until there is no more improvement in test error over 20 cycles to control model over-fitting.

Table 13: : Summary of performance of standard models with 100 validation cases

Model	*Standard Model Performance*	
	% Average Underestimate	% Average Overestimate
Model 1	-10.84%	8.35%
Model 2	-9.88%	7.16%
Model 3	-10.61%	8.86%
Model 4	-7.69%	7.09%
Model 5	-10.77%	9.36%
Model 6	-11.93%	10.96%
Model 7	-8.01%	7.05%
Model 8	-8.28%	7.78%
Model 9	-7.13%	7.89%
Model 10	-10.84%	8.95%
Averages	**-9.60%**	**8.35%**

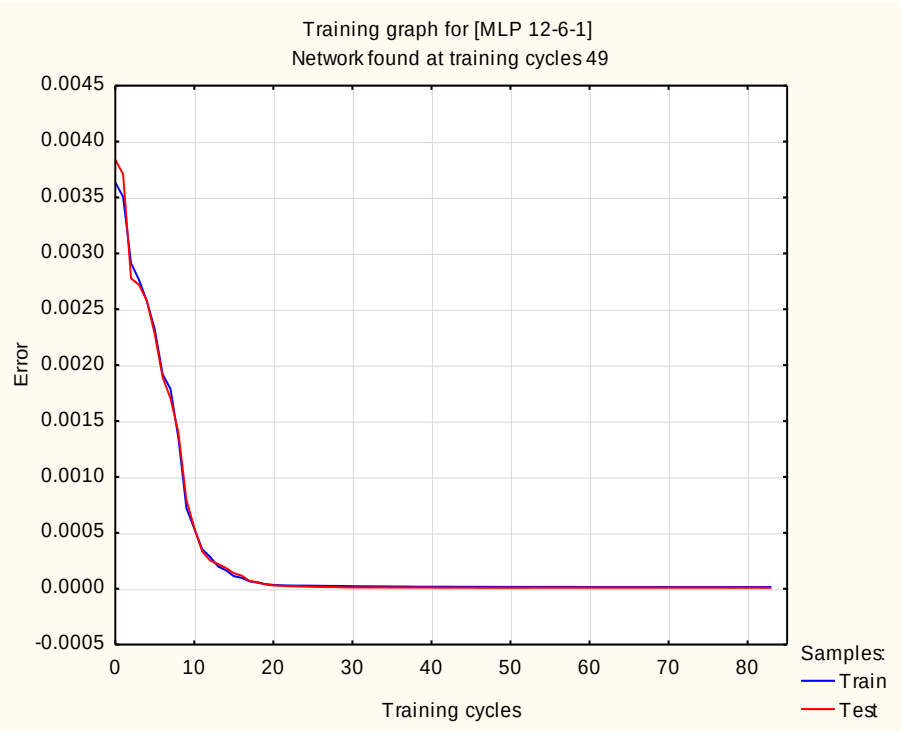


Figure 27: Training graph for Standard Model 4

4.4.2. Bootstrapping

During the development of the standard models, the dataset was divided into three exclusive subsets for training, testing and validating as is customary for neural network modelling. On second thought, this might actually not be getting the best out of the data as not all the data gets used for training, testing or validation. It is possible that some level of information is lost as the entire database is not used for the actual learning process.

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. Hair *et al.* (1998) summarise the steps taken in bootstrapping. This involves designating the original dataset to act as the population, then randomly re-sampling it a specific number of times to generate a large number of new samples. A different combination of new sub-samples is then used each time for training, testing and validating the model before averaging the model performance across the samples. By this approach, each data case gets used for training, testing and validation at least once during model development. This helps to glean as much information as possible from the entire dataset. Bootstrapping has been successfully applied to sunset detection, outdoor scene classification, and automatic image orientation detection by Jiebo *et al.* (2005) and for face detection in 3D by Schneiderman and Kanade (2000). Bootstrapping was used in both of these studies to overcome the problem of the large variety of input combinations and the limited amount of data available.

Furthermore, 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, non-parametric bootstrapping provides a way to estimate a statistic of population without explicitly deriving the sample distribution.

Statisticians, however, disagree 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.* (2010) 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 sample 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, 2,000 neural network models were again 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.

The bootstrapped models showed a far more consistent performance and produced smaller MSEs in comparison with the standard models as shown in Table 14. The results from the 100 validation cases of the bootstrapped models were also superior to those achieved by the standard models. While 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% (see Table 15).

Figure 28 shows a visual plot of the performance of the best 10 models from both the standard and bootstrapped models, validated over the 100 validation cases. It is obvious that the bootstrapped models far outperform the standard models. This performance improvement can be attributed to the fact that by using the 600 bootstrapped sub-

samples afforded the models a wider learning space than the standard models. These bootstrapped models were then carried forward for the final analysis in building the ensemble models.

Table 14: Summary of best 10 bootstrapped models

Model		Correlation Coefficient			MSE			Hidden activation	Output activation
		Training	Test	Validation	Training	Test	Validation		
1	MLP 12-3-1	0.9984	0.9980	0.9997	0.0013	0.0026	0.0007	Tanh	Identity
2	MLP 12-4-1	0.9984	0.9981	0.9997	0.0013	0.0028	0.0008	Logistic	Identity
3	MLP 12-5-1	0.9984	0.9980	0.9997	0.0012	0.0031	0.0012	Logistic	Identity
4	MLP 12-3-1	0.9985	0.9982	0.9998	0.0012	0.0023	0.0003	Tanh	Identity
5	MLP 12-5-1	0.9985	0.9981	0.9995	0.0012	0.0025	0.0019	Tanh	Identity
6	MLP 12-3-1	0.9984	0.9982	0.9998	0.0013	0.0025	0.0005	Tanh	Identity
7	MLP 12-3-1	0.9984	0.9982	0.9997	0.0012	0.0023	0.0008	Logistic	Identity
8	MLP 12-5-1	0.9984	0.9982	0.9995	0.0012	0.0031	0.0013	Logistic	Identity
9	MLP 12-5-1	0.9984	0.9983	0.9997	0.0013	0.0026	0.0010	Logistic	Identity
10	MLP 12-7-1	0.9985	0.9980	0.9994	0.0012	0.0026	0.0011	Tanh	Identity
Average		**0.9984**	**0.9981**	**0.9997**	**0.0012**	**0.0026**	**0.0010**	-	-

Table 15: Bootstrapped Model Performance

	Bootstrapped Model Performance	
	% Average Underestimate	% Average Overestimate
Model 1	-6.27%	3.21%
Model 2	-6.32%	4.42%
Model 3	-4.98%	4.31%
Model 4	-5.68%	3.54%
Model 5	-4.78%	2.84%
Model 6	-7.36%	3.79%
Model 7	-5.67%	3.15%
Model 8	-5.80%	4.21%
Model 9	-5.21%	4.89%
Model 10	-6.06%	4.07%
Average	**-5.81%**	**3.84%**

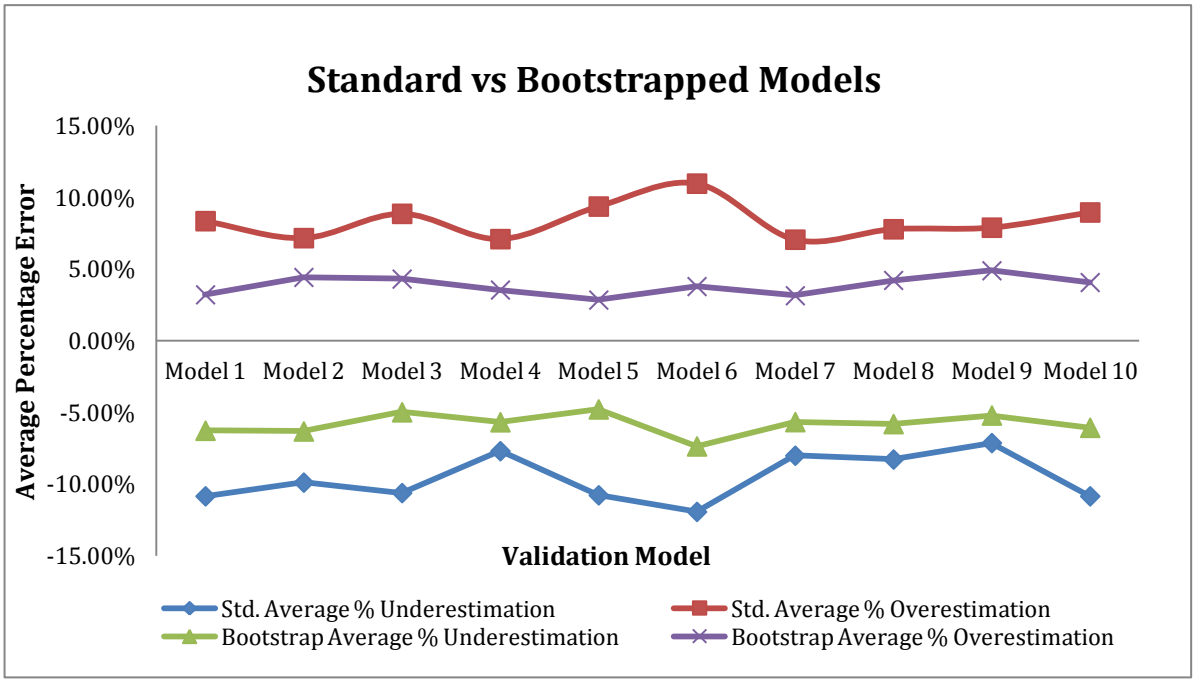


Figure 28: Comparison of Model Performance (Standard vs Bootstrapped)

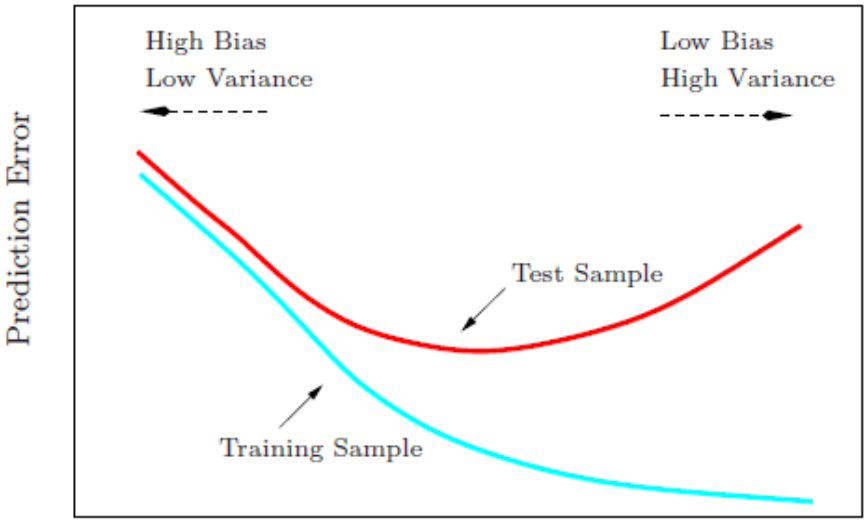
4.4.3. Ensemble Modelling

All modelling techniques are prone to two main types of error: bias and variance, according to Hastie *et al.* (2009). This is largely due to the fact that models essentially try to reduce complicated problems into simple forms and then attempt to solve the 'reduced' problem using an imperfect finite dataset. Bias generally refers to the difference between the mean of the estimated values and the true value while variance represents the variability over all estimated values (Zhang *et al.* 2014).

Skitmore *et al* (1990) describe a framework for reckoning the quality of estimates using three measures - bias, consistency and accuracy. While limiting their work to forecasts and contract bids, they defined bias as the average (mean) of the difference between actual tender price and the forecasted price. This bias has two main sources, according to Aibinu and Pasco (2008): bias associated with the estimating technique employed and environment as well as the bias contributed by the project itself. The lower the bias, the better the estimate. Consistency however refers to the 'degree of variation around the average'- the variance. The lesser this variance, the more consistent the estimate so that a low consistency might be equated to efficiency of the estimation process. Accuracy combines both bias and consistency so that an estimate with both low bias and variance measures is said to be accurate (Skitmore *et al.* 1990).

The relationship between variance and bias has been subject of studies by Geman *et al.* (1992), Zhou *et al.* (2002) and Hastie *et al.* (2009). This relationship can be summarised in Figure 29. Hastie *et al.* (2009) observe that as model complexity is increased, variance generally increases while the squared bias of the model decreases. The opposite also holds true. High variance models are over-fitted models that perform well on training sets but fail to generalise adequately when new data is presented to the model. High bias models, i.e. simple models, also under-fit the data and fail to learn effectively from the data.

This also unfortunately results in poor generalisation. It is important therefore to choose models that achieve a trade-off between variance and bias.



source: Hastie el 2009

Figure 29: Bias and Variance Trade-Off

The use of ensemble modelling helps to circumvent this problem by combining individual models in a way that achieves some sort of compromise between variance and bias. Also referred to as committee methods by Oza (2006), ensembles attempt to leverage the power of multiple models to achieve better prediction accuracy than any of the individual models could reach on their own. It is perhaps a way of consulting a committee of experts before reaching a final decision either by averaging, bagging, voting or by a 'winner-takes-all' procedure, whichever is most appropriate (Jordan and Jacobs 1994). 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 in Statistica® was applied to combine the 10 best bootstrapped models to trade off bias and variance. This also proved to be effective as performance from the bootstrapped models were further improved. Table 16 summarises the performance

of the ensemble models with the bootstrapped and standard models, visually illustrated in Figure 30. It is obvious that significant improvement has been achieved by applying the ensemble technique to the 10 bootstrapped models. On average, the ensembles overestimate final cost of the project by 2.33%. When they underestimate, they do so by an average of 3.83%.

Table 16: Summary of results (Standard, Bootstrap & Ensemble Models)

Model	*Average percentage error*	
	Overestimate	Underestimate
Standard models	8.35%	-9.60%
Bootstrapped models	3.84%	-5.81%
Ensemble model	2.33%	-3.83%

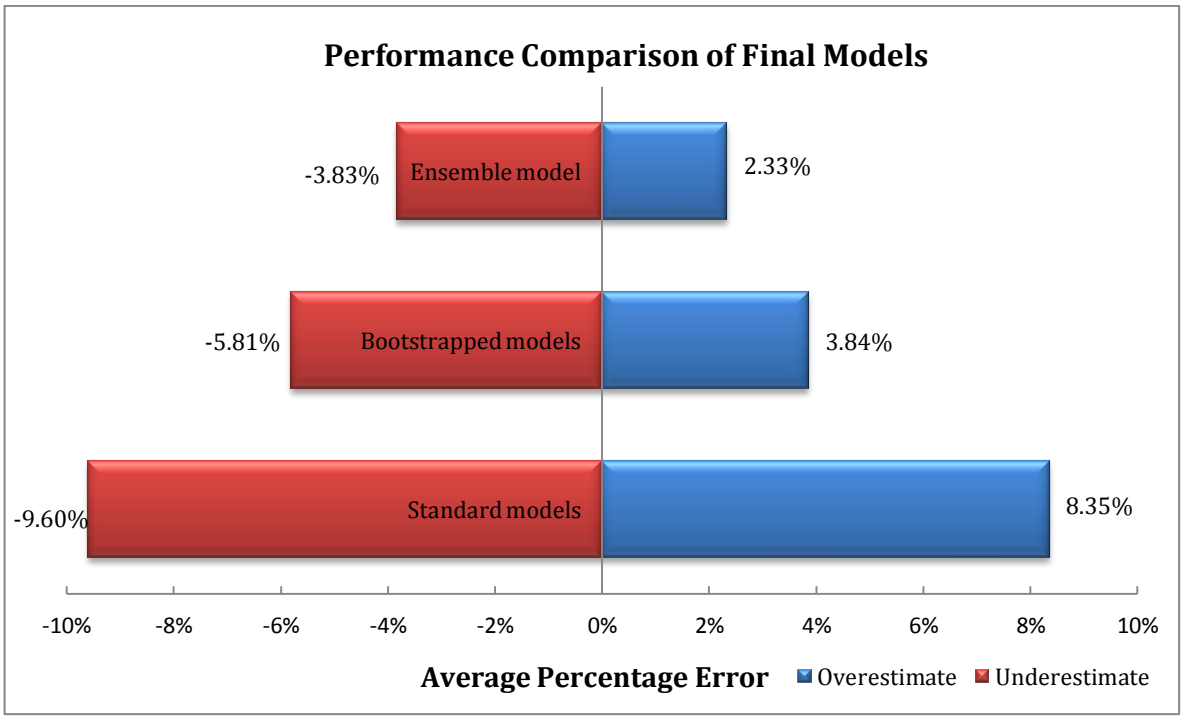


Figure 30: Bar chart showing the performance of the final models

For a more accessible comparison of model performance, Table 17 shows a random sample of 20 results out of the 100 validation cases used to test the ensemble model, further summarised in Table 17. It details a comparison between the ensemble's final cost prediction and the actual final cost of the project, with a measure of the actual monetary error observed. Overall, 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. The absolute percentage errors of the ensemble ranged between from 0.04% to 15.85% of final actual cost.

Table 17: 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 18: 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%**

4.5. CHAPTER 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 strategies. The possibilities are that these decision-support systems could help construction practitioners in making better informed and reliable decisions especially in the early planning stages of the project.

Hitherto, the scale and sources of cost overruns on construction projects have been thoroughly examined, particularly through the lens of the PsychoStrategists and Evolution Theorists. A distinction between the often conflated issues of overruns and underestimation was also clarified using the conceptual model in .

As already pointed out, 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. Cost overruns can thus be attributed to the unavailability of necessary information for reliable estimation. This chapter thus presents a practical means of transforming information embedded in existing construction data into decision-support systems that can complement traditional estimation methods for more reliable final cost forecasting.

Using standard artificial neural networks, non-parametric bootstrapping and ensemble modelling, final project cost forecasting models were developed with 1,600 completed projects. While the standard neural network models achieved acceptable validation performance of +8.35% and -9.60% average percentage over and underestimation respectively, the bootstrapped models significantly reduced this error to +3.84% and -5.81% error. The use of committee modelling in the ensembles to achieve a compromise between bias and variance, further improved the prediction results of the bootstrapped

models. 77% of the validation predictions using the ensemble models were within ±5% of actual final cost, an indication of the model's ability to learn the underlying structure and correlations within the database to predict final cost.

Although it is acknowledge that reliable cost planning and estimation form only one aspect of dealing with cost overrun in construction. A more holistic approach must include effective project governance, client leadership, accountability and measures of cost control. However, the results from the models show significant promise for future work on construction data mining to support the estimation process, thereby potentially producing more reliable and realistic final cost estimates.

The models 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 as well as their likely total financial commitment on each project. The models could further allow 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.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATION

"The outcome of any serious research

can only be to make two questions grow

where only one grew before."

Thorstein Veblen

“UNLESS someone like you cares a whole awful lot,

nothing is going to get better. It's not.”

Dr Seuss, The Lorax

5.0 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. Cost estimates prepared in the early stages of a project allow the client to perform a cost-benefit analysis, secure funding, evaluate most economical tenders as well as used as a basis for cost control during project delivery. Where the project is a commercial asset, the initial capital investment must be balanced with the cost of maintenance and operations over the life-time of the project to ensure that the project remains profitable and planned returns on capital investment is achieved over an estimated period. Decisions made at the early stages of the project development therefore carry far-more reaching economic consequences and can seal the financial fate of a project.

However, most construction projects routinely overrun their initial cost estimates leaving clients, financiers, contractors and the public dissatisfied. Clients may have to secure extra funding or suffer reputational detriments. Financiers may have to suffer the consequences of their investment not returning profits for a longer period. Contractors could face cash flow issues, liquidity and damage to their business image while the public, where the project is funded by taxpayer's money, have to pay more for a problem that was not their fault.

Hitherto, the thesis has addressed the "what?", "why?" and "how?" questions of construction cost overruns. A researcher has not completed his or her task unless they are confidently able to respond to a 'so what?' query as well as project forward the meaning of their work, placing it the context of the wider field of study. This chapter would thus make sense of the key findings and results of the study by tying together and synthesising the arguments and results in the body of the thesis to the aims and objectives set out in the introductory chapter. The chapter will answer the research questions and identify the

theoretical and practical implications of the study. Lessons for the infrastructure delivery and management of construction projects will also be made with some suggestions for future research.

5.1. REVIEW OF ORIGINAL AIMS AND OBJECTIVES

Following the arguments laid out in Chapter Two and the subsequent data analysis in Chapter Four using neural networks, bootstrapping and ensemble modelling, the following conclusions have been reached in line with the stated objectives of the research:

Objective 1a: *To ascertain through a critical review of the literature, the factors that contribute to the difference between the initially estimated cost and the resulting final costs at project completion*

The literature review in Chapter Two focused on contemporary mainstream arguments on the causes of overruns on construction projects. These include poorly managed risk and uncertainty, lack of information for reliable estimation in the early stages of the project, scope changes and rework. There was a strong case for the deliberate distortion or misstatement of the likely level of resources necessary to deliver a project and unjustifiable optimism bias - the tendency to evaluate possible negative future events in a fairer light than suggested by inference from the base rates. Specific evidence in support of the latter two sources were identified on the Scottish Parliament project and the Perth Arena project in Australia (Sections 2.2.2 and 2.2.3)

Objective 1b: *Explore the different theoretical schools of thought on the cost overruns*

The literature is quite clear - there are essentially two prevalent schools of thought on cost overruns, referred to in the thesis as the PsychoStrategists and the Evolution Theorists (Section 2.2). The PsychoStrategists attribute cost overruns mainly to deception (strategic misrepresentation) and delusion (optimism bias). The Evolution Theorists however posit that projects change in scope and definition

between the conception stage and project construction phase. These changes are the drivers of cost overruns according to the evolutionaries.

Objective 1c: *Synthesise the different schools of thought into a holistic conceptual model to help properly understand overrun.*

The PsychoStrategists and the Evolution Theorists hold opposing views on the sources of 'overruns'. They also measure 'overruns' from two different reference points - the former, from the cost at the time of the decision-to-build while the latter group measures from cost at contract award. This leads to large disparities in the size of cost 'overrun' reported from the two perspectives.

However, using the conceptual model in Figure 2, the two schools can be usefully viewed as two complementing sides of the issue - both valid and crucial to thoroughly understanding or tackling 'cost overrun'.

Based on Figure 2, underestimates have been described in this thesis as the difference between the estimated cost at project inception, where the decision-to-build is taken and the estimated cost at the end of the project definition stage. The main contributors to cost underestimation at this stage are a lack of information, significant scope changes, estimation error, strategic misrepresentation and optimism bias.

Overruns, however, are more appropriately described as the difference between cost at project completion and project definition stage. This is usually as a result of further scope changes (usually not as significant as those before detailed design), rework, ground conditions, technical and managerial difficulties, material price changes or estimation error.

Aim 1: To provide a better conceptual understanding of "cost overruns"

Aim 1 of the research can now be evaluated based on the conclusions reach in the objectives above. Much of the media furore and arguments in the literature on the so-called overruns hardly make a differentiation between the causes of 'underestimation' and 'overruns' as identified in

Figure 2. Media reportage, in particular, is usually based on a comparison between cost at inception of the project and resulting cost at completion of a project, ignoring the mediating phases of project gestation and definition. Very often, this comparison is between two projects that are significantly different in scope and design.

Furthermore, the conflating definitions and reference points for assessing cost performance on projects have resulted in large disparity in the level of overruns reported. Should the reference point for measurement cost growth be recalibrated to the point of contract award for example, it is very likely that not "nine out ten" projects actually "overrun their budgets" and infrastructure projects may not have an 86% likelihood of outrunning their budgets, as stated in the opening sentence of the thesis. It would also likely mean that the size of overrun, as reported by Flyvbjerg for example, will become significantly lower.

In addition, focussing on one side of the debate in dealing with overruns will do little to effectively tackle cost overruns in the management of construction projects. PsychoStrategic theorists neglect well documented issues like design problems, unforeseen ground conditions, scope changes and rework that drive up cost during the actual project construction. The unfortunate consequence of this perspective also brands planners, project promoters and estimators as unethical and suspicious without sufficient evidence to sustain the supposition. An evolutionary theorist perspective alone, on the other hand, would also rather naïvely not fully accommodate the strong influence and dynamics of business strategy, competition, power and organisation politics in setting unrealistic cost targets that will inevitably be unattainable.

Finally, as a result of the discussions in Chapter Two regarding the term 'overruns', it is proposed that the phrase 'cost growth' better describes the increase in estimated cost from one phase of the project to another. 'Overruns' indirectly imply that there exists a single, accurate and

deterministic estimated figure to which all others have to be compared. Perhaps it is for this very reason the word 'estimate' is used to refer to the projections of likely future cost based on known and available information at the time of forming those estimates. Consequently, cost certainty in terms of an estimate, would seem a rather false notion. Estimates can only get more accurate as more information is available.

The journal publications "*Rethinking construction cost overruns: Cognition, learning and estimation*" (Journal of Financial Management of Property and Construction) and "*Dealing with construction cost overruns using data mining*" (Construction Management and Economics) are based on the arguments that resulted from achieving Aim 1.

***Objective 2a**: Identify and collect a reliable dataset for the cost modelling process.*

Recognising the potential benefits of converting their existing database of projects into decision support tools for cost estimation, two collaborating industry partners got involved in this research after the concept and advantages of using data mining was presented to them. The initial data collection process involved shadowing of the tendering and estimation procedure in these firms as a quasi member of their tendering teams. This provided the opportunity to gain a first-hand understanding of how the data to be used for the modelling was generated and what different variables meant.

One dataset of 98 project cases with a total value of about £99 million was collected from Morrison Construction, a UK Civil Engineering contractor. These projects were completed between 2007 and 2011. This first dataset was used for developing trial models to test different modelling strategies and experiment with using neural networks for cost modelling.

The second dataset of approximately 1,600 projects was collected from a major public utility company in Scotland[1]. These projects were also completed fairly recently between 2009 and 2012. About 99% of the total number of project cases cost less than £25 million with only 3 projects costing more than this figure. The total value of these projects was over £800 million. 80% of the projects were completed within 3 years, with only 64 projects completed after 5 years. The average duration of the projects was about 24 months. See Section 4.3.

[1] Name withheld for confidentiality purposes

***Objective 2b**: Develop the cost models to estimate likely final cost of projects.*

The concept and approach of using artificial neural networks for construction cost modelling was initially piloted using the smaller dataset of 98 projects. Different training algorithms, transfer functions, neural network architectures and data transformations were experimented with using this dataset (See Section 4.3.1). Trial final cost estimation models were also developed using this dataset. The significant input variables for the model in this dataset include ground condition, project duration, tendering strategy, estimated cost at contract award and contractor's need for the project.

The performance of the final cost model from this dataset is indicated in Figure 12. The error range of this model was between -2% (underestimation) to 7.9% (overestimation) with an average error of -1.8% underestimation. This compares favourably with the -10% to +15% estimation error commonly found and accepted in practice (Potts 2008). The analysis and results of the trial models from this dataset have been published in the conference papers: "*Neural networks for modelling the final target cost of water projects*" and "*A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects*" (attached as Appendix A5 and A6)

Having achieved successfull results with the smaller dataset, the lessons, experience and approach used were then expanded to the larger dataset of 1,600 project cases. Three different modelling strategies, namely, standard neural networks, data boostrapping and ensemble modelling were adopted sequentially, with each new strategy designed to overcome a weaknesses of the previous. The overall aim of the modelling is to convert information embedded in historical construction data into cost models that could aid the estimation process in early stages of the project. Details of the analysis carried out under the three different modelling strategies are detailed from Section 4.4 of Chapter Four. Two published conference papers, "*My cost runneth*

over": Data mining to reduce construction cost overruns" and "*Dealing with construction cost overruns using data mining*" are based on the analysis in dataset 2.

***Objective 2c**: validate the models using new project cases.*

The models developed with dataset two were validated using 100 project cases that were not used in the model training. This was a way to ascertain whether the models had adequately discovered the underlying relationships and structure within the dataset in order to make reliable estimations of likely final cost of new projects.

The validation results have been presented in Figure 30 and Table 16. The results show that while the standard neural network models achieved acceptable validation performance range of -9.60% to +8.35% error on average, the bootstrapped models significantly reduced this error range to -5.81% to +3.84% (+ for over estimation, - for under estimation). The use of ensemble modelling to achieve a compromise between bias and variance, further improved the prediction results of the bootstrapped models to a range between -3.83% and +2.33%. 77% of the validation predictions using the ensemble models were within ±5% of actual final cost.

Aim 2: To develop cost models for estimating final cost of projects based on historical cost and project details of completed project.

The business landscape is continually experiencing a growing recognition of information as a key competitive tool. Companies that are able to successfully collect, analyze and understand the information available to them are among the winners in this information era. Most construction firms maintain copious information on each project undertaken. This data can usefully be transformed into decision support tools for the improvement in the reliability of cost.

Data mining, an analytic process for exploring large amounts of data in search of consistent patterns and systematic relationships between

variables, has been used in this research to develop final cost estimation models based on data collected from two industry partners. Neural networks, the chosen data mining technique for this research, was then used to scour these datasets to in order to find predictive knowledge for final cost estimation in the early stages of a project where the information required for a thorough estimation is largely unavailable.

The highly satisfactory level of performance of the models (as shown in Figure 30 and Table 18: Summary of validation performance of ensemble model) demonstrate the significant promise of effectively using neural networks and cost modelling techniques to increase the reliability of early stage cost estimates based on historical cost and project data. The journal paper "*Dealing with construction cost overruns using data mining*" (Construction Management and Economics) [Appendix A2] and the conference paper *"My cost runneth over": Data mining to reduce construction cost overruns*" [Appendix A3] were published to demonstrate the potential benefits of using data mining for early cost estimation in an attempt to reduce cost overruns.

5.2. ANSWERING THE RESEARCH QUESTIONS

Having achieved the aims and objectives of the research, a response is now provided for the research questions as below:

1. *Is the current understanding of construction cost overruns adequate?*

The study suggests that the current understanding of 'cost overruns' is inadequate and at best fragmented. There is no unanimity on the appropriate reference point from which cost performance on projects could be assessed leading to a large disparity in the scale of overruns reported in the literature. Some measure overruns as the difference between estimate at the time of decision to build and final completion cost, while others measure from the estimate at contract. As was evident in the arguments clearly laid out in Section 2.3, and supported by examples from real projects, the scope and design of some projects

tend to change significantly along the different phases of the project life cycle. This confusion sometimes leads to erroneous and misleading comparison of what has been termed 'apples and oranges' in some cases.

2. *What are the predominant schools of thought on the sources of construction cost overruns?*

Two predominant schools of thought have been identified in the study - The PsychoStrategists attribute cost overruns mainly to deception (strategic misrepresentation) and delusion (optimism bias). The Evolution Theorists, on the other hand, attribute overruns to change. They posit that projects essentially evolve in scope and definition between the conception stage and project construction phase, with attendant growing cost.

3. *Is there a conceptual difference between cost underestimation and cost overruns?*

Table 2 has been used to make a distinction between the two terms. Underestimates are appropriately measured as the difference between the estimated cost at project inception, where the decision-to-build is taken and the estimated cost at the end of the project definition stage. The main contributors to cost underestimation are a lack of reliable information to base the estimates on, significant scope changes as a result of project definition, estimation error, strategic misrepresentation and optimism bias.

Overruns, however, are more appropriately described as the difference between cost at project completion and project definition stage. This is usually as a result of further scope changes (usually not as significant as those before detailed design), rework, ground conditions, technical and managerial difficulties, material price changes or estimation error.

4. *Is neural networks an appropriate method of estimating the cost of construction projects?*

The satisfactory results achieved at the development stage of the research demonstrate the appropriateness and potential for using

neural networks for cost estimation in construction. Some of the advantages of using neural networks include its ease of use, power and ability to model complex non-linear relationships between a large number of variables without having to first establish any a priori conditions. They are particularly recommended where the relationships between variables are vaguely understood or difficult to describe by conventional approaches.

The most important point of caution might be that neural networks should be used when the goal of the modelling exercise is to measure *how well*, rather than thoroughly understand the *why* of a phenomenon. The why of the problem must be established in other ways as it is difficult to elucidate cause and effect relationships using neural networks.

5.3. SO WHAT? MAKING SENSE OF THE RESEARCH CONTRIBUTION

The aims and objectives have been achieved and research questions answered. But, it is imperative that sense is made of these outcomes in order to place the research within the wider context of construction management research and practice. As previously intimated, it is insufficient to simply address the "what?", "why?" and "how?" questions of construction cost overruns: we also need to know what difference the research makes.

5.3.1. Theoretical Contribution

Existing theories on the causes of overruns can be separated into two: one from an engineering and technical perspective, described as the Evolution Theorists, and the other emerges from an economic, psychological and strategic point of view, termed the PsychoStrategic Theory. These two views are both critical to holistically understanding and dealing with the problem of cost growth and therefore should be viewed as two complementing, rather than opposing, sides of the same issue.

The second theoretical contribution is summarised in the conceptual model shown in Figure 2 that helps to distinguish 'overruns' from underestimates, along with their causes. The two terms can be directly linked to the schools of thought above. PsychoStrategists such as Flyvbjerg *et al* (2002, 2008) and Wach (Wachs 1989, 1990) focus on underestimation while Evolution Theorist, such as Love *et al* (2012, 2014) and Odeck (Odeck 2004) focus their analysis on 'cost overruns'.

The third theoretical contribution is the introduction of the term 'cost growth' to describe the increase in estimated cost from one phase of the project to another, instead of the loosely used term 'cost overruns'. Overruns indirectly imply, and perhaps misleadingly so, that there exists a single, accurate and deterministic estimated figure to which all others have to be compared. This is practically unrealistic. Consequently, it is unreasonable to think of cost estimates on a construction project as 100% accurate. Their degree of realism and reliability would only be increased as more information becomes available for the estimation process.

Finally, the research contributes to existing knowledge on cost modelling approaches. It demonstrates the use of neural networks with data bootstrapping and ensemble modelling for developing final cost models based on historical data. As far as can be ascertained, this combination of modelling approaches has not been used in any construction cost estimation related research.

5.3.2. Contributions to Practice

The contribution of the study to practice can be viewed in two ways - a more holistic understanding of the sources of cost growth and the presentation of a neural network method for developing cost models from existing project information.

The industry is usually focussed on either one or the other schools of thought regarding cost growth. Focussing only on one side of the debate will probably not help much to effectively tackle the problem in the

procurement, governance and delivery of projects as PsychoStrategic view neglects prominent issues like design problems, unforeseen ground conditions, scope changes and rework that drive up cost during the actual project delivery. An evolutionary perspective alone, on the other hand, would also rather naïvely not fully accommodate the strong influence and dynamics of business strategy, competition, organisation politics in setting unrealistically low cost targets that will inevitably be unattainable. The industry may therefore need to start recognising and dealing with the sources of cost growth from both perspectives to deal effective with the problem.

Secondly, the use of artificial neural networks for cost modelling has been demonstrated as a possible avenue for converting existing data within construction organisations into decision support tools, especially where information is lacking or inadequate. The use of cost models can help clients, project managers, financiers or contractors to:

- pro-actively predict deviations in cost estimates;
- improve the reliability of their initial cost plans;
- enhance early identification of potential problems on a project;
- minimise late changes and their associated costs; and
- ultimately increase customer satisfaction.

The particular models developed in this research will be directly beneficial to the partnering firms in the research as the data used is company specific. When extending and translating the models to other environs and project types further specific data is clearly needed – it is not safe to assume the nature of the work considered here can be extrapolated without question – but the approach to data collection, data pre-processing and eventual model development using neural networks, bootstrapping and ensemble modelling can allow the extension to any relevant dataset or type of project very feasible.

5.4. RECOMMENDATIONS

The following recommendations have been made based on the key findings of the study:

1. **Understanding is crucial**

A thorough understanding of the sources of overruns on projects is an important precursor to effectively deal with the problem. Clients and industry need to revisit their thinking on overruns if any improvement in cost reliability is desired.

2. **Data is an asset**

The real value of data lies in using it to the advantage of any organisation. The construction industry thus needs to recognise and use the vast data available to them from past projects to support their cost estimation process.

3. **Client is key**

It is imperative to clarify questions about project scope and who has ultimate responsibility, on behalf of the client, to govern the project. This could have profound implications on cost growth from inception to completion of the project.

4. **Be realistic**

As was suggested by a Commercial Manager of a large construction company in the UK during the study, "winning a bid is easy. But winning at the right price is difficult". Clients should award contracts based on the *realistic* tenders submitted, rather than the lowest evaluated tender. Cultural changes within the industry towards the search for realistic targets might incentivise contractors to flag-up potential pitfalls early-on.

5.5. LIMITATIONS OF STUDY

1. Validation in practice

The models produced in this research have gone through a validation process using a sample data kept aside for that purpose. A further stage of validating the models would be to test them against a project yet to be undertaken. The scope and timeframe of the present research did not allow for such level of testing.

2. More detailed project attributes

The final models used high level project attributes like delivery partner, duration, scope of project, purpose of project and location. It is possible that the predictive performance of the models could further be improved should more detailed project level information on earthworks, concreting, plants, schedule of dayworks be used.

3. Data warehousing

The success of a data mining exercise depends heavily on the availability of business, operational and project data, stored in a meaningful and retrieval manner. For most construction companies, relevant data for modelling construction processes is sparse, fragmented or stored in ways that will make the use of data mining practically difficult, or even impossible. The poor culture of data collection and warehousing in the construction industry is expected to be perhaps one of the major limitations of using data mining in practice.

5.6. FURTHER RESEARCH

While tangible and worthwhile outcomes have been achieved from this research, there is still scope and room for further work and improvement. Long term progress and development in this field is not under consideration here. Instead, below are two research strategies that should be undertaken in the near future.

1. Neuro-Fuzzy Modelling

As discussed in Chapter Four, the modelling philosophy adopted is one of continuous and incremental improvements. In other words, only the

arrogant modeller assumes his or her work is complete in itself. For possible further improvement in the results already achieved, neuro-fuzzy modelling could be used. This approach allows the learning and generalisation capabilities of neural networks to be combined with the capacity for tolerance and imprecise knowledge representation of fuzzy set theory. It has the possibility of increasing the reliability and flexibility of the models. As indicated in Section 4.3.1, this approach was piloted using dataset 1 with initial results published in the conference paper "*A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects*" [Appendix A6].

A three-point fuzzy lower, upper and mean estimate of likely final cost was generated to provide a tolerance range for final cost rather than the traditional single point estimate. The performance of the final models using dataset 1 ranged from 3.3% underestimation to 1.6 % overestimation. Time constraints did not allow for the approach to be extended to dataset 2.

2. Validation In Practice

As discussed in the research limitations above, the work is currently only validated based on historical data. Validation in practice is a process which needs consideration to provide greater usefulness and acceptance of the models produced.

5.7. FINAL THOUGHTS

The scale of the problem of cost overruns in the global construction industry has been presented. Overruns occur irrespective of the size or type of project, its geographical location, procurement method or duration. There are essentially two schools of thought on the causes of overruns, referred to in the thesis as the PhychoStrategic and Evolution Theorists. These two perspectives have existed for many decades and it is inevitable that they will continue to dominate, causing the type of confusion and misunderstanding of construction of outcomes that have plagued the industry and have been outlined in detail in earlier

chapters of this thesis. A conceptual model that holds the two views as complementary, rather than opposing perspectives has been presented.

It is unlikely that the construction industry will be able to adequately deal with the problem of cost overruns if only one of these perspectives is focussed upon. *Unless* there is fresh thinking and a realisation in both academia and industry that previous thinking is over-simplistic, and conflated views of cost overrun are abandoned, there will be no progress to achieving better *value and satisfaction* for all construction stakeholders. The theoretical contribution of this thesis makes clear the detriments of continuing to separate the existing schools of thought and the benefits of rethinking cost overruns.

REFERENCES

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

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

Ahiaga-Dagbui, D D and Smith, S D (2012) Neural networks for modelling the final target cost of water projects. In: *Procs 28th Annual ARCOM Conference*, Smith, S D, Ed., Edinburgh, UK: Association of Researchers in Construction Management, 307-16.

Ahiaga-Dagbui, D D and Smith, S D (2013) "My cost runneth over": Data mining to reduce construction cost overruns. In: *Procs 29th Annual ARCOM Conference*, Smith, S D and Ahiaga-Dagbui, D D, Eds.), Reading, UK: Association of Researchers in Construction Management, 559-68.

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

Ahiaga-Dagbui, D D and Smith, S D (2014b) Dealing with construction cost overruns using data mining. *Construction Management & Economics*, **32**(7-8), 628-94.

Ahiaga-Dagbui, D D and Smith, S D (2014c) Exploring escalation of commitment in construction project management: Case study of the Scottish Parliament project. In: *Procs 30th Annual ARCOM Conference, 1-3 September, 2014*, Raiden, A B and Aboagye-Nimo, E, Eds.), Portsmouth, UK: Association of Researchers in Construction Management 755-64.

Ahiaga-Dagbui, D D, Tokede, O, Smith, S D and Wamuziri, S (2013) A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects. In: *Procs 29th Annual ARCOM Conference*, Smith, S D and Ahiaga-Dagbui, D D, Eds.), Reading, UK: Association of Researchers in Construction Management, 181-90.

Aibinu, A A and Pasco, T (2008) The accuracy of pre-tender building cost estimates in Australia. *Construction Management and Economics*, **26**(12), 1257 - 69.

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

Al-Tabtabai, H and Alex, P (1999) Preliminary Cost Estimation of Highway Construction Using Neural Networks. *Cost Engineering*, **41**(3), 19.

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**, 745-56.
Anderson, D and McNeill, G (1992) Artificial neural networks technology. *A DACS (Data & Analysis Center for Software) State-of-the-Art Report, Contract Number F30602-89-C-0082*, 87.
Anderson, J A (1995) *An Introduction to neural networks*. Cambridge, Massachusetts: MIT Press.
Audit Scotland (2000) *The new Scottish Parliament building, an examination of the management of the Holyrood project*, Edinburgh, UK: Audit Scotland.
Audit Scotland (2004) *'Management of Holyrood Building Project' (Audit Report prepared for the Auditor General of Scotland)*, Edinburgh, UK: Audit Scotland.
Auditor General (2010) *The Planning and Management of Perth Arena: Media Statement by the Auditor General for Western Australia, Colin Murphy*, Office of the Auditor General, Australia, http://tinyurl.com/n8fpzoh.
Auditor General of Western Australia (2012) *Managing Capital Projects*, Perth, Australia: Office of the Auditor General of Western Australia, http://tinyurl.com/l9ymlqu (last accessed in May 2014).
Baccarini, D (2005) Estimating project cost contingency – Beyond the 10% syndrome. In: *2005 Australian Institute of Project Management Conference*: AIPM.
Bode, J (2000) Artificial neural networks for cost estimation: simulations and pilot application. *International Journal of Production Research*, **38**(6), 1231-54.
Bordat, C, McCullouch, B G, Sinha, K C and Labi, S (2004) *An Analysis of Cost Overruns and Time Delays of INDOT Projects.*, Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, Indiana, Publication FHWA/IN/JTRP-2004/07.
Boussabaine, A and Elhag, T (1997) A neurofuzzy model for predicting cost and duration of construction projects. *RICS Research (9 p.). The Royal Institution of Chartered Surveyors*.
Bryman, A (2012) *Social research methods*. Oxford university press.
Burger, R (2003) Contingency, quantifying the uncertainty. *Cost Engineering*, **45**(8), 12-7.
Cao, Q, Parry, M E and Leggio, K B (2011) The three-factor model and artificial neural networks: predicting stock price movement in China. *Annals of Operations Research*, **185**(1), 25-44.
Chase, S, Weiss, M, Gibbs, P, Hillman, C and Urban, N *The Physics and Relativity FAQ*. [http://math.ucr.edu/home/baez/physics/General/occam.html.] Accessed 19th November, 2012
City of Edinburgh Council (2014) *The Tram Project*. [http://www.edinburgh.gov.uk/trams.] Accessed 28-05-2014
Creedy, G D (2006) *Risk factors leading to cost overrun in the delivery of highway construction projects*, PhD, School of Urban

Development, Faculty of Built Envornment and Engineering, Queensland University of Australia.

Creswell, J W (2009) *Research design: Qualitative, quantitative, and mixed methods approaches*. Third ed. Califonia, USA: Sage Publications, Inc.

Dahl, G, Yu, D, Deng, L and Acero, A (2010) Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*(99), 1-.

Demir, F (2008) Prediction of elastic modulus of normal and high strength concrete by artificial neural networks. *Construction and Building Materials*, **22**(7), 1428-35.

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

Efron, B (1992) Bootstrap methods: Another look at the jackknife. *In:* Kotz, S and Johnson, N L (Eds.), *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.

Fellows, R and Liu, A (2008) *Research methods for construction*. Third ed. Chichester, West Sussex, UK; Malden, MA, USA: Wiley-Blackwell Publication.

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

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 (2008) Curbing optimism bias and strategic misrepresentation in planning: Reference class forecasting in practice. *European Planning Studies*, **16**(1), 3-21.

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 and Stewart, A (2012) *Olympic Proportions: Cost and Cost Overrun at the Olympics 1960-2012*, Oxford, UK: Saïd Business School, University of Oxford

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.

Flyvbjerg, B, Skamris Holm, M K and Buhl, S L (2005) How (In)accurate Are Demand Forecasts in Public Works Projects?: The Case of Transportation. *Journal of the American Planning Association*, **71**(2), 131-46.

Gelinas, N (2007) Lessons of Boston's Big Dig. *City Journal*, **Autumn 2007**, Accessed on 8th May 2014, http://tinyurl.com/dxxrdf

Geman, S, Bienenstock, E and Doursat, R (1992) Neural networks and the bias/variance dilemma. *Neural Computation*, **4**(1), 1-58.

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.

Gunduz, M, Ugur, L O and Ozturk, E (2011) Parametric cost estimation system for light rail transit and metro trackworks. *Expert Systems with Applications*, **38**(3), 2873-7.

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

Handzic, M, Tjandrawibawa, F and Yeo, J (2003) How neural networks can help loan officers to make better informed application decisions. *Informing Science and Information Technology Education*, 97-109.

Hastie, T, Tibshirani, R and Friedman, J (2009) *The elements of statistical learning*. Vol. 2, Springer.

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

Hebb, D O (1949) *Organisation of Behavior: A Neuropsychological Theory*. New York: John Wiley.

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

Hinze, J, Selstead, G and Mahoney, J P (1992) Cost overruns on state of Washington construction contracts. *Transportation Research Record*(Issue Number 1351), 87-93.

Jiebo, L, Boutell, M, Gray, R T and Brown, C (2005) Image transform bootstrapping and its applications to semantic scene

classification. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, **35**(3), 563-70.
Jordan, M I and Jacobs, R A (1994) Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, **6**(2), 181-214.
Kahneman, D (2011) *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.
Kahneman, D and Tversky, A (1979) Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, **47**(2), 263-91.
Knight, F H (1921) *Risk, uncertainty and profit*. The Riverside Press.
Kodogiannis, V S, Lygouras, J N, Tarczynski, A and Chowdrey, H S (2008) Artificial odor discrimination system using electronic nose and neural networks for the identification of urinary tract infection. *Information Technology in Biomedicine, IEEE Transactions on*, **12**(6), 707-13.
Lovallo, D and Kahneman, D (2003) Delusions of Success: How Optimism Undermines Executives' Decisions. *Harvard Business Review*(July 2003).
Love, P E D, Edwards, D J and Smith, J (2005) Contract documentation and the incidence of rework in projects. *Architectural Engineering and Design Management*, **1**(4), 247-59.
Love, P E D, Edwards, D J and Irani, Z (2012) Moving beyond optimism bias and strategic misrepresentation: An explanation for social infrastructure project cost overruns. *IEEE Transactions on Engineering Management*, **59**(4), 560-71.
Love, P E D, Sing, C-P, Wang, X, Irani, Z and Thwala, D W (2012) Overruns in transportation infrastructure projects. *Structure and Infrastructure Engineering*, 1-19.
Love, P E D, Smith, J, Simpson, I, Regan, M, Sutrisna, M and Olatunji, O (2014) Understanding the Landscape of Overruns in Transport Infrastructure Projects. *Environment and Planning B: Planning and Design*, **in press**.
McCulloch, W S and Pitts, W (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biology*, **5**(4), 115-33.
Miller, D (2011) Edinburgh Trams: Half a line at double the cost. *BBC, 18th February, 2013, http://goo.gl/mfr96*, http://goo.gl/mfr96*18th February, 2013*
National Audit Office (2012) *The London 2012 Olympic Games and Paralympic Games: post-Games review* HC 794- Session 2012-13, National Audit Office, UK.
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.
Olden, J D and Jackson, D A (2002) Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, **154**(1-2), 135-50.

Osland, O and Strand, A (2010) The Politics and Institutions of Project Approval-a Critical-Constructive Comment on the Theory of Strategic Misrepresentation. *EJTIR*, **1**(10).

Oza, N C (2006) Ensemble data mining methods. *In:* Wang, J (Ed.), *Encyclopedia of Data Warehousing and Mining*, pp. 770-6: IGI Global.

Pattengale, N D, Alipour, M, Bininda-Emonds, O R, Moret, B M and Stamatakis, A (2010) How many bootstrap replicates are necessary? *Journal of Computational Biology*, **17**(3), 337-54.

Portas, J and AbouRizk, S (1997) Neural Network Model for Estimating Construction Productivity. *Journal of Construction Engineering and Management*, **123**(4), 399-410.

Potts, K (2008) *Construction cost management: learning from case studies*. Taylor & Francis Group.

Pradeep, J, Srinivasan, E and Himavathi, S (2011) Neural network based handwritten character recognition system without feature extraction. *In*. IEEE, 40-4.

Pyle, D (1999) *Data preparation for data mining*. Vol. 1, Morgan Kaufmann.

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

Ravdin, P M and Clark, G M (1992) A practical application of neural network analysis for predicting outcome of individual breast cancer patients. *Breast Cancer Research and Treatment*, **22**(3), 285-93.

Ripley, B D (1993) Statistical Aspects of Neural Networks. *In:* Barndorff-Nielsen, O E, Jensen, J L and Kendall, W S (Eds.), *Networks and chaos: statistical and probabilistic aspects*, pp. 40-111. London, UK: Chapman & Hall.

Rosenblatt, F (1958) The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review*, **65**(6), 386.

Ross, A and Williams, P (2013) *Financial Management in Construction Contracting*. Sussex, UK: Wiley-Blackwell.

Santos, R B, Ruppb, M, Bonzi, S J and Filetia, A M F (2013) Comparison Between Multilayer Feedforward Neural Networks and a Radial Basis Function Network to Detect and Locate Leaks in Pipelines Transporting Gas.

Saridemir, M (2009) Prediction of compressive strength of concretes containing metakaolin and silica fume by artificial neural networks. *Advances in Engineering Software*, **40**(5), 350-5.

Sarle, W S (1994) Neural Networks and Statistical Models. In: *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, Cary, North Carolina, USA: SAS Institute Inc, 1538–50.

Schneiderman, H and Kanade, T (2000) A statistical method for 3D object detection applied to faces and cars. *In*, *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*. IEEE, Vol. 1, 746-51.

Scotland's Parliament (1997) White Paper presented to Parliament by the Secretary of State for Scotland by Command of Her Majesty, Cm 3658. In, Edinburgh.

Skitmore, R M, Stradling, S, Tuohy, A and Mkwezalamba, H (1990) *The Accuracy of construction price forecasts : a study of quantity surveyors' performance in early stage estimating*, Salford, UK: Department of Surveying, University of Salford.

Sonmez, R and Rowings, J (1998) Construction Labor Productivity Modeling with Neural Networks. *Journal of Construction Engineering and Management*, **124**(6), 498-504.

Stadium Database (2014) *How much did Brazil spend on World Cup stadiums?* [http://tinyurl.com/ngkjw9j.] Accessed 15th July 2014

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

Tokede, O, Ahiaga-Dagbui, D D, Smith, S D and Wamuziri, S (2014) Mapping Relational Efficiency in Neuro-Fuzzy Hybrid Cost Models. In: *2014 Construction Research Congress*, Castro-Lacouture, D, Ed., Atlanta, GA, USA: American Society of Civil Engineers (ASCE), 1458-67.

Topcu, I B and Saridemir, M (2008) Prediction of compressive strength of concrete containing fly ash using artificial neural networks and fuzzy logic. *Computational Materials Science*, **41**(3), 305-11.

Vidalis, S and Najafi, F (2002) Cost and time overruns in highway construction. *In*, *4th Transportation Speciality Conference of the Canadian Society for Civil Engineering*, 5-8.

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.

Wang, X Y, Xu, W B, Sun, J and Zhao, Q (2010) Foreign exchange rates forecasting based on VLRBP artificial neural networks. *Computer Engineering and Design*, **31**(1), 167-71.

Wang, Y-R, Yu, C-Y and Chan, H-H (2012) Predicting construction cost and schedule success using artificial neural networks ensemble and support vector machines classification models. *International Journal of Project Management*, **30**(4), 470-8.

Wei, Y and Chen, M C (2012) Forecasting the short-term metro passenger flow with empirical mode decomposition and neural networks. *Transportation Research Part C: Emerging Technologies*, **21**(1), 148-62.

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

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

Yin, R K (2009) *Case study research: Design and methods*. Vol. 5, *Applied Social Research Methods Series*, Thousand Oaks, CA: Sage Publications.

Zandieh, M, Azadeh, A, Hadadi, B and Saberi, M (2009) Application of Artificial Neural Networks for Airline Number of Passenger Estimation in Time Series State. *Applied Sci*, **9**, 1001-13.

Zhang, P, Song, D, Wang, J and Hou, Y (2014) Bias–variance analysis in estimating true query model for information retrieval. *Information Processing & Management*, **50**(1), 199-217.

Zhou, Z-H, Wu, J and Tang, W (2002) Ensembling neural networks: many could be better than all. *Artificial intelligence*, **137**(1), 239-63.

APPENDIX A: PUBLICATIONS

A total of 7 publications resulted from the research reported in this thesis - 2 journal papers and 5 peer-reviewed conference papers. The details of these papers are listed below and full copies appended.

Peer-reviewed journal publications:

1. 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.

2. Ahiaga-Dagbui, D D and Smith, S D (2014) Dealing with construction cost overruns using data mining. *Construction Management & Economics*, **32**(7-8), 628-94.

International peer-reviewed conference papers:

3. Ahiaga-Dagbui, D D and Smith, S D (2013) "My cost runneth over": Data mining to reduce construction cost overruns. In: *Procs 29th Annual ARCOM Conference*, Smith, S D and Ahiaga-Dagbui, D D, Eds.), Reading, UK: Association of Researchers in Construction Management, 559-68.

4. Ahiaga-Dagbui, D D and Smith, S D (2014) Exploring escalation of commitment in construction project management: Case study of the Scottish Parliament project. In: *Procs 30th Annual ARCOM Conference*, Raiden, A, Ed., Portsmouth, UK: Association of Researchers in Construction Management (accepted, in proof).

5. Ahiaga-Dagbui, D D and Smith, S D (2012) Neural networks for modelling the final target cost of water projects. In: *Procs 28th Annual ARCOM Conference*, Smith, S D, Ed., Edinburgh, UK: Association of Researchers in Construction Management, 307-16.

6. Ahiaga-Dagbui, D D, Tokede, O, Smith, S D and Wamuziri, S (2013) A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects. In: *Procs 29th Annual ARCOM Conference*, Smith, S D and Ahiaga-Dagbui, D D, Eds.), Reading, UK: Association of Researchers in Construction Management, 181-90.

7. Tokede, O, Ahiaga-Dagbui, D D, Smith, S D and Wamuziri, S (2014) Mapping Relational Efficiency in Neuro-Fuzzy Hybrid Cost Models. In: *2014 Construction Research Congress*, Castro-Lacouture, D, Ed., Atlanta, GA, USA: American Society of Civil Engineers (ASCE), 1458-67.

Appendix A1

Rethinking construction cost overruns: cognition, learning and estimation

Ahiaga-Dagbui D. Dominic and Simon D. Smith
School of Engineering, University of Edinburgh, Edinburgh, UK

Abstract

Purpose – Drawing on mainstream arguments in the literature, the paper presents a coherent and holistic view on the causes of cost overruns, and the dynamics between cognitive dispositions, learning and estimation. A cost prediction model has also been developed using data mining for estimating final cost of projects. The paper aims to discuss these issues.

Design/methodology/approach – A mixed-method approach was adopted: a qualitative exploration of the causes of cost overrun followed by an empirical development of a final cost model using artificial neural networks.

Findings – A conceptual model to distinguish between the often conflated causes of underestimation and cost overruns on large publicly funded projects. The empirical model developed in this paper achieved an average absolute percentage error of 3.67 percent with 87 percent of the model predictions within a range of ± 5 percent of the actual final cost.

Practical implications – The model developed can be converted to a desktop package for quick cost predictions and the generation of various alternative solutions for a construction project in a sort of what-if analysis for the purposes of comparison. The use of the model could also greatly reduce the time and resources spent on estimation.

Originality/value – A thorough discussion on the dynamics between cognitive dispositions, learning and cost estimation has been presented. It also presents a conceptual model for understanding two often conflated issues of cost overrun and under-estimation.

Keywords Data mining, Prospect theory, Cost overruns, Dunning-Kruger effects, Optimism bias, Referenced class forecasting

Paper type Research paper

Introduction

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). Reliable cost estimates are important for several reasons – for organisational budgeting purposes, for loan application if a project has to be funded through credit facilities, to estimate likely cost of financing loans (interest payments), for estimating commercial feasibility or viability of the project. The present economic conditions also impose a parsimonious approach to spending on most organisations and governments. 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. These include type of project, material costs, likely design and scope changes, ground conditions, duration, size of project, type of client, tendering method and so on (Ahiaga-Dagbui and Smith, 2012). 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 futile; while ignoring them altogether creates a recipe for cost overruns, disputes, law suits and even project

Journal of Financial Management of Property and Construction
Vol. 19 No. 1, 2014
pp. 38-54

1366-4387
DOI 10.1108/JFMPC-06-2013-0027

termination in some cases. There is also a high level of uncertainty around most of these factors at the initial stages of the project as noted by Jennings (2012).

Table I shows major public projects that have experienced significant cost growth. Flyvbjerg *et al.* (2004) report that nine out ten infrastructure projects overrun their budgets and that infrastructure projects have an 86 percent 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; was adjusted to about £9.3 billion in 2007 after significant scope changes; and 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).

Cost overrun in the construction industry has been attributed to a number of sources including technical error in design or estimation, managerial incompetency, risk and uncertainty, suspicions of foul-play, deception and delusion, and even corruption. A recent debate on the Construction Network of Building Researchers (CNBR) on whether or not construction cost overruns could be attributed to error in estimation, or lies by project sponsors and estimators, raised more questions than answers (See the November 2012 CNBR archive online). For instance: how accurate or reliable can cost estimates be? What is the best measure of cost overrun? Might there be need to change how cost performance is presently measured? Should the estimator be absolved of the responsibility of producing reasonably accurate estimates? Should the industry even bother about cost overruns at all, if project goals are met in the long run?

While drawing on the works of some contemporary authorities on the subject, different schools of thought on causes of construction cost overruns have been synthesized in this study, to provide a coherent and holistic view of the problem. Recurring themes have been expanded upon, challenging traditional paradigms of assessing cost performance on construction projects while offering emerging frameworks of reckoning cost growth. It is proposed that there is a conflation of two quite different issues in the understanding of cost growth: cost underestimation and cost overrun. The paper presents a conceptual model for understanding these issues and then presents the development of a validated cost model using data mining and artificial neural networks (ANN). It is hoped that the continuous and effective application of data mining techniques might be one of the possible avenues for alleviating the problem of project cost overruns within the construction.

Project	Estimated cost (in millions)	Final cost (in millions)	% overrun
Sydney Opera House	AUD 7	AUD 102	1,357
Nat West Tower	£15	£115	667
Thames Barrier Project	£23	£461	1,904
Scottish Parliament	£195[a]	£414	112
British Library	£142	£511	260

Notes: [a]September 2000 estimate; initially stated cost was about £40 million
Source: Audit Scotland (2004)

Table I.
Some examples of cost growth in construction projects

Sources of cost growth

Causes of cost growth have been attributed to several sources including improperly managed risk and uncertainty (Okmen and Öztas, 2010), scope creep (Love *et al.*, 2011; Gil and Lundrigan, 2012), optimism bias (Lovallo and Kahneman, 2003; Jennings, 2012) and suspicions of foul-play and corruption (Wachs, 1990; Flyvbjerg, 2009). While not attempting to provide a definitive list of all possible sources, the following section of the paper provides a synthesis of mainstream arguments on the causes of cost growth to provide a holistic view of the subject.

Risk and uncertainty

The nature of a construction project makes it particularly prone to the effects of risk and uncertainty – it is complex and dynamic; each project has many parties with differing business and project objectives; projects are exposed to the weather (not in a controlled environment); and total project duration can spread over several years. It is no surprise then that risk, simply defined here as the measure of exposure to financial loss, or gain (Akintoye, 2000), has been heavily cited as one of the main causes of failure to meet cost targets on construction projects (Skitmore and Ng, 2003; Öztas, 2004; Okmen and Öztas, 2010). Arguably, the construction industry is perhaps one of the most risk prone industries, with project cost being one of main areas susceptible to its effects. Almost all types of risk (including scope changes, inclement weather, unsuitable ground conditions, disputes, client's cash flow problems, etc.) present financial ramifications.

Ahiaga-Dagbui and Smith (2012) noted that effective cost planning relates the design of 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 limit of expenditure. This stage in a project life-cycle is particularly crucial as decisions made during the early stages of the development process carry more far-reaching economic consequences than the relatively limited decisions which can be made later in the process. Despite the importance of cost estimation, it is undeniably not simple, nor straightforward, because of the lack of information in the early stages of the project (Hegazy, 2002). To achieve accuracy, the estimator has to be able to predict the future – something even the best technologies cannot achieve with certainty. This is because accurate reasoning is only possible in a world where information is complete and certain, and where cause and effect links are accurately known. Risk and uncertainty thus deeply pervade the construction industry and continue to cause unending controversy and debate. As Baccarini (2005) suggests, all too often risks are either ignored or dealt with in a completely arbitrary manner using rules-of-thumb or percentages. Flanagan and Norman (1993) also point out that the task of risk management or response in most cases is thus so poorly performed, that far too much risk is passively retained, ultimately resulting in cost escalation during project delivery.

Strategic misrepresentation and optimism bias

Some authorities on the subject of cost overrun have proposed more depressing explanations to the phenomenon. Flyvbjerg *et al.*, suggest that overruns are chiefly due to "strategic misrepresentations", i.e. outright lying (Flyvbjerg *et al.*, 2002) and "optimism bias" (Flyvbjerg, 2007). Flyvbjerg *et al.* compared the cost of projects at the time of the decision to build to the cost at completion and found inaccuracies in cost forecasts for transportation infrastructure projects to be on average 44.7 percent for

rail, 33.8 percent for bridges and tunnels, 20.4 percent for roads – concluding that nine out of ten projects outrun their cost targets. Overruns beyond 100 percent of original cost are also not uncommon (Trost and Oberlander, 2003; Odeck, 2004).

In order to get a project approved, sponsors and estimators, especially on public works, tend to intentionally underestimate the true cost of the project in what has been described as the "Machiavelli factor" (Flyvberg, 2003). "By routinely overestimating benefits and underestimating costs, promoters make their projects look good on paper, which helps get them approved and built" (Flyvbjerg *et al.*, 2005). It makes little reasoning to stop the project once a considerable amount of money has already been spent to get it started, Flyvberg (2004) claims. Wachs (1989) was even more forthright in his paper "When planners lie with numbers" and later advocated for better ethics in forecasting for public works (Wachs, 1990).

If cost overruns cannot be explained by intentional underestimation, optimism bias might be a likely culprit (Flyvbjerg, 2007). Optimism bias can be explained as the cognitive disposition to evaluate future events in a fairer light than they might actually be in reality (Lovallo and Kahneman, 2003). Unlike strategic misrepresentation, this might not be born out of deceptive intent, but also often leads to underestimating true cost, overestimation of benefits, and overlooking the potential of error and uncertainty. The potential gains of the project thus become overwhelmingly enticing, and almost blinding to likely pitfalls. It also leads to underestimating the full extent of certain risk events, should they occur.

In effect, delusion and deception are complementary explanations of the failure of large infrastructure projects, causing works such as diverting existing utilities, environmental impacts and foreseeable risks to be continually underestimated in construction (Flyvbjerg, 2009). This line of diagnosis of the problem of cost overrun might seem appealing, at least on first thought, especially in terms of large capital intensive public projects or those that are likely to make to make high political statements. Flyvbjerg's far-reaching work on cost overruns led to the endorsement of his "reference class forecasting (RCF)" by the American Planning Association in 2005 (APA, 2005; Flyvbjerg, 2007). This will be discussed in more detail in this paper.

Going beyond strategic misrepresentation and optimism bias

Even though deception and delusion might be plausible explanations for cost overruns, particularly in large publicly funded or politically motivated projects, they are not easily generalisable to all types of projects undertaken within the construction industry. Researchers, including Love *et al.* (2012), rebut Flyvbjerg's conclusions as simplistic, largely misleading and not an accurate reflection of reality. Love *et al.*'s rejoinder suggests a move beyond optimism bias and strategic misrepresentation to focus on intermediary events, actions, the so-called "pathogens" that occur between project inception and completion. At the core of Love's argument is that many events and actions that are not accounted for in initial estimates, tend to drive up cost. This school of thought is largely supported by Aibinu and Pasco (2008), Odeck (2004) and Odeyinka *et al* (2012). Love's case study of social infrastructure projects suggest that foul-play, as suggested by Flyvbjerg and Wach, might not be best explanations of cost overruns; and that the fingers point at events that occur before and during the project delivery stage (Love *et al.*, 2011). Besides, it is almost impossible to draw valid distinctions along a continuum of motivation when promoting a project from reasonable optimism, through over-enthusiasm, culpable error, to deliberate deceit using statistical analysis, as adopted in the Flyvbjerg's works.

Research on leadership and governance of construction projects by Gil and Lundrigan (2012), perhaps offers a more holistic assessment of cost growth that aligns closely with the views of Love *et al.* above. That projects evolve, is essentially, the core of their defence. Very often, construction projects change considerably in scope and design between conception, to inception and completion, often due to a client's proposed changes or technically imposed changes. This suggests that it might be erroneous to simply compare the cost of a project at inception, A, with the cost at completion, B, and wherever $B > A$, then overruns have occurred and estimators of A either lied or were incompetent. A and B are essentially very different. More robust explanations of overruns need to factor-in process and product, as well as sources of changes to scope. For Love *et al* (2011) and Gil and Lundrigan (2012) (*op. cit.*), project overruns are not really a case of projects not going according to plan (budget), but the other way round – plans not going according to project.

Gil and Lundrigan (2012) propose a "relay race" framework for understanding cost growth, particularly on mega projects such as the London Olympics project, Scottish Parliament or Terminal 2 project at Heathrow Airport, all of which seemed to have suffered the curse of cost growth, at least on a perfunctory examination. In the relay race of construction delivery, the baton of project leadership is passed on from one person(s) or organisation at the different stages of the project delivery. The aims and scope of the project, as well as skills and competencies of the project sponsors and promoters (project governors) at the conceptual stage, are often very different from their counterparts at the project design or delivery stage. Also, it is not unusual for most public projects to have long gestation periods, stretching over several years before final approval is reached, by which time project budget would also have changed a number of times. The Scottish Parliament Building is a paragon in this respect – the *circa* £40 million submitted by the Scottish Office as likely final cost did not take into consideration project location, or the building of a completely new parliament building. It is no wonder the final cost of the project was ten times this initial proposed cost (Fraser, 2004).

Perception and measuring overruns

Perhaps our perception of cost overruns needs to change altogether. What is described as cost overruns at the moment might not be overruns after all if reckoned through the eyes of different procurement routes, for example. It is possibly one of the reasons why cost overrun is not often reported in projects procured through joint ventures or alliancing. Typically, in traditional contracting, design and estimates are first prepared by the client's estimator (CE) and then bids are invited from contractors. The lowest bidder often wins the job with the lowest tender value becoming the cost estimate at the beginning of the project (A). The contractor undertakes then to deliver the project at cost, A, and all add-ons are dealt with through change orders or claims until project completion at cost, B. Whenever $B > A$, overruns are reported. It is easy to identify how competition, market conditions, optimism bias and the selection by lowest bidder combine to drive down the initial estimate, A, creating a somewhat unrealistic target as likely final cost. For the contractor therefore, winning work at the right price (realistic cost) becomes a very difficult task. To be thorough in estimation would mean including likely cost of most/all risk events in the tender, consequently pricing the contractor out of competition. Most contractors may therefore not include potential risk events in their tenders, so as to increase their likelihood of winning the contract. This was evident in related studies in modelling final cost of construction projects (Ahiaga-Dagbui and Smith, 2012).

Some have suggested that the industry move beyond its fixation on measuring project success largely in terms of cost (Bassioni *et al.*, 2004; Yeung *et al.*, 2008). The CNBR debate was frequently punctuated by the question, "why care about cost overruns anyway? If projects run over budget but deliver what the client wants, should not everyone be happy?" After all, cost overruns only represent our human inability to predict future events accurately, or identify risks and quantify their likely impact and cost. Others think perhaps there is a need for a paradigm shift in how projects are evaluated to cover a combination of social, economic, social, usability or value for money (Toor and Ogunlana, 2010). The Sydney Opera House experienced large overruns at the time of construction but it is now generally considered a twenty-first century icon of buildings and a popular destination for tourists and opera concerts. Similarly, in spite of the controversies about cost overruns, the Scottish Parliament Building has won several awards, including the coveted Stirling Award in 2005 by the Royal Institute of British Architects for its audacious, highly conceptual and iconic design. Even if cost should be a major factor for assessment, it certainly should not be a simplistic or statistical comparison between awarded contract sum and cost at final accounts.

Cognition, bias and learning

Can a science that combines intuition and analysis ever be precise or unbiased? A qualified "no" is probably the answer to that question, according to Kahneman and Tversky (1979), formulators of Prospect Theory – decision making under risk and uncertainty. The theory suggests people make decisions based on the likely gains, or loss, of a venture, and not necessarily based on the real outcome of the decision. It further proposes that decision making is often flawed by systematic biases and that error in judgement is often systematic and predictable, rather than random. Kahneman, a Noble Prize winner for his works on decision making and behavioural economics, delineates decision making and the illusion of understanding, stating that we often exhibit an excessive confidence in what we believe we know about any situation, and that our inability to acknowledge the full extent of our ignorance and the uncertainty of the world we live in makes us prone to overestimate how much we really understand (Kahneman, 2011). Kahneman's work with Lovallo and Kahneman (2003) provides further defence of the Prospect Theory from different business areas. Kahneman's theory holds profound extensions for decision making in the construction industry, especially for large public projects where the effects and cost of risk and uncertainty are particular heightened. It would also provide large support of Flyvbjerg's arguments on strategic misrepresentation and optimism bias already discussed in this paper. Conceivably, this is one reason why it is easy to err on the side of optimism when promoting a project, or when estimating the outcome of a risk event.

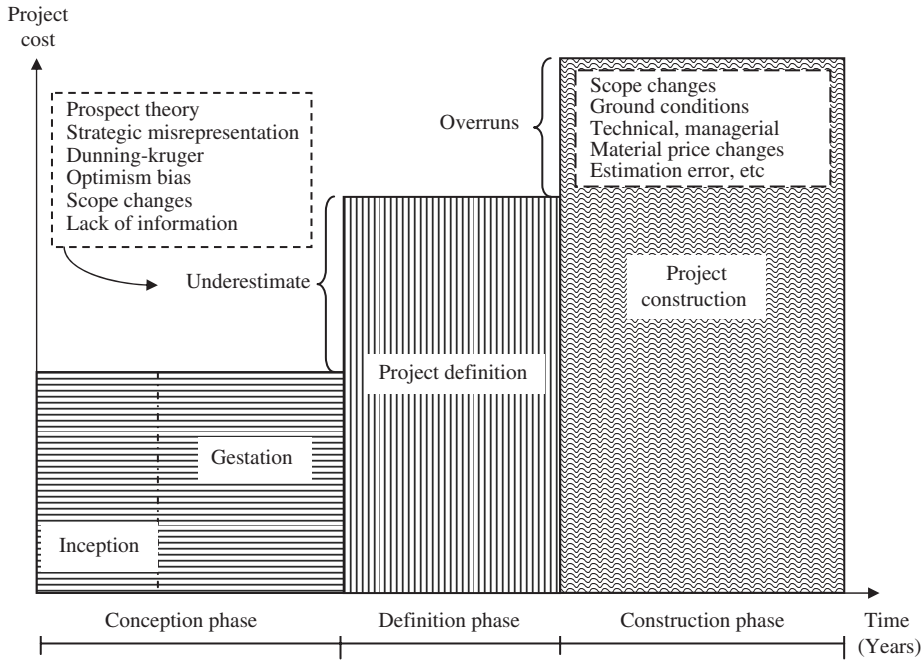
Perhaps even more controversial are the conclusions reached by Kruger and Dunning (2009), that incompetence does not only cause poor performance but also has the dual effect of robbing people of the ability to recognise poor performance. They posit that the metacognitive skills required to judge the accuracy of a decision is the same required to evaluate the error in the same decision – to lack the former, is to fall short in the latter as well (Kruger and Dunning, 1999). The result thereof is that the "incompetent will tend to grossly overestimate their skills and abilities" (Kruger and Dunning, 2009). They tied their conclusion to Darwin's (1871) pronouncement: "ignorance more frequently begets confidence than does knowledge", a theory largely supported by Ehrlinger *et al.* (2008) and Maki *et al.* (1994).

Herein lies the estimation complex – a combination of optimism bias and Prospect Theory predisposes us to underestimate true cost, discounting the real effect of uncertainty and error while doing so. At the same time, Dunning-Kruger tendencies blind forecasters to the error in reaching unrealistic estimates for project cost. Juxtapose these with the effect of risk and uncertainty, competition embedded within the culture of lowest-bidder tendering, as well as strategic misrepresentation, and the overruns reported in Table I become less surprising. It is easier to understand how most cost estimates can be prepared, or at least reported, with an unjustifiable confidence in their accuracy. If this is the case, then perhaps we might not have to move beyond optimism bias just yet, as suggested by Love *et al.* (2011). If we are indeed systematically prone to err towards optimism bias in our reasoning, then it might be wise to rethink how that affects our estimates and what needs to be done about it.

Flyvbjerg (2005) also noted that "no learning" seemed to be taking place in the construction industry over the 70 years prior to his study, and that estimation accuracy has not seen much improvement even with the advancements in technology and the proliferation of cost models and project management approaches. Kruger and Dunning (2009), as well as Ehrlinger *et al.* (2008) attribute lack of performance improvement to the lack of accurate and constructive feedback. They however observed, that an awareness of limitations of skills and decision making within an environment of uncertainty, helped to improve performance and self-calibration. A lack of learning in the construction industry could be explained in a number of ways: that the mitigating factors causing overruns are ones that the industry absolutely cannot overcome and therefore, has to accept cost overruns as normal part of practice; or, that there is simply very little incentive to reach realistic target inception; or further still that the industry seems largely to miss the opportunities offered by effective knowledge transfer and feedback from previously completed projects (Hartmann and Dorée, 2013). How is explicit and tacit knowledge captured and utilised within the industry presently? How do project closure reports feed back into the development of new projects for continuous improvement?

Rethinking overruns

For the purposes of cost modelling or estimation, it is important to clarify an important point. Existing literature, and recent CNBR debate, on "cost overruns" seems to conflate two related, but different issues – overruns and underestimation. Unfortunately, a lot of cost models do not make this distinction either and thus become limited in their application in practice. As already pointed out, most large publicly funded 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 likeness to the defined project. 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; the case of the initial circa £40 million estimate for the Scottish Parliament. For many large publicly funded projects, this is normally when project sponsors garner for project approval and funding. It is perhaps at this stage the effects of Prospect Theory, Dunning-Kruger effect, optimism bias and strategic misrepresentation are particularly heightened, to keep cost at an attractive low and benefits of undertaking the project high. This might be what accounts for what the authors refer to as underestimation of likely cost – the difference between estimated cost at project inception and cost at the end of project definition phase in Figure 1.



Source: Ahiaga-Dagbui and Smith (2013)

Figure 1.
Conceptual model for understanding cost growth on large public projects

Overruns however, are aptly described as the difference in cost at project completion and project definition stage (Figure 1). This is usually as a result of further scope changes, normally not as significant as those at project definition stage, ground conditions, technical and managerial difficulties, material or labour price changes or estimation error. These are the factors that Love *et al.* (2011) describe as "pathogens". So, whereas, Flyvberg's work mainly deals with underestimation, Love's explanations for cost growth largely covers the latter phases of the construction project. It is important to note however that Figure 1 is not necessarily wholly applicable for small, non-political and routine projects where the effects of the political and cognitive causes of cost growth are less heightened. Much of the media hype on cost overruns however is often based on a comparison between cost at inception and cost at completion, almost ignoring the mediating phases of project gestation and definition.

Reference class forecasting

Flyvbjerg developed a practical method for forecasting cost of large projects based on RCF formulated by Kahneman and Tversky (1979) and Kahneman (1994). RCF attempts to use "distributional information" (knowledge) from previous projects similar to the new project being undertaken, the so-called taking of an "outside view" of planned actions, based on actual past performance. Kahneman and Flyvbjerg reckon this approach might somehow help to bypass optimism bias and strategic misrepresentation in decision making (Flyvbjerg, 2007). The methodology involves three steps, summarised simply here as:

(1) identify a reference class of past, similar projects;
(2) estimate a probability distribution for the selected reference class; and
(3) establish likely cost of the new project using the reference class distribution.

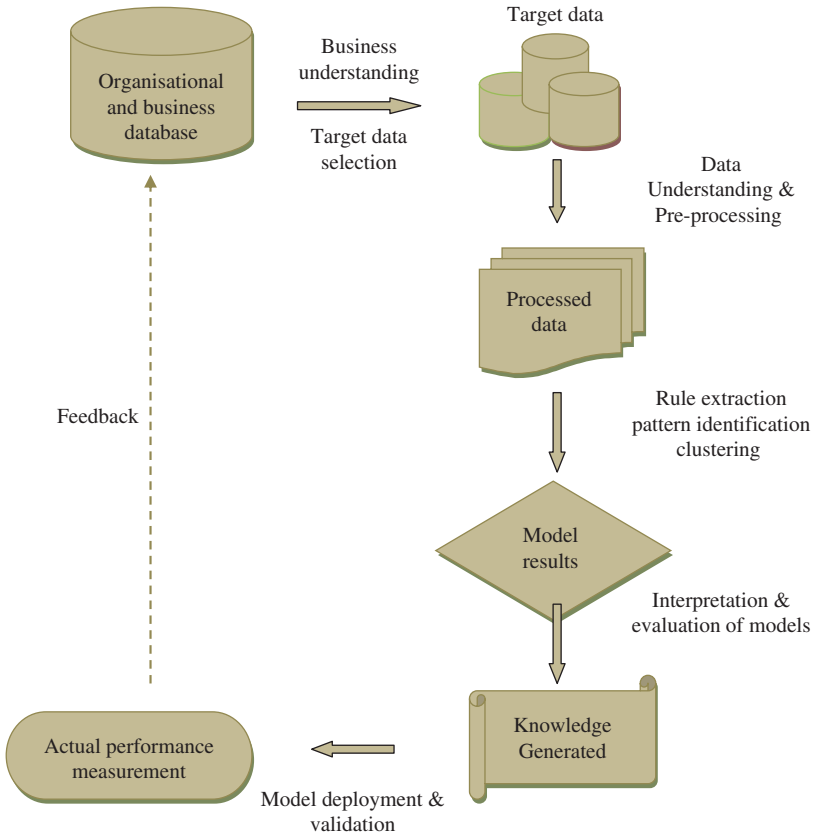
The first instance of its application was on Edinburgh Tram project by the UK Government – the original forecast by the Transport Initiatives Edinburgh (tie), the project promoter was about £255 million but the RCF indicated this could rise up to £400 million and warned that the final cost could even be exceedingly higher (Flyvbjerg, 2007). Recent estimates now indicate that the final construction cost of the Trams could be around £776 million (Miller, 2011; *Railnews*, 2012). The RCF has reportedly been applied to the £15 billion London Crossrail and £7.5 million Taunton Third Way projects in the UK (Flyvbjerg, 2007).

Even though RCF remains to be widely tested or adopted, it might be a step in the right direction especially in dealing with the root causes of underestimation, (as opposed to cost overrun) as shown in Figure 1, i.e. optimism bias, Prospect Theory, Dunning-Kruger effect and strategic misrepresentation. However, as pointed out by Flyvbjerg, RCF is largely applicable to large, non-routine or one-off projects such as stadiums, museums, dams, etc. On smaller, less political, or frequent projects however, a fairly similar but more established method of forecasting that employs previous experience and incremental learning is data mining. This has been extensively used in other industries including finance (Kovalerchuk and Vityaev, 2000), medicine (Bellazzi and Zupan, 2008; Koh and Tan, 2011) and business (Apte *et al.*, 2002), but is yet to see widespread adoption in the construction industry. Notwithstanding, it has been applied to construction knowledge management (Yu and Lin, 2006), for estimating the productivity of construction equipment (Yang *et al.*, 2003), study of occupational injuries (Cheng *et al.*, 2012a), alternative dispute resolution (Fan and Li, 2013) and prediction of the compressive strength high performance concrete (Cheng *et al.*, 2012b). Data mining is used to develop final cost models in the next section of this paper, in a manner that addresses the overruns part of Figure 1.

Final cost model development using data mining

Data mining is the 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. Similar to RCF, data mining 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. Data pre-processing typically involves steps such as sub-sampling, clustering, transformation, de-noising, normalisation or feature extraction (StatSoft Inc., 2011), to ensure that the data are structured and presented to the model in the most suitable way for effective modelling.

The next stage, as shown in Figure 2, 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 results obtained are then evaluated and presented into some meaningful form to aid business decision making. This final step might involve



Source: Ahiaga-Dagbui and Smith (2013)

Figure 2.
The generic data mining process

graphical representation or visualisation of the model for easy communication. ANN is used for the modelling aspect of this study mainly because of its learning and generalisation capabilities (Anderson, 1995).

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. Approximately 1,600 projects completed between 2004 and 2012, with cost range of between £4,000 and £15 million, comprising newly built, upgrade, repair or refurbishment projects. Fifteen project cases were selected using stratified random sampling to be used for independent testing of the final models. The remaining data were then split in an 80:20 percent ratio for training and testing of the models, respectively.

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:

$$zScore = \frac{x_i - \mu}{\sigma} \quad (1)$$

where: *zScore* is the standardized value of a numerical input, x_i; μ is the mean of the numerical predictor; and σ is the standard deviation of the numerical predictor.

This allowed numerical inputs to be squashed into a smaller range of variability, potentially improving the numerical condition of the optimization process of the model (StatSoft Inc., 2008). If one input has a range of 0-1, while another has a range of 0-30 million, as was the case in the data that were used in this analysis, the neural 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 (0,1) coding system. Data screening using scree test, factor analysis and optimal binning allowed for the selection of six initial predictors (primary purpose of project, project scope, project delivery partners, operating region, project duration, and initial estimated cost) to be used for the actual modelling using ANN. Invariant variables, such as payment method, fluctuation measure and type of client, were removed from the variable set as they would only increase the model complexity and yet offer no useful information for model performance.

Model development

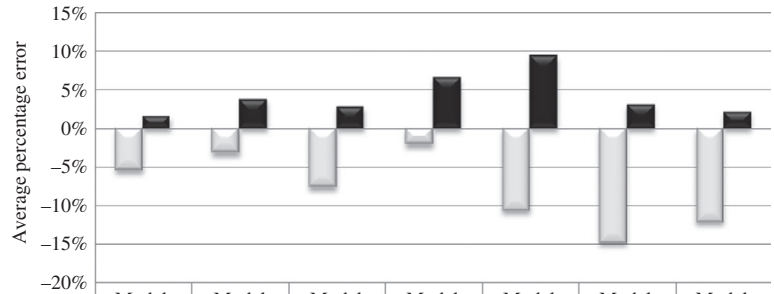
The final model was developed after an iterative process of fine-tuning the network parameters and/or inputs until acceptable error levels were achieved or when the model showed no further improvement. First, the automatic network search function of Statistica 10® software was used to optimise the search for the best network parameters, after which customized networks were developed using the optimal parameters identified. Five activation functions[1] were iterated in both hidden and output layers, using gradient descent, conjugate descent and Quasi-Newton (BFGS) training algorithms. About 2,000 multilayer perceptron networks were trained at each iteration stage, retaining the five best before further tweaking to investigate possible model improvement.

Early stopping, the process of halting training when the test error stops decreasing, was used to prevent memorising or over-fitting the dataset in order to improve generalization. Over-fitted models perform very well on training and testing data, but fail to generalise satisfactorily when new "unseen" cases are used to validate their performance. The best networks at each stage were selected based on their overall performance, 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:

$$SOS = \sum (T_i - O_i)^2 \quad (2)$$

where: O_i is the predicted final cost of the ith data case (output); and 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. Overall, about 30 networks were retained, which were then validated using the 15 separate projects that were selected using stratified sampling at the beginning of the modelling exercise. Figure 3 shows the performance of the best seven out the 30 validated models.



	Model 33 %Error	Model 37 %Error	Model 40 %Error	Model 46 %Error	Model 50 %Error	Model 61 %Error	Model 70 %Error
Average under estimation	–5%	–3%	–8%	–2%	–11%	–15%	–12%
Average over estimation	2%	4%	3%	7%	99%	3%	2%

Figure 3.
Performance of retained models

Table II shows the performance of overall best model (model 33). It compares the final cost forecasts reached by the model with the actual final cost recorded at the end of the project. This model was an MLP 8-11-1, i.e. a multilayer perceptron with eight nodes in the input layer, 11 hidden units and one output (final cost). It was trained with a Quasi-Newton (BFGS) training algorithms and had a hyperbolic tangent (tanH) activation function in both hidden and output layers. The tanh activation function, defined in equation (3), squashes continuous variables into a range of $(-1, +1)$ for more effective training of the neural network models:

$$f(x) = \frac{sin\, h\, x}{\cos h\, x} = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} = \frac{e^{2x} - 1}{e^{2x} + 1} \tag{3}$$

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

Table II.
Validation results of the best model (model 33)

The final predictors in this model were the purpose of the project, the construction delivery partner used by the client, the estimated duration, an early scheme estimate of final cost and scope of the project. The average APE achieved by this model was 3.67 percent across the 15 validation cases. Its APEs ranged between 0.04 percent and 15.85 percent. It was observed that the worst performances of the model were achieved on projects with the smallest values in the validation set (cases 13 and 15). This might be because a majority of the projects used for the model training had values in excess of £5 million. However, the actual monetary errors on these predictions were deemed satisfactory as they were relatively small (about £3,500 and £2,500 for models 13 and 15, respectively). 87 percent of the validation predictions of the best model were within ± 5 percent of the actual cost of the project. 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.

Conclusion

Cost estimate reliability and accuracy on construction projects continues to receive a lot of attention from both industry and academia. The industry is faced with a complex web of causes, which we propose fall into two distinct yet often conflated categories – cost underestimation and cost overrun summarised as follows.

Underestimation

- optimism bias – a propensity to believe and act on a notion that all will go well leading to the underestimation the role of uncertainty in outcomes;
- prospect theory – making decisions based on likely gains and loss rather than the actual outcome of the decision;
- strategic misrepresentation – outright lying and corruption; and
- Dunning-Kruger effect – the bend to overestimate competency or accuracy in judgement and the inability to see past our own errors; competition to win projects.

Overrun

- scope changes, whether mandated by circumstances or requested by client;
- managerial and technical difficulties;
- risk and uncertainty; and
- ground conditions, price changes (etc.).

Most of these, especially the cognitive and psychological factors, tend to work together to drive down the true cost of the project during the initial stages, creating a false and unreliable estimate as target to reach. We have attempted to provide a holistic view of the problem of cost growth, while presenting a conceptual model to distinguish between these often conflated ideas of underestimation and overruns on construction projects. RCF was discussed as a possible means of addressing underestimation, particularly on large publicly funded projects. The development of a final cost prediction model using data mining and ANN was then presented as a possible avenue of addressing cost overruns in the construction industry. The best model achieved an

average absolute percentage error of 3.67 percent with 87 percent of the validation predictions falling within an error range of ± 5 percent. These methods can be used to develop decision support systems especially at early stages of the construction project as well as complement traditional methods of estimation in order to reach more accurate and reliable cost estimates.

Clients can play a crucial role in ensuring the quality and reliability of cost estimates in the construction industry. As indicated by the commercial manager of one the biggest construction companies in the UK, "winning a tender is easy. But winning at the right price is difficult". Unless clients start demanding realistic estimates, rather than the lowest estimates at the early stages of a project, the problem of cost overrun might remain with the industry for a long time to come. Cultural changes within the industry towards the search for realistic targets might incentivise contractors to flag up potential estimating pitfalls early-on. Questions about who has the responsibility on behalf of the client to govern the project always has profound implications on cost growth from inception to completion and needs to be addressed very early on a project. This is particularly important on mega projects.

Project knowledge capture and its utilisation would also be crucial in tackling cost overruns. Some data mining techniques like neural networks are particularly useful in modelling both explicit and tacit knowledge within extensive databases. This can be used to complement traditional cost estimation methods or RFC to reach more realistic and reliable estimates. Finally, and perhaps more importantly, is the creation of a culture of critical questioning, measures of accountability, with checks and balances to make sure that cost is managed to be within reasonable budget limits.

Note

1. Identity, logistic, tanh, exponential and sine activation functions.

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, Association of Researchers in Construction Management, Edinburgh, 3-5 September*, pp. 307-316.

Ahiaga-Dagbui, D.D. and Smith, S.D. (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*, Association of Researchers in Construction Management, Reading, pp. 559-568.

Aibinu, A.A. and Pasco, T. (2008), "The accuracy of pre-tender building cost estimates in Australia", *Construction Management and Economics*, Vol. 26 No. 12, pp. 1257-1269.

Akintoye, A. (2000), "Analysis of factors influencing project cost estimating practice", *Construction Management & Economics*, Vol. 18 No. 1, pp. 77-89.

Anderson, J.A. (1995), *An Introduction to Neural Networks*, MIT Press, Cambridge, MA.

APA (2005), *JAPA Article Calls on Planners to Help End Inaccuracies in Public Project Revenue Forecasting*, American Planning Association (APA), available at: http://goo.gl/I7DLA

Apte, C., Liu, B., Pednault, E.P.D. and Smyth, P. (2002), "Business applications of data mining", *Communications of the ACM*, Vol. 45 No. 8, pp. 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*, Vol. 17 No. 6, pp. 337-342.

Baccarini, D. (2005), "Estimating project cost contingency – beyond the 10% syndrome", *2005 Australian Institute of Project Management Conference*, AIPM, available at: www.aipm.com.au/resource/Baccarini-AIPMconf05.pdf (accessed 12 August 2011).

Bassioni, H., Price, A. and Hassan, T. (2004), "Performance measurement in construction", *Journal of Management in Engineering*, Vol. 20 No. 2, pp. 42-50.

BCIS (2012), *BIS Construction Price and Cost Indices*, Building Cost Information Services, available at: www.bcis.co.uk

Bellazzi, R. and Zupan, B. (2008), "Predictive data mining in clinical medicine: current issues and guidelines", *International Journal of Medical Informatics*, Vol. 77 No. 2, pp. 81-97.

Chan, A.P. and Chan, A.P. (2004), "Key performance indicators for measuring construction success", *Benchmarking: An International Journal*, Vol. 11 No. 2, pp. 203-221.

Cheng, C.-W., Leu, S.-S., Cheng, Y.-M., Wu, T.-C. and Lin, C.-C. (2012a), "Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry", *Accident Analysis & Prevention*, Vol. 48, pp. 214-222.

Cheng, M.-Y., Chou, J.-S., Roy, A.F.V. and Wu, Y.-W. (2012b), "High-performance concrete compressive strength prediction using time-weighted evolutionary fuzzy support vector machines inference model", *Automation in Construction*, Vol. 28, pp. 106-115.

Darwin, C. (1871), *The Decent of Man*, John Murray, London.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D. and Kruger, J. (2008), "Why the unskilled are unaware: further explorations of (absent) self-insight among the incompetent", *Organizational Behavior and Human Decision Processes*, Vol. 105 No. 1, pp. 98-121.

Fan, H. and Li, H. (2013), "Retrieving similar cases for alternative dispute resolution in construction accidents using text mining techniques", *Automation in Construction*, Vol. 34, pp. 85-91.

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

Flyvberg, B. (2003), "Machiavellian tunnelling", *World Tunnelling*, p. 43.

Flyvbjerg, B. (2005), "Design by deception: the politics of megaproject approval", *Harvard Design Magazine*, Vol. 22, pp. 50-59.

Flyvbjerg, B. (2007), "Curbing optimism bias and strategic misrepresentation in planning: reference class forecasting in practice", *European Planning Studies*, Vol. 16 No. 1, pp. 3-21.

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*, Vol. 25 No. 3, pp. 344-367.

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*, Vol. 68, pp. 279-295.

Flyvbjerg, B., Holm, M.K.S. and Buhl, S.L. (2004), "What causes cost overrun in transport infrastructure projects?", *Transport Reviews*, Vol. 24 No. 1, pp. 3-18.

Flyvbjerg, B., Skamris Holm, M.K. and Buhl, S.L. (2005), "How (in)accurate are demand forecasts in public works projects? The case of transportation", *Journal of the American Planning Association*, Vol. 71 No. 2, pp. 131-146.

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, Scottish Parliamentary Corporate Body, available at: www.holyroodinquiry.org/

Gidson, O. (2012), "London 2012 Olympics will cost a total of £8.921bn", *The Guardian*, 23 October, available at: http://goo.gl/sxatK (accessed 22 April 2013).

Gil, N. and Lundrigan, C. (2012), *The Leadership and Governance of Megaprojects*, Centre for Infrastructure Development (CID), Manchester Business School, The University of Manchester, available at: http://goo.gl/cf2ST

Hartmann, A. and Dorée, A. (2013), "Messages in bottles: the fallacy of transferring knowledge between construction projects", in Smith, S.D. and Ahiaga-Dagbui, D.D. (Eds), *Procs 29th Annual ARCOM Conference, Association of Researchers in Construction Management, Reading, 2-4 September* (in press).

Hegazy, T. (2002), *Computer-Based Construction Project Management*, Prentice-Hall, Upper Saddle River, NJ.

Jennings, W. (2012), "Why costs overrun: risk, optimism and uncertainty in budgeting for the London 2012 Olympic games", *Construction Management and Economics*, Vol. 30 No. 6, pp. 455-462.

Kahneman, D. (1994), "New challenges to the rationality assumption", *Journal of Institutional and Theoretical Economics*, Vol. 150, pp. 18-36.

Kahneman, D. (2011), *Thinking, Fast and Slow*, Farrar, Straus and Giroux, New York, NY.

Kahneman, D. and Tversky, A. (1979), "Prospect theory: an analysis of decision under risk", *Econometrica*, Vol. 47 No. 2, pp. 263-291.

Koh, H.C. and Tan, G. (2011), "Data mining applications in healthcare", *Journal of Healthcare Information Management*, Vol. 19 No. 2, p. 65.

Kovalerchuk, B. and Vityaev, E. (2000), *Data Mining in Finance*, Kluwer Academic Publisher, Hingham, MA.

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*, Vol. 77 No. 6, pp. 1121-1134.

Kruger, J. and Dunning, D. (2009), "Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments", *Psychology*, Vol. 1, pp. 30-46.

Lovallo, D. and Kahneman, D. (2003), "Delusions of success: how optimism undermines executives' decisions", *Harvard Business Review*, July.

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", *IEEE Transactions on Engineering Management*, Vol. 59 No. 4, pp. 560-571.

Love, P.E.D., Davis, P., Ellis, J. and Cheung, S.O. (2010), "Dispute causation: identification of pathogenic influences in construction", *Engineering, Construction and Architectural Management*, Vol. 17 No. 4, pp. 404-423.

Love, P.E.D., Sing, C.-P., Wang, X., Irani, Z. and Thwala, D.W. (2012), "Overruns in transportation infrastructure projects", *Structure and Infrastructure Engineering*, pp. 1-19.

Maki, R.H., Jonas, D. and Kallod, M. (1994), "The relationship between comprehension and metacomprehension ability", *Psychonomic Bulletin & Review*, Vol. 1 No. 1, pp. 126-129.

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

NAO (2012), "The London 2012 Olympic games and paralympic games: post-games review", paper presented at 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*, Vol. 36 No. 2, pp. 2592-2602.

Odeck, J. (2004), "Cost overruns in road construction – what are their sizes and determinants?", *Transport Policy*, Vol. 11 No. 1, pp. 43-53.

Odeyinka, H., Larkin, K., Weatherup, R., Cunningham, G., McKane, M. and Bogle, G. (2012), *Modelling Risk Impacts on the Variability Between Contract Sum and Final Account* (*A Research Report Submitted to RICS)*, Royal Institution of Chartered Surveyors, London.

Okmen, O. and Öztas, A. (2010), "Construction cost analysis under uncertainty with correlated cost risk analysis model", *Construction Management and Economics*, Vol. 28 No. 2, pp. 203-212.

Öztas, A. (2004), "Risk analysis in fixed-price design-build construction projects", *Building and Environment*, Vol. 39 No. 2, pp. 229-237.

Railnews (2012), "Edinburgh Tram costs soar again", *Railnews*, available at: http://goo.gl/M5uZ7 (accessed 18 February 2013).

Skitmore, M.R. and Ng, T.S. (2003), "Forecast models for actual construction time and cost", *Building and Environment*, Vol. 38 No. 8, pp. 1075-1083.

StatSoft Inc. (2008), *A Short Course in Data Mining*, StatSoft, Inc., available at: www.statsoft.com/Portals/0/Products/Data-Mining/data_mining_tutorial.pdf (accessed 25 January 2012).

StatSoft Inc. (2011), *STATISTICA 10* (*Data Analysis Software System*), available at: www.statsoft.com

Toor, S.-U.-R. and Ogunlana, S.O. (2010), "Beyond the 'iron triangle': stakeholder perception of key performance indicators (KPIs) for large-scale public sector development projects", *International Journal of Project Management*, Vol. 28 No. 3, pp. 228-236.

Trost, S.M. and Oberlander, G. (2003), "Predicting accuracy of early cost estimates using factor analysis and multivariate regression", *Journal of Construction Engineering and Management*, Vol. 129 No. 2.

Wachs, M. (1989), "When planners lie with numbers", *Journal of the American Planning Association*, Vol. 55 No. 4, pp. 476-479.

Wachs, M. (1990), "Ethics and advocacy in forecasting for public policy", *Business and Professional Ethics Journal*, Vol. 9 Nos 1/2, pp. 141-157.

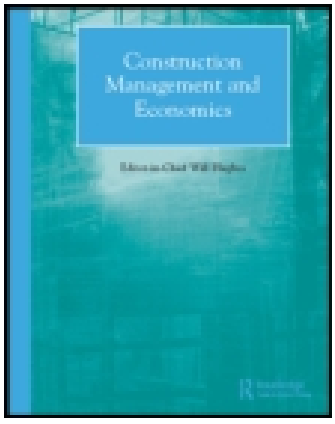
Yang, J., Edwards, D.J. and Love, P.E.D. (2003), "A computational intelligent fuzzy model approach for excavator cycle time simulation", *Automation in Construction*, Vol. 12 No. 6, pp. 725-735.

Yeung, J.F., Chan, A.P. and Chan, D.W. (2008), "Establishing quantitative indicators for measuring the partnering performance of construction projects in Hong Kong", *Construction Management and Economics*, Vol. 26 No. 3, pp. 277-301.

Yu, W.-D. and Lin, H.-W. (2006), "A VaFALCON neuro-fuzzy system for mining of incomplete construction databases", *Automation in Construction*, Vol. 15 No. 1, pp. 20-32.

Corresponding author
Ahiaga-Dagbui D. Dominic can be contacted at: D.Ahiaga-Dagbui@ed.ac.uk

Appendix A2



Construction Management and Economics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/rcme20

Dealing with construction cost overruns using data mining

Dominic D. Ahiaga-Dagbui[a] & Simon D. Smith[a]
[a] School of Engineering, University of Edinburgh, Edinburgh EH9 3JL, UK
Published online: 24 Jul 2014.

To cite this article: Dominic D. Ahiaga-Dagbui & Simon D. Smith (2014) Dealing with construction cost overruns using data mining, Construction Management and Economics, 32:7-8, 682-694, DOI: 10.1080/01446193.2014.933854

To link to this article: http://dx.doi.org/10.1080/01446193.2014.933854

Construction Management and Economics, 2014
Vol. 32, Nos. 7–8, 682–694, http://dx.doi.org/10.1080/01446193.2014.933854

Routledge
Taylor & Francis Group

Dealing with construction cost overruns using data mining

DOMINIC D. AHIAGA-DAGBUI* and SIMON D. SMITH

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

Received 4 November 2013; accepted 9 June 2014

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 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 non-parametric bootstrapping and ensemble modelling in artificial neural networks, final project cost-forecasting models were developed with 1600 completed projects. This helped to extract information embedded in data on completed construction projects, in an attempt to address the problem of the dearth of information in the early stages of a project. It was found that 92% of the 100 validation predictions were within ±10% of the actual final cost of the project while 77% were within ±5% of actual final cost. This indicates the model's ability to generalize 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

In a construction project, the main obligations of a project team towards their client are usually reduced to concerns around functional requirements, specific quality, and delivery within an acceptable budget and time frame. Usually for most clients, the cost aspect of these requirements seems to rank highest. Thus, the estimates prepared at the initial stages of a project can play several important 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 more far-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, Nicholas (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.*, 2012) 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

*Author for correspondence. E-mail: D.Ahiaga-Dagbui@ed.ac.uk

these firms presents both a challenge and an opportunity: a challenge to traditional methods of data analysis since the data are often complex, and usually, 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 policymaking. Advances in the fields of data warehousing, artificial intelligence, statistics, visualization techniques and machine learning now make it possible for data to be transformed into a 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 extensive application in industries such as business (cf. Apte *et al.*, 2002) and medicine (cf. Koh and Tan, 2005). 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 the 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 (ANNs). 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 of 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 (Ahiaga-Dagbui and Smith, 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 can be an exhausting task, if not altogether futile. Ignoring most of these factors also creates a recipe for possible cost growth, disputes, lawsuits 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 Auditor General of Western Australia assessed the management and performance of 20 capital-intensive non-residential projects including sports venues, schools and hospitals, undertaken within the state. The expected cost of all the projects at the time was A$6.157 billion, a staggering $3.275 billion (114%) more than the total original approved budget estimates. Fifteen of the 20 projects were expected to exceed their original approved budgets, of which four were expected to exceed their budgets by more than 200% (Auditor General of Western Australia, 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 City of Boston's Central Artery project (popularly referred

to as the Big Dig) was to cost US$2.6 billion but was completed at US$14.8 billion and seven years late in 2006 (Gelinas, 2007). 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 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.

Sources of overrun

Cost overrun 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 because of doubts as to their suitability and lack of knowledge and understanding of these methods. 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 inclement weather 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), as well as Wachs (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 categorized the sources of cost overruns on construction projects into four groups: technical (error), psychological, economic and political. They concluded that cost escalation could not be adequately explained by estimation error, but was more likely caused by strategic misrepresentation: an intentional attempt to mislead. They observed that nine out of 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.* (2002, p. 279) 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 chains 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. Flyvberg makes this distinction between the two concepts with the terms 'deception' and 'delusion' respectively (Flyvbjerg, 2008). It is plausible to reckon how strategic misrepresentation and optimism bias might 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.

The proponents of another school of thought on cost overruns, referred to as the 'evolution theorists', include Love *et al.* (2012) as well as Gil and Lundrigan (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.* (2012, p. 560) provide a counter-perspective to the delusion and deception 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 generalizable to all projects undertaken within the industry. Their argument would seem sustainable, especially in respect of 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 overconfidence and delusion, culpable error, to deliberate deceit using just statistical analysis, the method adopted in Flyvbjerg's works.

Love *et al.* (2005) also conducted a questionnaire survey of 161 construction professionals in the Australian construction industry and found that rework was one of the main contributors to escalation of cost. The main sources of rework as found in their work are ineffective use of information technology, staff turnover/allocation to other projects, incomplete design at the time of tender, insufficient time to prepare contract documentation and poor coordination between design team members. This conclusion is similar to that reached by Bordat *et al.* (2004) who found that the 'dominant' source of cost overrun was change orders due mainly to 'errors and omissions' in design. In a more recent research, Love *et al.* (2014) challenged the strategic misrepresentation and optimism bias perspective by Flyvbjerg (2008) as lacking in verifiable causality, and therefore limited in their application.

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).

Measuring overruns

It may be important to note here that much of the current literature and media furore on cost overrun seems to oversimplify its rather complex causes. 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 defined project, as was the case of the New Children Hospital in Western Australia (Auditor General of Western Australia, 2012). The initially approved budget for the hospital was A$207 million. The scope at this stage was to relocate the Princess Margaret Hospital to the Royal Perth Hospital. However, this scope completely changed during project definition to the construction of a totally new Medical Center at A$962 million, a cost increase of A$755 if taken on cursory examination. The Holyrood Project in Edinburgh also experienced a similar significant scope change, and thereby the astonishing cost growth recorded (see Audit Scotland, 2000, 2004). It seems erroneous, therefore, to make a direct comparison between the initial 'estimate' A and its final completion cost B: the two schemes are usually very different. 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.* (2012, 2014), but probably justifiable as estimate A is usually the estimate used to get project approval when publicly funded projects are being appraised. As it is often practically difficult to discontinue a project once a considerable amount of money has already been spent to get it started, it may thus be crucial for the industry to find more effective ways of project approval that deal better with underestimation of true costs and the setting of unrealistic cost targets.

Going forward: estimating final cost

Alex *et al.* (2010) reviewed the cost performance on more than 800 construction projects of 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, equipment and materials, and making allowance for profits and overheads for individual construction items, is deterministic by nature. It therefore largely neglects and deals poorly with uncertainties and their correlation effects on cost, and is thereby

deemed inadequate in reaching a reliable and realistic final cost. As an alternative to traditional estimation approaches, data mining, using the learning and generalization 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 for 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 then validating 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, job title and account balance, Huang *et al.* (2007) developed a support vector machine credit scoring model to assess loan applicants' creditworthiness in an attempt to limit a financing firm's exposure to 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, classifying these applications into four groups of customer identification, attraction, retention and development. One-to-one marketing and loyalty programmes 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 reuses knowledge from previous value engineering exercises in an attempt to streamline future exercises, making them more systematic, organized and problem-focused. 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 in 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 association rule-based data mining to identify the characteristics of occupational injuries reported between 1999 and 2004 in the construction industry of Taiwan. 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 organization and business transactions of the firm. The selected dataset 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 offering 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, normalization 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', regression models generate an

Table 1 Framework for selecting a data mining technique

Data mining category	Data mining requirement	Data mining technique	Technique characteristics
Regression	Prediction	Regression	Flexibility
Clustering	Pattern discovery	Support vector machine (SVM)	Accuracy (precision)
Classification	Surveillance	Self-organizing maps	Power
Visualization	Performance	Genetic algorithm, etc.	'Interpretability'
Summarization	Measurement		Ease of deployment
	Business		
	Understanding		

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.

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 twofold: to develop decision support systems from existing data to complement the existing estimation process within our collaborating organization 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 organization 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 allow the organization 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 organization. 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 for the modelling were 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 upgrades. Project values ranged from a mere £1000 to £30 million and durations from three months to five 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 two 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 fairly 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 were 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 are found in Table 2.

Table 2 Overview of data used for model development

Size	Types of project	Type of organization	Cost range	Duration range	Year span
*c.*1600	Water mains, manholes, combined sewer overflows, repairs, upgrades	Client	£4000 to £15 million	1 month to 5 years	2004–12

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 offering the modeller the chance to get to understand the data thoroughly. Cost values were normalized to a 2012 baseline using the infrastructure resources cost indices by the Building Cost Information Services (2012) with a base year 2000. This allowed for cost values to be quite comparable across different years. Numerical predictors were further standardized to *zScores* using

$$zscore = \frac{x_i - \mu}{\sigma} \quad (1)$$

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 were 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 while 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.

Cost model development

Data visualization 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 were categorical, rather than the usual numeric type. It was thus decided to use artificial neural networks (ANNs) for the actual modelling because of their ability to cope with non-linear relationships and categorical variables (cf. Anderson, 1995). An artificial neural network is an abstraction of the human brain with abilities to learn from experience and generalize 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 other thereby resulting in spurious predictions when both of those variables are included in the model (Marsh *et al.*, 2004). Neural networks have 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 tried at this stage. RBF models the relationship between inputs and targets in two phases: it first performs a probability distribution of the inputs before 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 3–10 nodes.

Consequently, using a custom range of 3–10 hidden nodes in one hidden layer, a dataset size split of 70% for training, 15% for testing and another 15% for first

Table 3 Initial list of variables for model development

	Type of data	Category 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

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

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 mean squared errors (MSE). MSE is defined here as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(O_{i-T_i})^2 \quad (2)$$

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) and n is the sample size.

The higher the MSE value, the poorer the network at generalization, 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 the Broyden-Fletcher-Goldfarb-Shanno (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 memorizing or over-fitting the dataset in order to improve generalization. Over-fitted models perform very well on training and testing data, but fail to generalize 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, bivariate 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 projects 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 from five input factors. These five significant input factors are purpose of project (wastewater, water or general), scope of project (new-build, upgrade or replacement), ground condition (contaminated or non-contaminated ground), delivery partner (anonymized as X, Y, Z) and estimated project duration. The models also had between three and seven nodes in a single hidden layer with one output, i.e. final cost. They were trained with a tanh or logistic activation function between their input and hidden layers, and 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, non-parametric 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 get used for training, testing or validation, and thus 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 bootstrap sample size without specific guidance on an optimum size. Efron and Tibshirani (1993), as well as Pattengale *et al.* (2009), however, suggest that a 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 1 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. While 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

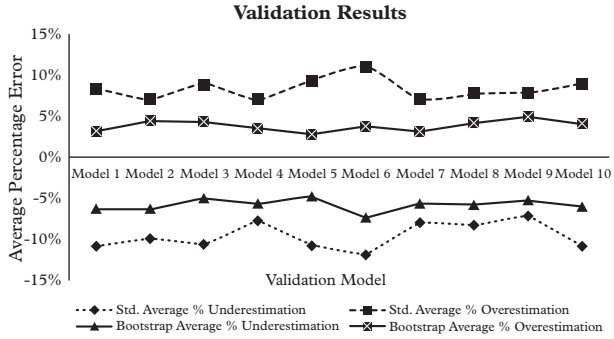


Figure 1 Validation results (standard models vs bootstrapping)

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.

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,

Table 4 Summary of results (standard, bootstrapped and ensemble models)

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

Table 5 Sample results from ensemble model validation

Case	Actual final cost (000)	Ensemble prediction (000)	Ensemble error (000)	Ensemble absolute % error
1	4846	4990	(144)	2.97
2	1586	1590	(4)	0.25
3	24 986	23 760	1226	4.91
4	11 143	10 934	209	1.88
5	5328	5765	(437)	8.20
6	3787	3723	64	1.69
7	17 346	16 967	379	2.18
8	4136	4033	103	2.49
9	3117	2994	123	3.95
10	1000	939	61	6.10
11	1773	1674	99	5.58
12	3779	3600	179	4.74
13	209	192	17	8.13
14	3960	3810	150	3.79
15	294	300	(6)	2.04
16	2296	2220	76	3.31
17	2104	2038	66	3.14
18	248	247	1	0.40
19	208	192	16	7.69
20	201	197	4	1.99

Table 6 Summary of validation performance of ensemble model

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

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.

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 6 shows a summary the performance of the ensemble model for all the 100 validation cases. Note that 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. Only eight out of the 100 validations had predictions beyond ±10% of the final cost of the project case.

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 retrievable manner for a number of projects over time, a vast database of potentially valuable assets 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 reducing 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 of 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 non-parametric bootstrapping and ensemble modelling in artificial neural networks, cost models were developed to estimate the final construction cost of water infrastructure projects. It was found that 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 will be particularly useful at the pre-contract stage of the partnering construction firm that participated 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 these are relevant business and operational data to solve the problem at hand.

References

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. (2012) Neural networks for modelling the final target cost of water projects, in Smith, S.D. (ed.) *Proceedings 28th Annual ARCOM Conference*, Association of Researchers in Construction Management, Edingburgh, UK, pp. 307–16.

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 and 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**(7), 745–56.

Anderson, J.A. (1995) *An Introduction to Neural Networks*, MIT Press, Cambridge, MA.

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.

Audit Scotland (2000) *The New Scottish Parliament Building: An Examination of the Management of the Holyrood Project*, Audit Scotland, Edinburgh, UK.

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

Auditor General of Western Australia (2012) *Managing Capital Projects*, Office of the Auditor General of Western Australia, Perth, Australia, available at http://tinyurl.com/l9ymlqu (accessed May 2014).

Bordat, C., McCullouch, B.G., Sinha, K.C. and Labi, S. (2004) *An Analysis of Cost Overruns and Time Delays of INDOT Projects*, Publication FHWA/IN/JTRP-2004/07, Joint Transportation Research Program, Indiana Department of Transportation and Purdue University, West Lafayette, IN.

Breiman, L. (1996) Bagging predictors. *Machine Learning*, **24**(2), 123–40.

Building Cost Information Services (2012) *BIS Construction Price and Cost Indices*, available at www.bcis.co.uk (accessed November 2012).

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**, 106–15.

Efron, B. (1992) Bootstrap methods: another look at the jackknife, in Kotz, S. and Johnson, N.L. (eds) *Breakthroughs in Statistics*, Springer, pp. 569–93.

Efron, B. and Tibshirani, R. (1993) *An Introduction to the Bootstrap*, Chapman and Hall, New York.

Egan, J. (1998) *Rethinking Construction*, Construction Task Force Report for the Department of the Environment, Transport and the Regions, HMSO, London.

Elhag, T.M.S. and Boussabaine, A.H. (1998) An artificial neural system for cost estimation of construction projects, in Hughes, W. (ed.) *Proceedings 14th Annual ARCOM Conference*, University of Reading, 9–11 September, Association of Researchers in Construction Management, Reading, pp. 219–26.

Emsley, M.W., Lowe, D.J., Duff, A., Harding, A. and Hickson, A. (2002) Data modelling and the application of a neural network approach to the prediction of total construction costs. *Construction Management and Economics*, **20**(6), 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*, Blackwell Science, Oxford.

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

Flyvbjerg, B. (2008) 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) Underestimating costs in public works projects: error or lie? *Journal of the American Planning Association*, **68**(3), 279–95.

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.

Gelinas, N. (2007) Lessons of Boston's Big Dig. *City Journal*, Autumn, available at http://tinyurl.com/dxxrdf (accessed 8 May 2014).

General Accounting Office (1997) *Transportation Infrastructure: Managing the Costs of Large-Dollar Highway Projects*, United States General Accounting Office (GAO), Washington DC.

Gil, N. and Lundrigan, C. (2012) *The Leadership and Governance of Megaprojects*, CID Technical Report No. 3/2012, Centre for Infrastructure Development (CID), Manchester Business School, The University of Manchester.

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*, Springer, Berlin, pp. 283–302.

Gurney, K. (1997) *An Introduction to Neural Networks*, UCL Press, London.

Hair, J., Tatham, R., Anderson, R. and Black, W. (1998) *Multivariate Data Analysis*, 5th edn, Prentice Hall, Upper Saddle River, NJ, USA.

Hegazy, T. (2002) *Computer-Based Construction Project Management*, Prentice Hall, Upper Saddle River, NJ.

Hoffman, P., Grinstein, G., Marx, K., Grosse, I. and Stanley, E. (1997) DNA visual and analytic data mining, in *Visualization '97, Proceedings*, Phoenix, AZ, 24–24 October, pp. 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 edn, John Wiley & Sons, Oxford.

Koh, H.C. and Tan, G. (2005) Data mining applications in healthcare. *Journal of Healthcare Information Management*, 19(2), 64–72.

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 Smith, J. (2005) Contract documentation and the incidence of rework in projects. *Architectural Engineering and Design Management*, **1**(4), 247–59.

Love, P.E.D., Edwards, D.J. and Irani, Z. (2012) Moving beyond optimism bias and strategic misrepresentation: an explanation for social infrastructure project cost overruns. *IEEE Transactions on Engineering Management*, 59(4), 560–71.

Love, P.E.D., Smith, J., Simpson, I., Regan, M., Sutrisna, M. and Olatunji, O. (2014) Understanding the landscape of overruns in transport infrastructure projects. *Environment and Planning B: Planning and Design* (forthcoming).

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), 518–22.

McKim, R.A. (1993) Neural networks and the identification and estimation of risk, *Transaction of the 37th Annual Meeting of the American Association of Cost Engineers*, 11–14 July, Dearborn, MI, pp. 5.1–5.10.

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, TSO, London.

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*, 2nd edn, Elsevier Butterworth-Heinemann, Oxford.

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, in Wang, J. (ed.), *Encyclopedia of Data Warehousing and Mining*, Idea Group Inc., IGI Global, pp. 770–6.

Pattengale, N.D., Alipour, M., Bininda-Emonds, O.R., Moret, B.M. and Stamatakis, A. (2009) How many bootstrap replicates are necessary? *Journal of Computational Biology*, 17(3), 337–54.

Sarle, W.S. (1994) Neural networks and statistical models, in *Proceedings of the Nineteenth Annual SAS Users Group International Conference*, SAS Institute Inc, North Carolina, USA, pp. 1538–50.

Skapura, D.M. (1996) Building Neural Networks, ACM Press, New York, USA.

StatSoft Inc. (2008) *A Short Course in Data Mining*, StatSoft Inc, Tulsa, OK.

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

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), 806–20.

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.

Appendix A3

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.

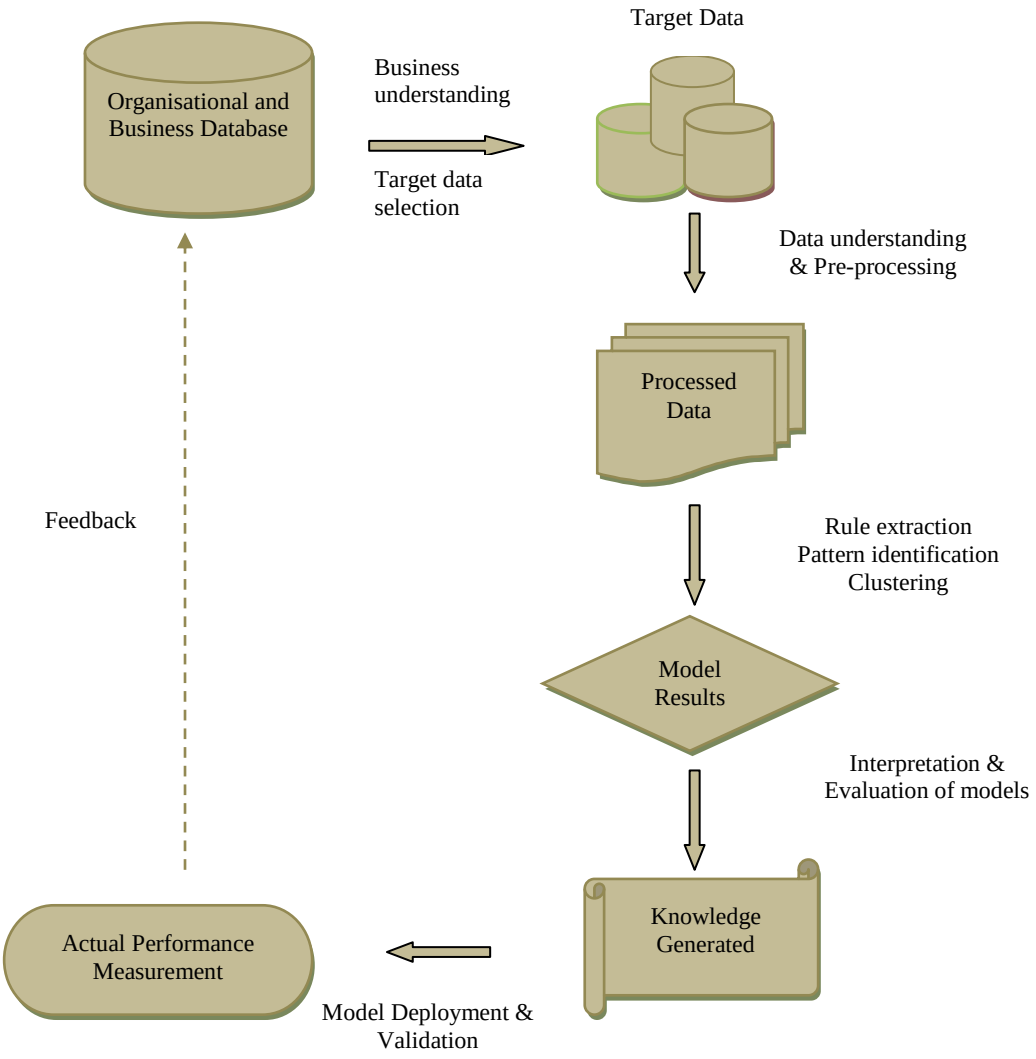


Figure 1: The generic data mining process

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

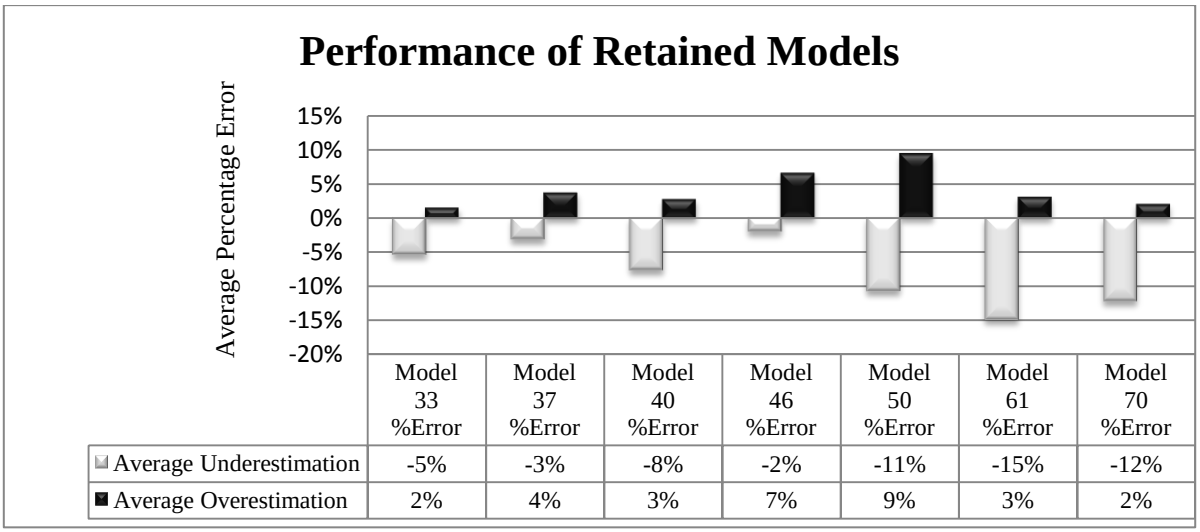


Figure 2: Performance of selected models

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.

Appendix A4

EXPLORING ESCALATION OF COMMITMENT IN CONSTRUCTION PROJECT MANAGEMENT: CASE STUDY OF THE SCOTTISH PARLIAMENT PROJECT

(This is the accepted copy of the article)

Citation:

Ahiaga-Dagbui, D.D and Smith, S.D (2014) Exploring escalation of commitment in construction project management: Case study of the Scottish Parliament project In: *Procs 30th Annual ARCOM Conference*, Raiden A (Ed.) Portsmouth, UK: Association of Researchers in Construction Management.

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

EXPLORING ESCALATION OF COMMITMENT IN CONSTRUCTION PROJECT MANAGEMENT: CASE STUDY OF THE SCOTTISH PARLIAMENT PROJECT

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

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

Successfully managing large construction projects within defined budget and time constraints has always been a major challenge largely because crucial decisions about the project's ultimate fate have to be made within an environment of significant uncertainty at the beginning of the project. It is not surprising that cost and time overruns are commonplace on construction projects. Existing literature often suggests economical, technical, political or managerial roots to this phenomenon. A less explored possible cause within construction management framework is the escalation of commitment to a course of action. This theory, grounded in social psychology and organisation behaviour, suggests the tendency of people and organisations to become 'locked-in' and 'entrapped' in a particular course of action and thereby 'throw good money after bad' to make the venture succeed. This defies conventional rationality behind subjective expected utility theory. Through a critical analysis of the literature, we identify different frequently cited enablers of escalation of commitment. Using a hindsight constructivist approach, we then demonstrate references to some of these enablers on the Scottish Parliament project. We found strong evidence in support of possible strategic misrepresentation, confirmation bias, self-justification and optimism bias. We highlight the importance of setting realistic time and budget constraints to circumvent escalation and make several recommendations to attenuate unwarranted escalation of commitment, including the use of an objective outsider to evaluate responses to disconfirming information and the structuring of incentive systems that do not punish for inconsistency in order to curb the effects of self-justification and reputation management.

Keywords: cost overruns, confirmation bias, escalation of commitment, self-justification, strategic misrepresentation.

INTRODUCTION

Literature in social psychology and organisational behaviour suggests that after investing time, money, energy and other resources in a chosen course of action, individuals and decision makers often become "locked-in" or "entrapped" in that course of action, sometimes even if the venture is failing. Staw's (1976, 1981) seminal work on escalation of commitment seeks to explain why decision makers sometimes embark on a questionable course of action and then persist with them above and beyond what the objective facts suggest. The thesis of his work suggests that negative feedback on a previous decision often tends to rouse the feeling of self-justification and regret of that particular decision, thereby resulting in a reinforcement of additional

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

resources (money, time or effort) to try and make the course of action pay off. Consider the following situations:

1. A representative of an equity firm makes a decision to invest £5 million in a new IT start-up that is expected to take about 3 years to develop and implement. It emerges after two years that the IT firm is having liquidity issues and that the product might require additional funds of £2.5 million and a year's extension. The equity firm must decide whether to write-off the initial £5 million investment or commit the additional funds to give the project a chance of success. Should they cut their losses now, risk losing a total £7.5 million, or stake their chance at gaining much more should the project eventually succeed?
2. A Government proposes an grand project that will represent the essence and ideals of a people and be a symbol of devolution and national distinctiveness at £40 million. Two years later, it becomes obvious that it is impossible to complete the project at that cost and a new estimate was set at £119 million, with legislators imposing a cap of £195 million in the third year. By the 4th year, cost had increased to £241 million, rising twice in the 5th year to £295 million amidst several controversies. By the 6th year the cost reaches £376 million before project completion at £431 million in the 7th year.

Although each of the cases above presents different decision making situations, they both have a common trait - sequential decision patterns with one decision being made based on a previous. In each case also, a considerable amount of time, money and effort has already been committed to the venture and the results do not seem to be going as initially intended. Arkes and Blumer (1985) suggested that investment of resources often sets in motion non-rational sequential decision making process, with one form of commitment begetting further commitment. They further suggest that the more responsibility a person has for the outcome of an initial decision, the greater is the inertia towards further commitment. This tendency however, as noted by Bazerman and Moore (2008) defies the conventional rationality behind subjective expected utility theory which suggests that sunk costs or past losses should not enter into decisions regarding future gain (Bazerman and Moore 2008).

Using the theoretical framework described in the discourse above, this paper will explore the sources of escalation of commitment using the case study of the Scottish Parliament project. We examine official government publications and documentary evidence from the public enquiry that followed the controversies surrounding the project using a hindsight constructivist research approach. We focus on the events before and during the construction that created an environment for escalation and how these possibly led to the inevitable cost and duration overrun on the project. The next section of the paper explores the theory of escalation more closely, before we examine the Holyrood project for evidence of the locked-in syndrome. We then reveal some lessons learnt from the case study for construction project management with recommendations on how to attenuate unwarranted escalation tendencies.

THEORETICAL FRAMEWORK: ESCALATION OF COMMITMENT

Decision making experiments have provided a lot of evidence that individuals have a systematic bias towards escalation of commitment. Some of the reasons provided include a failure to treat previous investments as sunk cost (Arkes and Blumer 1985), self-justification (Staw 1981) and anticipated regret (Sarangee *et al.* 2013). In some cases, decision makers have used escalation of investments as opportunity to redeem a

previous sub-optimal choice (Kahneman 1994) whiles Brockner (1992) posits that escalation tendencies may be buoyed by personal responsibility for negative consequences. Traditional economic decision making models suggest that people are rational and would make decisions in an attempt to maximise expected utility. Sunk costs (past investments) must essentially therefore be considered as historical and irrecoverable, thus should not be considered in decisions regarding future course of action (Bazerman and Moore 2008). However, Barnes' (1984) work supports the supposition that decision making is often biased in favour of retrospective rationality- the sunk cost effect.

Organisations also demonstrate escalation tendencies, albeit in a more complex manner, according to Guler (2007). The presence of multiple members for decision making in organisations normally should increase the likelihood of recognising the irrationality of escalating commitment to a failing course of action. Bazerman *et al* (1984) thus found that groups are less likely than individuals to escalate commitment. They however added that where groups do escalate, they tend to do so to a greater degree than individuals, possibly because group dynamics tends to increase the level of justification to continue to support an initial venture. We refer to this here as the *strength in numbers* effect.

A tale of two schools

There are essentially two schools of thought on escalation phenomenon. Decision error theorists, after Staw (1976), maintain that escalation is a result of a systematic bias in decision making where people, especially those that have personal responsibility for the outcome of the project or have a vested interest in the project, interpret feedback to support their point of view (Caldwell and O'Reilly 1982). According to Nickerson (1998), this can either be intentional or that the decision maker unknowingly falls to the curse of a confirmation bias - the seeking and interpretation of feedback in ways that are partial to existing beliefs or expectations.

Decision dilemma theorists, after Bowen (1987), however point to uncertainty of information and argue that feedback is often equivocal and that it is impossible to accurately predict how any venture will eventually turn out. Hantula and DeNicolis Bragger (1999) posit that these uncertainties could explain why it may be a prudent, at least at the time of making the decision, to continue to give the project a chance. Whether the project eventually fails or succeed is not necessarily a result of one wrong decision to rectify a previous sub-optimal choice, but simply a decision made amongst many alternatives in an environment of uncertainty.

Sequential investment and escalation

Sequential investment projects are particular susceptible to escalation tendencies because the venture does not generate intermediate financial payoffs until its complete. There is also some level of uncertainty over the amount and timing of the investment that will be required over the life of the project. Each investment stage therefore presents more opportunity cost as well as a milestone to either escalate commitment or pursue an alternative course of action. As found by Shepherd and Cardon (2009) however, terminating unsuccessful projects often comes with negative attending consequences including loss of job or losing face within an organisation. Decision makers often thus attempt to keep projects running by using end-gaming and using future-perfect strategies (Clegg *et al.* 2006). Strategic misrepresentation, the deliberate distortion or misstatement of the amount of time or resources necessary to complete the venture is not an uncommon tactic either (see Jones and Euske 1991).

Table 1 summarise some of the factors that create an environment that enable escalation of commitment.

Table 1: Escalation enablers

	Category	Description	Sources
1	Sunk-cost effects	Tendency to continue an endeavour because some amount of money, time or effort has already been invested in it. Investment begetting more investment.	Arkes and Blumer (1985), Brockner *et al* (1986)
2	Optimism bias	Overestimating the likelihood of positive events while downplaying the occurrence or severity of negative events.	Tversky and Kahneman (1974), Flyvbjerg (2007)
3	Future-perfect strategies (End-gaming)	Forward looking projection of ends with a visualization of the means by which that projected future may be accomplished	Weick and Kiesler (1979) , Clegg *et al* (2006)
4	Strategic Misrepresentation	Deliberate distortion or misstatement of the amount of resources or time necessary to achieve an aim	Jones and Euske (1991), Flyvbjerg (2007)
5	Confirmation bias	Tendency to seek or interpret information in ways that are partial to existing beliefs or expectations	Cadwell and O'Reilly (1982)
6	Norms of consistency	Consistent and decisive leaders are often viewed as better leaders. Decisions makers tend to stick to their initial decisions to main this consistency.	Staw and Ross (1980), Wellen *et al* (1998)
7	Image/Reputation Management	Not wanting to appear indecisive or incompetent Driven by feelings of personal responsibility	Smith and Terry (2003), Shepherd and Cardon (2009)
8	Self-justification	Unwillingness to admit to oneself, and/or others that a previous decision was sub-optimal.	Festinger (1962), Brockner (1992)
9	Organisational & Political influences	Coercive and normative pressures using institutional power or authority	Pfeffer (1992), Guler (2007)

Construction projects normally involve a series of sequential decisions before actual construction begins. Most projects will go through long feasibility and gestation periods before project approval and eventual delivery. These phases involve an iterative process of information acquisition and incremental commitment over a period of time, presenting a conducive environment for escalation of commitment. Where a project has commercial interest and is subject to sequential investment, the project often tends to be perceived as an end in itself according to Winch (2013), and therefore must be completed, no matter what, in order to recoup any initial investments.

RESEARCH APPROACH

Winch (2013) explored the three-pronged effects of future perfect strategising, strategic misrepresentation and escalation of commitment on the Channel Fixed Link project in an attempt to develop a broader organisational perspective on cost escalation in major projects. He proposed a hindsight constructivist or historical approach as research method to help fully comprehend the organisational complexities that led to overruns. Winch suggests that this approach will help comprehend the idiosyncratic embeddedness of major construction. We adopt a similar approach in this paper as it best helps for sense-making of the political and social construct of our case study - the Scottish Parliament building (Holyrood Project). We explore escalation of commitment using official documentary evidence from the government commissioned public enquiry that followed the controversies surrounding the construction of the Holyrood project (Fraser 2004). We also examine the Auditor General's reports (2000, 2004) and the Spencely Report (2000) submitted to the Scottish Parliamentary Corporate Body.

CASE STUDY: HOLYROOD PROJECT

Completed 3 years late in 2004, at a cost of £431million, The Holyrood Building in Edinburgh houses the Members of the Scottish Parliament (MSPs). Its final cost is approximately ten times more than the headline final cost of £40million announced in the Government's devolution White Paper, Scotland's Parliament (1997). The Government commissioned the Spencely Report (2000) to investigate cost and time overruns on the project. This was followed by two major probes by the Auditor General (2000, 2004) before the defining public enquiry, chaired by Lord Fraser of Carmyllie (2004) after project hand-over to investigate key decisions undertaken throughout the project delivery. There were 66 witnesses and more than 13,000 documents examined for the Public Enquiry (PE) alone. A full transcript of the transactions at the enquiry can be found at www.holyroodinquiry.org. These reports, as well as minutes of parliamentary proceedings, provide a rich source of documentary evidence to support the empirical analysis conducted in this paper.

The Act of Union of 1707 merged the Parliaments of Scotland and England into the Parliament of Great Britain, housed in the Palace of Westminster in London. Scotland was now effectively directly governed from London as a result (Colley 1992). However, in September 1997, the people of Scotland voted "Yes" in a referendum that would see the creation of the first Scottish Parliament in almost 300 years. Donald Dewer was appointed Secretary of State with the mandate to oversee the construction of a the parliament house. He became the main project champion, a key player and driver of what was to represent Scottish identity and aspirations. But the euphoria surrounding the referendum at this time led to many ill-considered decisions that created a conducive environment for escalation.

Optimism bias

First was the unrealistic cost ceiling of £40million. This turned out to be a rather optimistic estimate, or better still, a guesstimate of final cost of the project by non-construction professionals. Recall that a central theme of escalation theory is the increase in resources devoted to a venture in an attempt to redeem a previous sub-optimal choice. A member of the Scottish Parliament Corporate Body, Andrew Welsh MSP, stated that "*right from the very start, the budgets were totally unrealistic. The original budgets we inherited were for a fictional building*" [11 February 2004]. Rusell Hillhouse, former Permanent Under-Secretary at the Scottish Office and a member of the team that estimated the cost of the project at £40million said "*we couldn't possibly have done a thorough job, and this was very difficult because it was a time when people were working extremely hard on other aspects of the White Paper*" [PE 30th October 2003][2]. Sam Galbraith, former Under-Secretary of State at the Scottish Office also told the public enquiry, "*the figure of £40million in the white document, was never for Holyrood. That was for a bog-standard building on a greenfield site."* [PE 28 October 2003]. When asked how he knew the figure was not for Holyrood project, he responded "*That's what Donald [Dewer] told me*" suggesting that the project champion at this stage may have been aware that the cost of the project announced to the public was unrealistic.

Self justification, Reputation management and Norms of Consistency

Another sub-optimal decision that was made at the beginning of the project was the unrealistic completion date imposed on the project. Speed to build was priority for the

[2] Abbreviations: PE- Public Enquiry; MS/SE - Documentary evidences submitted to the public enquiry

project promoters who wanted the project completed within two years. This was strongly criticised by the opposition leaders. In a letter to all MSPs, Donald Gorrie MSP criticised the decision of the Scottish Office and the Secretary of State, Donald Dewer, writing "*There is no need for this haste...There has been widespread informed criticism of the fast timetable, for which there is no need. Professionals and organisations favouring the Holyrood site, favour a delay while the plans, timescale and budget are revised*" [MS/16/042 - 043]. Alex Salmond MSP also insisted that there was no need to try and deliver the project within such a short duration. He wrote to Donald Dewer, *"...it is quite impossible to have any new debating chamber of quality... ready by the time of the elections to the Scottish Parliament in 1999*" [MS/1/071 – 079]. Ignoring these warnings, however, the project sponsors still proceeded with the 2 year duration.

At least three enablers of escalation might have been at play at this stage - political reputation management, self justification and maintaining norms of consistency. Negative feedback on a past decisions calls the validity of the original decision into question and is dissonant with a decision maker's natural desire to see himself as competent. Many decision makers would often escalate commitment to their previous decision in order to prove that the initial decision was valid. In the case of the promoters of the Holyrood project, choosing a fast tract delivery method suddenly became very appealing if they had to meet 2 year deadline. Construction management procurement method was thus chosen as it has the advantage of allowing both design and project construction to occur concurrently. Using conventional construction methods of design before building would have added an extra 18months to the duration, according to William Armstrong, the Project Manager [PE 3 December 2003]. However, using construction management may well have been the single most important decision that was largely responsible for the cost and time overrun experienced on the Holyrood project. The client bears all financial risks associated with delays and design changes and final cost of the project could not be realistically known until all designs were completed. In addition, there is little incentive for the design team to keep cost low when such a method is used. Paul Grice, Clerk and Chief Executive of the Scottish Parliament told the public enquiry *'It is a fact of construction management - until you let the last tender, and settled the last claim, you can't know the final amount*' [PE 10 February 2004]. Robert Brown MSP, a member of the Scottish Parliament Corporate Body that was in charge of the project at one point aptly explains the source of the problems on the project. He noted, "*the signature design, the contractual method, and the process of developing the design detail, I increasingly came to the view that most of our difficulties [experienced on the project] were in a sense inevitable once the button was pressed at the beginning by the Scottish Office when they let the contract in the first place.*"

Strategic misrepresentation

There was evidence of strategic misrepresentation, the deliberate distortion or misstatement of the amount of time and resources necessary to achieve an aim, at many stages during the procurement of the project. Five weeks after their election 1999, the new MSPs had to vote on whether or not to continue the project. At this stage, Alex Salmond MSP, leader of the main opposition party wrote to Sir David Steel MSP, the Presiding Officer of the Scottish Parliament, requesting that the project be suspended and that an estimate of possible cancellation cost be produced "*in order to properly debate the future of the Holyrood project or other alternatives*" (MS/1/083). He further wrote in a follow-up letter, *"It is now possible that we may have to consider cancelling the Holyrood project; in the circumstances it is essential*

that no further actions should be taken which would add to the cost of cancellation if this were the decision which Parliament reached." [MS/1/084]

Faced with the dire prospect of possible project cancellation, civil servants in the Scottish Office, led by Barbara Doig, the Project Sponsor, decided to hide the fact that costs were going to be significantly higher than what the MSPs were to vote upon. In a classic example of strategic misrepresentation, the Project Sponsor did not include an extra £27million for risk in the estimates submitted to the MSPs. She later insisted that she was *'confident the £27million could be managed out'* and therefore was not to be included in the information given to the members of the Scottish Parliament

The proposed vote for an amendment urging a termination of the project was defeated by only three votes. Alex Salmond MSP, later told the public enquiry that the vote was based on false information, adding, *"it is inconceivable that had the proper information been given to the members of the Scottish Parliament, that there wouldn't have been at least a delay for taking stock and reassessment... the figures, the facts, the timeline shows that when the Parliament were told they were inheriting a project of £109 million, it was actually well over £200 million and was totally out of control... Parliamentarians being misled and misinformed is a very serious issue indeed."* [PE 13 November 2003]

Lord Fraser himself makes a strong case for strategic misrepresentation on the Holyrood Project by stating "*As at the point of hand-over, where there is a very tight vote in the Parliament on whether to proceed with this particular project or not, that figure was specifically kept away from them. It looks rather as though, those who were involved in this were determined to keep the figure down as low as possible, even to the point of concealing it from the Parliament, in the hope that the project would go ahead.*"

Political end-gaming and future-perfecting strategies

There was a lot of evidence supporting political end-gaming and future-perfecting strategies in the early stages of the project as well. Donald Dewer and the project team seem to have capitalised on the newly found nationalistic sentiments and euphoria around the referendum. The project was continuously presented to the public as one that will represent the essence of Scottish devolution and be an "important symbol for Scotland" that will "pay tribute to the country's past achievements and signal its future aspirations" (Scotland's Parliament 1997). Riding on these sentiments, Donald Dewer probably felt the need to build momentum and get the project started quickly. Consensus regarding some key decisions was ignored as he bypassed the consent of MSPs at many strategic stages, including the choosing of a site of the project [See MS/1/071 – 079]. It emerged during the public enquiry that he felt he had to 'endow' the MSPs with the new building and that if the decision of location of the building was not made quickly enough, the MSPs will never get around to doing it themselves. He probably also was aware that once the first concrete was poured, the project would become like a moving train that could not be stopped.

Confirmation bias

Confirmation bias, the tendency to seek or interpret information in ways that are partial to existing beliefs or expectations, played a key role in escalation on the Holyrood project. William Armstrong, an experienced project professional was the First Project Manager for the Holyrood Project at the Scottish Office. He resigned from his role because of frustrations he experienced regarding the spiralling cost and time delays. He was critical of the performance and commitment of the Architect,

Enrique Miralles writing to Project Sponsor, Barbara Doig, "*There is no indication that Miralles [can] remedy the deficiencies in time, cost and design to meet the programme.*" [PE SE-4-044]. His resignation letter prophesied that if measures were not quickly taken to properly control and manage the project, the *"programme will drift, the cost will increase, the design team will make claims, the contractors will make claims, and the project will become a disaster*" [PE SE-4-044]. As indicated by Caldwell and O'Reilly (1982) and Kahneman (2011), confirmation bias leads a decision maker to underplay, and in some cases, even ignore disconfirming feedback on performance of any venture. William Armstrong's strong warnings were blatantly ignored by the project sponsor, who later stated that *'I was comfortable that a great deal was being done to ensure that we continue to be on program, that we got the cost sorted out and that we got the design to the quality required"* [PE 4 December, 2003]. She decided instead that it was better that William Armstrong be removed from his post. He resigned before he could be fired.

Political and organisational influences

There were very strong political and organisational influences at many stages of the project as well. For example, opposition MSPs requested a two month delay in the project to examine the whole project more closely and explore other possible options. Margo MacDonald MSP insisted during a parliamentary debate that "*too many questions are unanswered at this stage, and we plead with you [Donald Dewer] for the time to find adequate answers*" [17 June 1999]. As is usually the case, those responsible for the negative outcome of a particular decision tend to maintain the norms of consistency in order not to appear indecisive or appear politically weak. Donald Dewer thus responded that such a delays requested by the opposition parties would "*cost more than £3million in contract penalties*". He added, *"this Parliament would look like a laughing stock"* if the opposition party got its way during the debate in Parliament. When it became apparent that the opposition might be fighting a lost cause, Donald Gorrie MSP said in reference to Donald Dewer, *"it is a despotism, we have one man says what happens and we all obediently follow him"* [17 June 1999)].

There were other sources of problems on the Holyrood project including significant scope changes, the death of the architect Enric Miralles, shortly followed by the death of project champion Donald Dewer. However, we have only concerned ourselves with some of the factors that may have contributed to escalation of commitment with its attending significant cost and time overruns.

CONCLUSIONS

The present study concerns the escalation of commitment to a particular course of action in decision making. We identified different enablers of escalation from the literature including sunk costs, self-justification, confirmation bias and strategic misrepresentation. We then examined official documentary evidence on the Holyrood project using a hindsight constructivist approach for possible causes of escalation that ultimately resulted in the cost and time overruns experienced on the project. We found overwhelming evidence in support of the use of strategic misrepresentation, self-justification and reputation management during the project. The study also uncovered evidence of optimism bias on the part of project sponsors in defining the budget and time constraints for the project.

The case study suggests that escalation of commitment is a complex phenomenon with additive causes from different sources. We also highlight the importance of the early stages of a project, as decisions taken at this stage become increasingly difficult to

reverse. In general, it is important for project sponsors and decision makers to be aware of the fact that their decisions will tend to be biased by previous decisions, and that we all tend to have a natural inertia towards escalation of commitment, particularly after receiving negative feedback.

RECOMMENDATIONS

Knowing why and when escalation occurs can help managers avoid this common decision bias. However, as escalation may not always be readily obvious, it is important to put in place organisational structures that will help attenuate unwarranted escalation. The use of an objective outsider to evaluate our responses to disconfirming information, especially in situations of sequential decision making can be helpful in reducing escalation tendencies. It might be helpful to structure incentives so that decision makers are not punished for supposed inconsistency in order to curb the effect of self-justification. Increased monitoring, accountability, budget controls and scrutiny might also be helpful especially on large and complex projects.

While this paper deals with the sources of escalation and how it might be curbed, it is important to mention that escalation should not necessarily be considered as a negative tendency. There are situations where it might be economically rational to escalate commitment to keep options open or maintain personal and future business relationships. On cursory examination, this might sound divergent to the core of the foregone discussions in this paper. However, what is proposed in this paper instead is that decision makers should be aware of the difficulty of separating initial decisions from related future decisions. It might be prudent to actively search for disconfirming information to provide a balanced perspective on confirming information that we are more likely to intuitively seek.

REFERENCES

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

Audit Scotland (2000) *The new Scottish Parliament building, an examination of the management of the Holyrood project*, Audit Scotland.

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

Barnes, J H (1984) Cognitive biases and their impact on strategic planning. *Strategic Management Journal*, **5**(2), 129-37.

Bazerman, M H and Moore, D A (2008) *Judgment in managerial decision making*. 7 ed. New York: Wiley.

Bazerman, M H, Giuliano, T and Appelman, A (1984) Escalation of commitment in individual and group decision making. *Organizational behavior and human performance*, **33**(2), 141-52.

Bowen, M G (1987) The escalation phenomenon reconsidered: decision dilemmas or decision errors? *Academy of management Review*, **12**(1), 52-66.

Brockner, J (1992) The escalation of commitment to a failing course of action: Toward theoretical progress. *Academy of management Review*, **17**(1), 39-61.

Brockner, J, Houser, R, Birnbaum, G, Lloyd, K, Deitcher, J, Nathanson, S and Rubin, J Z (1986) Escalation of commitment to an ineffective course of action: The effect of feedback having negative implications for self-identity. *Administrative Science Quarterly*, 109-26.

Caldwell, D F and O'Reilly, C A (1982) Responses to failure: The effects of choice and responsibility on impression management. *Academy of management journal*, **25**(1), 121-36.

Clegg, S R, Pitsis, T S, Marosszeky, M and Rura-Polley, T (2006) Making the Future Perfect : Constructing the Olympic Dream. *In:* Hodgson, D and Cicmil, S (Eds.), *Making Projects Critical*. Basingstoke: Palgrave Macmillan.

Colley, L (1992) *Britons: forging the nation, 1707-1837*. Yale University Press.

Festinger, L (1962) *A theory of cognitive dissonance*. Vol. 2, Stanford university press.

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

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.

Guler, I (2007) Throwing good money after bad? Political and institutional influences on sequential decision making in the venture capital industry. *Administrative Science Quarterly*, **52**(2), 248-85.

Hantula, D A and DeNicolis Bragger, J L (1999) The Effects of Feedback Equivocality on Escalation of Commitment: An Empirical Investigation of Decision Dilemma Theory. *Journal of Applied Social Psychology*, **29**(2), 424-44.

Jones, L R and Euske, K J (1991) Strategic Misrepresentation in Budgeting. *Journal of Public Administration Research and Theory*, **1**(4), 437-60.

Kahneman, D (1994) New challenges to the rationality assumption. *Journal of Institutional and Theoretical Economics*, **150**, 18 – 36.

Kahneman, D (2011) *Thinking, Fast and Slow*. New York: Farrar, Straus and Giroux.

Nickerson, R S (1998) Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, **2**(2), 175.

Pfeffer, J (1992) Understanding power in organizations. *California Management Review*, **34**(2), 29-50.

Sarangee, K, Schmidt, J B and Wallman, J P (2013) Clinging to Slim Chances: The Dynamics of Anticipating Regret When Developing New Products. *Journal of Product Innovation Management*, **30**(5), 980-93.

Scotland's Parliament (1997) White Paper presented to Parliament by the Secretary of State for Scotland by Command of Her Majesty, Cm 3658. In, Edinburgh.

Shepherd, D A and Cardon, M S (2009) Negative Emotional Reactions to Project Failure and the Self-Compassion to Learn from the Experience. *Journal of Management Studies*, **46**(6), 923-49.

Smith, J R and Terry, D J (2003) Attitude-behaviour consistency: the role of group norms, attitude accessibility, and mode of behavioural decision-making. *European Journal of Social Psychology*, **33**(5), 591-608.

Spencely, J, D (2000) *A report on the Holyrood Project*, Edinburgh: Scottish Parliamentary Corporate Body

Staw, B M (1976) Knee-deep in the big muddy: A study of escalating commitment to a chosen course of action. *Organizational behavior and human performance*, **16**(1), 27-44.

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

Staw, B M and Ross, J (1980) Commitment in an experimenting society: A study of the attribution of leadership from administrative scenarios. *Journal of Applied Psychology*, **65**(3), 249.

Tversky, A and Kahneman, D (1974) Judgment under uncertainty: Heuristics and biases. *science*, **185**(4157), 1124-31.

Weick, K E and Kiesler, C A (1979) *The social psychology of organizing*. Vol. 2, Random House New York.

Wellen, J M, Hogg, M A and Terry, D J (1998) Group norms and attitude–behavior consistency: The role of group salience and mood. *Group Dynamics: Theory, Research, and Practice*, **2**(1), 48.

Winch, G M (2013) Escalation in major projects: Lessons from the Channel Fixed Link. *International Journal of Project Management*, **31**(5), 724-34.

Appendix A5

NEURAL NETWORKS FOR MODELLING THE FINAL TARGET COST OF WATER PROJECTS

(Author's Copy)

Citation:

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

Corresponding email: d.ahiaga-dagbui@ed.ac.uk

NEURAL NETWORKS FOR MODELLING THE FINAL TARGET COST OF WATER PROJECTS

Ahiaga-Dagbui, DD[1] and Smith, SD[2]

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

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

Keywords: Cost Estimation, Cost Modelling, Neural Networks.

INTRODUCTION

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

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

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

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

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

COST PLANNING AND ESTIMATION

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

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

COST MODELS

Ferry et al (1999) also describe cost models as the symbolic representation of a system, expressing the content of that system in terms of the factors which influence its costs. The models may be in the form of mathematical equations (eg. Regression models) or a set of defined steps to estimate the cost of a particular item (eg. Storey enclosure method). Cost models can be very useful in strategic level decisions such as bid/not to bid decisions, with potential saving of time and effort on non-viable projects. They are furthermore appealing because of current harsh economic climate with tough competition and limited resources. However, the production of reasonably

accurate, acceptable and timely parametric cost estimates can be a difficult task. For example, using only 4 different parameters for a project and considering three alternative values for each, and varying one at a time will produce 81 different project solutions or alternatives. This can be done rather rapidly using an computer-based model but will undoubted be a laborious task using traditional cost estimation (Sequeira 1999). The time, effort and resource level required for this task would mostly be unjustifiable at planning stages of a project, perhaps a strong suggestion that detailed cost estimates at strategic level are often far from the optimal solutions because of time and resource constrains.

ARTIFICIAL NEURAL NETWORKS

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

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

Applications of neural networks

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

Training the Neural Network

A neural network, like the human brain, learns from experience (Hinton 1992). Experience here refers to past data within the domain of the problem under study. The aim of any training regime is to help the network to continuously reduce the error of its predictions by varying the weights between its connections (Setyawati et al. 2003).

Examples of the training set are presented to the network in its input layer. These are then transferred to the hidden layer by some form of activation function, normally a linear activation function. Random weights are applied to these input values in the hidden layer and then their cumulative weighted values transferred to the output layer. If the training algorithm adopted is a supervised one, the result of the training, called the output, is compared to the target (the expected real value) at the output layer and the error (difference between the output and the expected value, normally measured as the root mean squared error RMSE) is computed. This is then sent as feedback to the network and an error function is used to try and minimize the value of the error in the next cycle of training. The most common form of learning is the back propagation method, which is a supervised learning method (Setyawati, Creese and Sahirman 2003).

Neural Network Problems

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

DATA

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

MODEL DEVELOPMENT

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

Data Pre-processing

The aim of data pre-processing is to structure and present the data to the model in the most suitable way as well as offer the modeller the chance to get to know the data thoroughly. For this research, extreme values and outliers were either re-coded or deleted from the sample set and missing values replaced with the mean or mode. Input errors were corrected and all cost values were normalised to 2010 with the base year 1995 using the BIS cost indices. Invariant variables, such as procurement option,

payment method, fluctuation measure and type of client, were removed from the variable set as they would only increase the model complexity and yet offer no useful information for model performance. Finally, categorical variables such as type of project, need for project, etc. were coded using the one-of-N coding, resulting in 4 sub-variables for type of soil for example (Good, Moderate, Poor, Not Applicable). Twenty-eight sub-variables resulted out of the initial 11 input variables. This coding allowed the model to infer importance on its own without the modeller imposing weightings or subjective ratings to the variables. Ninety project cases remained after the pre-processing stage and were then passed on for the modelling proper.

Phase One: TC and FTC

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

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

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

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

Where Oi is the prediction (network outputs)

T_i is the target (actual value) of the *i*th data case.

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

The results from the best network for this phase was rather unsatisfactory as the errors observed were very high (See table 1.0), most likely due to the use of the raw data for

the modelling. The best network at this stage was an MLP with 25 input variables, 31 nodes in the hidden layer. It was trained using a BFGS training algorithm, tanH and logistic activation functions in the hidden and output layers respectively.

Table 1: Network Performance: TC and FTC

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

Phase two: LogTC and logFTC

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

Table 2: Network performance: Log TC and logFTC

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

Phase three: log FTC and logTC with Weight Decay

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

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

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

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

Table 4: Relative Importance of Variables

CONCLUSIONS

Artificial Neural Network is used to develop a cost estimation model for water projects in this paper. Their ability to capture and generalise non-linear relationships are exploited to detect the interactions in qualitative variables like tendering method, contractors need for the project, location, site access and project type in developing cost models to predict the final target cost of water projects. The use of weight decay regularisation to encourage the development of a parsimonious network to improve the model's performance and reliability was also investigated. This showed significant

promise for future analysis if combined with other techniques like pruning and sensitivity analysis of predictor variables. As a prototype of a wider research, the results achieved are very satisfactory and will potentially be improved with a larger dataset in this on-going research

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Westney, R E (1992) *Computerized management of multiple small projects*. Marcel Dekker, Inc.
Wilmot, C G and Mei, B (2005) Neural network modeling of highway construction costs. *Journal of Construction Engineering and Management*, **131**, 765.

Appendix A6

A NEURO-FUZZY HYBRID MODEL FOR PREDICTING FINAL COST OF WATER INFRASTRUCTURE PROJECTS

Dominic D. Ahiaga-Dagbui[1], Olubukola Tokede[2], Simon D. Smith[1], Sam Wamuziri[3]

[1] *School of Engineering, University of Edinburgh, EH9 3JL, UK*

[2] *School of Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, EH10 5DT, UK*

[3] *Centre for Management Research, Glyndŵr University, Wrexham, LL11 2AW, UK*

Nine out of ten infrastructure projects exceed their initial cost estimates. Accuracy of construction cost estimates remains a contentious area of debate within both academia and industry. Explanations for this have ranged from scope changes, risk and uncertainty, optimism bias, technical and managerial difficulties, suspicions of corruption, lying and insufficient required information for accurate estimation. The capacity for tolerance and imprecise knowledge representation of fuzzy set theory is combined with the learning and generalising capabilities of neural networks to develop neuro-fuzzy hybrid cost models in this paper to predict likely final cost of water infrastructure projects. The will help to increase reliability, flexibility and accuracy of initial cost estimates. Neural networks is first used to develop relative numerical weightings of cost predictors extracted from primary data collected on 98 completed projects. These were then standardised into fuzzy sets to establish a consistent framework for combining the effect of each variable on the overall final cost. A three-point fuzzy lower, upper and mean estimate of likely final cost is generated to provide a tolerance range for final cost rather than the traditional single point estimate. The performance of the final models ranged from 3.3% underestimation to 1.6 % overestimation. The best models however averaged an error of 0.6% underestimation and 0.8% overestimation of final cost of the project. The results are now being extended to a larger database of about 4500 projects in collaboration with an industry partner.

Keywords: artificial neural network, cost estimation, cost modelling, cost overrun, fuzzy set theory.

INTRODUCTION

Infrastructure projects have an 86% likelihood of exceeding the initial cost estimates and 9 out of 10 of them exceed their budgets (Flyvbjerg *et al.* 2002). A key example

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

[2] O.Tokede@napier.ac.uk

Ahiaga-Dagbui DD, Tokede O, Smith SD and Wamuziri S (2013) A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects *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, 181-190.

is the case of the stadiums built for the 2010 FIFA World Cup games in South Africa. With overruns ranging between 5 to 94% of original cost, none of the 10 stadiums were completed within budget (Baloyi and Bekker 2011). There is overwhelming evidence in literature, and practice, which support the conclusion that cost overrun is endemic within the construction industry, irrespective of size, type, sector or geographical location of the project (see Jackson 2002; Flyvbjerg *et al.* 2004; Odeck 2004; Baloyi and Bekker 2011). Cost remains arguably one of the most important key performance indicators on most projects (Chan and Chan 2004; Yeung *et al.* 2008) so that statistics, such as the ones above, leaves most clients grossly dissatisfied, giving the industry a poor reputation regarding budget reliability (Agyakwa-Baah 2009).

Despite its importance, cost estimation is undeniably not simple, nor straightforward, largely due to the dearth of information required for detailed estimation. It is even made worse by the cloud of uncertainty that shrouds cost drivers in the early stages of the project (Hegazy 2002) and the changes that occur in scope and design of the project once construction actually begins (Love *et al.* 2011; Gil and Lundrigan 2012). It is an inexact science and estimators have to make decisions within an environment of uncertainty. Moreover, even though it is accepted that factors such as tendering method, type of client, location of project, procurement method, size of project etc. have an effect on final cost of a project, it is difficult to establish their measured financial impact (Ahiaga-Dagbui and Smith 2012). This complex web of cost influencing variables would make it seem that the decision-to-build, for most projects, is based on a somewhat unrealistic cost estimate that will inevitably be exceeded.

Against this backdrop, debates have not waned on causes and measures of cost overruns. A recent discussion on the Construction Network of Building Researchers (CNBR) left a number of unresolved questions. How accurate can estimates be? Is there an acceptable way to compare final cost of project to cost estimates? What is the most acceptable measure of cost performance on a construction project? Is it even possible to achieve certainty of cost estimates, when the very estimates are made in an environment of uncertainty? (see the Nov 2012 CNBR archive online).While the answers to these can be varied; even sometimes strongly opposing; it is difficult to disagree that clients and project financiers still require some form of reasonably accurate estimate of their likely financial commitment for a project before the project begins.

In this paper, the authors attempt to model the final cost of water infrastructure projects using gathered cost data and other project details such as location, procurement method, size of project, type of client, etc of 98 water infrastructure projects. This paper, a sequel to a previous that uses only neural networks for modelling final cost (see Ahiaga-Dagbui and Smith 2012) employs Neuro-Fuzzy (NF) hybrid models - a combination of neural networks and fuzzy set theory, drawing on synergies from the two techniques in an attempt to develop more accurate, reliable and consistent final cost models. The next section of the paper provides an overview of the two modelling techniques used in the paper- neural networks and fuzzy set theory, and then proceeds to develop a neuro-fuzzy cost estimation hybrid model before concluding with results achieved and potential extensions of this research.

NEURAL NETWORKS

Work on artificial neural networks stemmed from the curiosity to understand how the brain processes information. Haykin (1994) described the brain as a highly complex

and parallel information processing system, capable of performing very complex computations many times faster than many types of computer processors. Artificial neural network (ANN) is thus just a simplistic abstraction of the biological neural networks of the brain, endowed with the capability to learn from experience (or examples) and then generalise for new cases using the acquired knowledge even within sparse or incomplete data (Anderson 1995). They are able to adapt to changing environments (or datasets) and are often referred to as universal approximators because of their ability to closely map input to output spaces in different types of problem domains (Fausett 1994). They essentially seek underlying relationships between variables and are particularly suited for complex, hard-to-learn problems, where no formal underlying theories or classical mathematical and traditional procedures exist (Adeli 2001). Neural networks are very sophisticated modelling techniques capable of modelling extremely complex functions. In particular, neural networks are non-linear (Denton and Hung 1996). For many years linear modelling (Regression), has been the commonly used technique in most modelling domains since they have well-known optimization strategies. Where the linear approximation was not valid, which was frequently the case (Boussabaine and Kirkham 2008), the models suffered accordingly.

Arguably, the strongest argument against the use of ANN is its supposed 'black-boxness'(Olden and Jackson 2002)- it is difficult to extract knowledge from the neural network model or fully understand how it reaches its conclusions. In regression, for example, an equation with explainable physical properties is produced. This is not the case in ANN modelling - no equation results out of the model and the network weights and connections make little sense. How the inputs interact to produce the output is at best, only known to the model. In a previous model using the same data, only neural network is used to model final cost projects (Ahiaga-Dagbui and Smith 2012). In an attempt to illuminate the black-box of ANNs, the authors combine the learning and generalisation abilities of neural networks with the capacity for tolerance and imprecise knowledge representation of fuzzy set theory to develop a hybrid neuro-fuzzy cost model for cost prediction.

FUZZY SET THEORY

Fuzzy set theory is an aspect of contemporary mathematics which focuses on the ambiguities in describing events or classes. It is an attempt to formalise human abilities of conversation, reasoning, and decision-making in an environment of imprecision, uncertainty as well as conflicting and/or incomplete information (Zadeh 2008). It incorporates 'matter of degree' rather than crisp boundaries into decision variables (Tokede and Wamuziri 2012). Fuzzy set theory allows an approximate interpolation between observed inputs and output situations (Ross 2009) and provides a means for modelling human vagueness in judgment. It basically requires encoding certain decision parameters as fuzzy sets (Zadeh 2008).

The defining characteristic of a fuzzy set is embodied in its membership function (MF). According to Kim *et al.* (2006), an MF provides an effective way to translate subjective terms into mathematical measures. A variable in fuzzy logic could have a set of values, characterised in linguistic terms, such as short, medium or long duration of project, or poor, moderate and good ground conditions. MFs can be generated in a number of ways either using intuition or some other algorithmic or logical operations (see Ross (2009) on how to use genetic algorithm, neural networks, rank ordering or inductive reasoning in developing MFs).

Ross (2009) stipulates that fuzzy relations are analogous to classical mathematical functions and basically represent mappings for sets. Fuzzy relations share the mapping potentials exhibited by neural networks and hence provide a compatible interphase in problem solving. Relations exhibit mathematical properties such as reflexivity, transitivity and symmetry which ultimately helps in interpreting attributes in fuzzy systems (Zadeh 1994). Chen and Huang (2007) used fuzzy relations in estimating the possibility-of-meeting the completion time of a construction project.

Fuzzy relations could be also employed in establishing the strength and possible association between different pairs. This can be achieved through the composition operator - a mathematical operation that seeks to establish the relationship between similar elements in different universe of discourse (Zimmermann 2001). Two common variants of the composition operator are the max-product and max-min. According to Zimmermann (2001), the most frequently used composition operator is the max-min; though both procedures produce comparable results in many instances. The max-min composition operation basically implements the strength of one chain as equal to the strength of its weakest link; the maximum of this then represents the overall chain strength in the fuzzy system (Ross 2009). Applications in civil engineering and construction research have been reported in Ayyub (1997). For cost and risk evaluation, fuzzy sets helps in quantification of variables, whose nature could be considered as complex and fit for description within a range of options (Tokede and Wamuziri 2012). An overview of fuzzy logic applications in construction management is provided by Chan *et al.* (2009)

NEURO-FUZZY

Neural networks solves problems by identifying the underlying patterns between the variables in the data it receives (Ross 2009) and then makes predictions based on the knowledge acquired (Adya and Collopy 1998). They are powerful, easy to use (StatSoft Inc. 2011) and can deal with large number of variables and non-linear relationships (Denton and Hung 1996). Yet, they are limited by their 'black-box' nature (Patterson 1996; Olden and Jackson 2002). They also perform best when using numerical or continuous data (StatSoft Inc. 2011). The majority of the data used in this research happen to be categorical in nature - location, type of client, procurement method, etc. Fuzzy sets represent composition of graded categories using mathematics based on logical reasoning (Belohlavek *et al.* 2009). It attempts to formalise decision making in an environment of uncertainty and incomplete information (Zadeh 2008), the kind that aptly describes cost estimation of construction projects.

Tokede and Wamuziri (2012) suggest that fuzzy set theory may not function at its optimal best as a stand-alone mathematical framework. Its practicality and utility is enhanced by combining its logic with pre-existent mathematical formulations. NF hybrid models thus have the potential to effectively represent modes of reasoning and decision making that are approximate rather than exact (Zadeh 1994), the case of construction cost estimation. Yu and Lin (2006) present an NF model for mining information from incomplete construction databases whilst Bilgehan (2010) uses NF models predict concrete compressive strength. Boussabaine (2001) similarly presents NF models for modelling the likely duration of construction projects

MODEL DEVELOPMENT

The NF models reported in this paper have been developed in three main stages - the first using statistical methods to pre-process the collected data, the second using

neural networks to develop relative final cost weightings of predictors and lastly using fuzzy sets to predict final cost. These stages are detailed below.

Stage One: Data and Data Pre-processing

Details on 98 water infrastructure projects completed in Scotland between 2007 and 2011were collected. The nature of the projects ranged from construction of water mains, water treatment plants, Combined Sewer Overflows (CSOs), installation of manholes or water pumps and upgrades and repairs to sewers. All the projects were target cost contracts with values between £9,000-£14 million and durations from 1-22 months.

The collected data is processed so as to structure and present the data to the model in the most suitable way. For this research, extreme values and outliers were either re-coded or deleted from the sample set and missing values replaced with the mean or mode. Input errors were corrected and all cost values were normalised to 2010 with the base year 1995 using the infrastructure resources cost indices by the Building Cost Information Services (BCIS 2012). Screening of variables to the smallest number is desirable because simpler models are easier to deploy - a model with 15 variables means information has to be known about all these variables before the model can be used for prediction. Redundant predictors - variables 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 ranking, bi-variate histograms or cross-tabulation. Further variable screening using scree test, mean plots and optimal binning in Statistica 10 software, suggested the optimal number of variables for predicting final cost to be between 5-7 predictors.

Stage Two: Neural Network Modelling

The neural network stage of the model developed was to determine a consistent numerical weighting for all the predictors depending on their relative contribution to determining the final cost of the project. Ten initial predictors[2] were used as inputs in a 3-layered feed-forward back-propagation neural network architecture with Final Target Cost as output of the model. The 98 project cases were split in a 75:15:10% ratio for training, testing and validation respectively. The best model was developed through an iterative procedure of continually tweaking the neural network parameters i.e. hidden nodes and activation functions, to produce improved model performance. Model performance was measured using the correlation coefficient between predicted and output values as well as the Sum of Squares (SOS) of errors below:

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

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

The ten best networks were retained and further tested using the validation set to produce *Figure 2*. The validation set was not used in the training of the model so can be considered as an independent verification of the model's ability to generalise on new data. This gave a quick indication of the average error level of each of the models.

[2] Initial list of predictors for the neural network model: Type of Soil, Site Access, Type of Location, Contractor's Need for the Project, Frequency of Project, Type of Deadline, Awarded Target cost (transformed as logTC), Type of project, Tendering Strategy, Duration (transformed as logD)

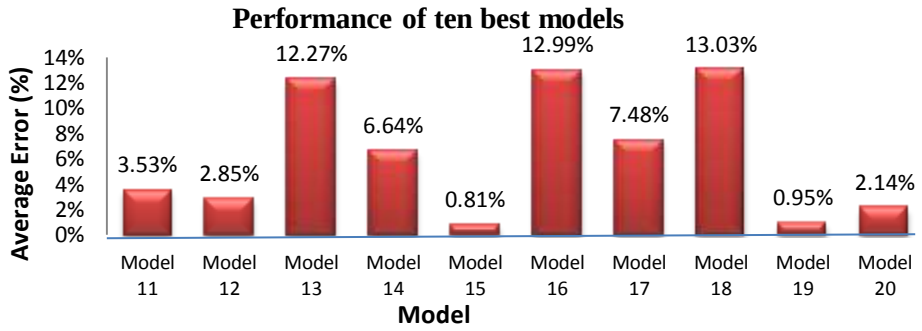


Figure 2: Performance of the ten best models

A sensitivity analysis was then carried out using the three best validated models in order to determine the contribution of each predictor to the model's performance. This was partly based on a test for parsimony using Ockham's Razor principle - one should not increase, beyond what is necessary, the number of entities required to explain anything and that all things being equal, preference should be given to the simplest hypothesis (Chase *et al.* 1996). This principle of simplicity is used to prune down the number of variables required in the model to predict the final cost, thus reducing inconsistencies, ambiguities and potential redundancies in the model. An initial ranking of all the predictors was generated based on their contribution to the model's performance. Then starting from the least important, one predictor was removed from the model at a time whilst measuring the performance of the model without that predictor. This was done until the model showed no further improvement or began to decay. The best set of predictors of final target cost after this stage are tendering strategy, site access, location, type of project, contractor's need for the project, type of soil, as well as estimated initial cost and duration (the common log of these were used in the model)

Table 1: Sensitivity analysis to determine relative ranking of predictors

Model	logTC	Tendering Strategy	Site Access	Type of Location	Project Type	Contractor's Need	Soil Type	logD
15. MLP 18-5-1	4.80	2.22	8.44	2.04	1.50	3.80	1.22	1.09
19. MLP 18-3-1	7.71	9.08	8.91	11.82	7.93	4.77	7.07	0.68
20. MLP 18-3-1	8.21	9.18	2.64	3.24	1.89	2.55	2.56	1.21
Average Weighting	6.90	6.83	6.66	5.70	3.77	3.71	3.61	0.99

Stage Three: Fuzzy Sets Modelling

Fuzzy set theory is applied at this stage of the modelling exercise to evaluate the subjective measures for each of the cost predictors in order to predict final cost. Using $\sum \ Normalized\ ranking = \frac{w_i}{\sum\ W}$=1 Eqn. 2. 2, the average weighted ranking for each of the variables from *Table* was normalized to unity in order to generate a standardised index for the subsequent fuzzy set computations (see *Table 4*)

$$\sum Normalized\ ranking = \frac{w_i}{\sum W} = 1 \qquad \text{Eqn. 2}$$

Where w*i* is the average relative weighting of the ith predictor
∑W is the sum of relative weighting of all predictors

Table 4: Normalized weighted values of the cost predictors from the neural network analysis

Factors	Tendering strategy	Site Access	Type of Location	Project Type	Contractor Need	Soil Type	Log Duration
Normalized ranking	0.22	0.21	0.18	0.12	0.12	0.11	0.04

With mean target cost to predictor plots, all predictors were fuzzified using the range set below:

$x \geq 5.8,$	Influence is Rather High
$5.6 \geq x \geq 5.8$	Influence is High
$5.4 \geq x \geq 5.6$	Influence is Medium
$x \leq 5.4,$	Influence is Low

The next stage of the fuzzy modelling involved developing membership functions. In developing these, the tolerance index is particularly relevant in evaluating and constraining the range of possibilities subject to a complex set of influencing variables, quantitatively and/or qualitatively defined. The tolerance index is vital in order to model the uncertainty in the cost values within a realistic continuum as opposed to a single figure-of-merit. For this study, the tolerances, β, were adapted to follow those indicated by Ayyub (1997) and reported in the table below.

Table 5: Values of tolerance. Source: adapted from Ayyub (1997)

β	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Poor/Low	1.0	0.9	0.7	0.4	0	0	0	0	0	0	0
Median	0	0	0.4	0.7	0.9	1.0	0.9	0.7	0.4	0	0
High	0	0	0	0	0	0	0	0.4	0.7	0.9	0
Rather High	0	0	0	0	0	0.4	0.7	0.9	1.0	0.9	0.7

Each of the project variables in the validation set was converted into fuzzy set variables using *Table 5*. According to Ross (2009), the fuzzy relation, $\tilde{T}$ of two sets, $\tilde{R}$ and $\tilde{S}$ can be defined by the set-theoretic and membership function-theoretic, mathematically expressed as:

$$\tilde{T} = \tilde{R} \circ \tilde{S} \qquad \text{Eqn. 3}$$

$$\tilde{\mu}_{T(x,z)} = \bigvee_{y \in Y} [\tilde{\mu}_{R(x,y)} \wedge \tilde{\mu}_{S(y,z)}] \qquad \text{Eqn. 4}$$

In Eqn. 3 above, R is a fuzzy relation on the Cartesian space X x Y. S is a fuzzy relation on Y x Z, and T is fuzzy relation on X x Z. In this cost estimation problem, R represents the set of cost predictors and S refers to the set of standard values of tolerance for linguistic descriptors of project attributes. The max-min composition operator is employed to deduce the strength and degree of relationship between specific relational pairs, which in this case, depicts the overall project cost as a fuzzy relationship of the normalised cost predictor weightings in *Table 4*, and based on the associated fuzzified project attributes deducible from *Table 5*.

The tolerance of each of the cost values in the validation set was computed, using Eqn.4 and defuzzified to obtain a 3-point estimate representing the fuzzy mean, fuzzy upper and fuzzy lower values as shown in *Table 6*. These three values provided a range of likely final cost rather than the customary single value estimate. *Table 6* shows the performance of the NF hybrid models in predicting the final cost of 10 different projects used in the validation set. This is summarised in

Table 7 along with the average model performance of the neural network model only.

The Fuzzy Upper best predicts the final cost and have the smallest percentage errors, ranging from 0.6% average underestimation to 0.8% overestimation of the likely final cost of the project. This represents an appreciable improvement in the results achieved using the neural network models only, also shown in

Table 7. The best three models at the neural network stage averaged a 1.2% under-estimation and 4.6% over-estimation of the actual final cost of the projects in the validation dataset. These results show significant promise in using neuro-fuzzy hybrid models to learn the underlying relationships between variables such as tendering strategy, site access, project location, type of soil or type of project and final cost of construction project.

Table 6: Neuro-fuzzy model validation results

		Model Prediction (log)					
Validation Cases	Actual Final Cost (log)	Fuzzy Lower (FL)	% error (FL)	Fuzzy Mean (FM)	% error (FM)	Fuzzy Upper (FU)	% error (FU)
1	5.78	5.65	2.4%	5.68	1.8%	5.75	0.5%
2	6.90	6.75	2.2%	6.77	1.9%	6.86	0.7%
3	5.41	5.35	1.1%	5.39	0.5%	5.46	-0.9%
4	5.22	5.09	2.6%	5.12	1.9%	5.20	0.5%
5	6.51	6.38	2.0%	6.41	1.6%	6.48	0.4%
6	5.95	5.85	1.7%	5.87	1.4%	5.95	-0.1%
7	6.91	6.78	1.9%	6.80	1.6%	6.89	0.4%
8	4.67	4.58	1.8%	4.62	1.1%	4.69	-0.5%
9	5.00	4.97	0.6%	4.99	0.1%	5.07	-1.6%
10	4.49	4.34	3.3%	4.36	2.9%	4.45	0.9%

Table 7: Summary of results from neuro-fuzzy model validation

	Summary of results			
	Neuro-fuzzy Lower (FL)	Neuro-fuzzy Mean (FM)	Neuro-fuzzy Upper (FU)	Neural Network Only
Average % Under-estimation	2%	1.50%	0.60%	1.2%
Average % Over-estimation	N/A	N/A	0.80%	4.6%

As already stated, even though it is agreeable that these factors affect the final cost on a project, it is difficult to assign cost measures to them as their relationship to cost are not thoroughly understood. The neuro-fuzzy hybrid models are possibly a step in the right direction in producing more accurate and realistic cost estimates at the initial stages of a construction project in an attempt to alleviate the problem of cost overruns

CONCLUSION

The research reported in this paper combines the learning and generalisation capabilities of artificial neural networks with fuzzy logic's ability to formalise human reasoning and decision making within an environment of uncertainty and incomplete information to develop neuro-fuzzy hybrid cost models for predicting the final cost of small water infrastructure projects. In particular, the research attempts to use some non-traditional cost predictors such as site access, location, tendering strategy, project and soil type to estimate likely final cost. The authors present a three-point range of possible likely final cost outcomes instead of the classical single point estimate. This

might allow estimators and clients to more accurately estimate likely contingency needs for their projects. In their extended form, these models can readily be converted into stand-alone desktop applications that can allow quick simulation of what-if scenarios and also allow the easy generation of different cost estimates should project parameters change. As a sequel to a previous paper that used only neural networks, the results here show an improvement in the predictive performance and thus the results are now being extended to a database of 4500 projects with an industry partner.

REFERENCES

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

Adya, M and Collopy, F (1998) How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, **17**(5-6), 481-95.

Agyakwa-Baah, A B (2009) *Risk Management in the Ghanaian Construction Industry* Master thesis Unpublished Thesis, Sheffield Hallam University, UK.

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.

Ayyub, B M (1997) *Uncertainty modeling and analysis in civil engineering*. CRC.

Baloyi, L and Bekker, M (2011) Causes of construction cost and time overruns: The 2010 FIFA World Cup stadia in South Africa. *Acta Structilia*, **18**(1).

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

Belohlavek, R, Klir, G J, Lewis III, H W and Way, E C (2009) Concepts and fuzzy sets: Misunderstandings, misconceptions, and oversights. *International journal of approximate reasoning*, **51**(1), 23-34.

Bilgehan, M (2010) A comparative study for the concrete compressive strength estimation using neural network and neuro-fuzzy modelling approaches. *Nondestructive Testing and Evaluation*, **26**(1), 35-55.

Boussabaine, A H (2001) Neurofuzzy modelling of construction projects' duration I: principles. *Engineering Construction and Architectural Management*, **8**(2), 104-13.

Boussabaine, A H and Kirkham, R (2008) Artificial Neural Network Modeling Techniques for Applied Civil and Construction Engineering Research. *In:* Knight, A and Ruddock, L (Eds.), *Advanced research Methods in the Built Environment*. London: Wiley-Blackwell.

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

Chan, A P, Chan, D W and Yeung, J F (2009) Overview of the application of "fuzzy techniques" in construction management research. *Journal of Construction Engineering and Management*, **135**(11), 1241-52.

Chase, S, Weiss, M, Gibbs, P, Hillman, C and Urban, N. *The Physics and Relativity FAQ* 1996 [cited 19th November, 2012. Available from http://math.ucr.edu/home/baez/physics/General/occam.html.

Denton, J W and Hung, M S (1996) A comparison of nonlinear optimization methods for supervised learning in multilayer feedforward neural networks. *European Journal of Operational Research*, **93**(2), 358-68.

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

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

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.

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

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

Jackson, S (2002) Project cost overruns and risk management. *In:* Greenwood, D (Ed.), *Proceedings 18th Annual ARCOM Conference*, 2-4 September 2002 Newcastle, Northumbria University, UK. Association of Researchers in Construction Management, (Vol. 1) 99–108.

Kim, J, Lee, S, Hong, T and Han, S (2006) Activity vulnerability index for delay risk forecasting. *Canadian Journal of Civil Engineering*, **33**(10), 1261-70.

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.

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

Olden, J D and Jackson, D A (2002) Illuminating the "black box": a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling*, **154**(1-2), 135-50.

Patterson, D W (1996) *Artificial Neural Networks: Theory and Applications*. Singapore: Prentice Hall.

Ross, T J (2009) *Fuzzy logic with engineering applications*. 3ed. Chichester: John Wiley & Sons.

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

Tokede, O and Wamuziri, S (2012) Perceptions of fuzzy set theory in construction risk analysis. *In:* Simon, S (Ed.), *Procs 28th Annual ARCOM Conference*, 3-5 September 2012, Edinburgh, UK. Association of Researchers in Construction Management, 1197-207.

Yeung, J F, Chan, A P and Chan, D W (2008) Establishing quantitative indicators for measuring the partnering performance of construction projects in Hong Kong. *Construction Management and Economics*, **26**(3), 277-301.

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.

Zadeh, L A (1994) Fuzzy Logic, Neural Networks, and Soft Computing. *Communications of the ACM*, **37**(3), 77-84.

Zadeh, L A (2008) Is there a need for fuzzy logic? *Information Sciences*, **178**(13), 2751-79.

Zimmermann, H-J (2001) *Fuzzy set theory and its applications*. Springer.

Appendix A7

Mapping Relational Efficiency in Neuro-Fuzzy Hybrid Cost Models

Olubukola TOKEDE, S.M.ASCE[1]; Dominic AHIAGA-DAGBUI[2]; Simon SMITH[2] and Sam WAMUZIRI, M.ASCE[3]

[1] School of Engineering and the Built Environment, Edinburgh Napier University, Edinburgh, EH10 5DT, UKPH (44) 131-455-2837; email: O.Tokede@napier.ac.uk

[2] School of Engineering, University of Edinburgh, EH9 3JL, UK; email: D.Ahiaga-Dabui@ed.ac.uk, S.Smith@ed.ac.uk

[3] Centre for Management Research, Glyndŵr University, Wrexham, LL11 2AW, UK, S.Wamuziri@glyndwr.ac.uk

ABSTRACT

Significant improvements are achievable in the accuracy of cost estimates if cost models adequately incorporate issues of flexibility and uncertainty. This study evaluates the relational efficiencies of the fuzzy composition operators – the max-min and max-product, in establishing the final cost of water infrastructure projects. Cost and project data was collected on 1600 water infrastructure projects completed in Scotland between 2000 and 2011. Neural network is first used to develop relative weightings of relevant cost predictors. These were then standardized into fuzzy sets to establish a consistent effect of each variable on the overall target cost. The strength and degree of relationship of the normalized cost predictor weightings and the fuzzified project attributes were combined using the max-min and max-product composition operators to obtain project cost predictions. The predictions from the two composition operators are compared with the actual cost figures. Results show comparable performance in the efficiency of the composition operators. Based on statistical correlations, the max-product composition operator achieved on average a deviation of 1.71% while the max-min composition had an average deviation of 1.86%. Improvements in the relational efficiency of neuro-fuzzy hybrid cost models could assist in developing a robust framework for realistic cost targets on construction projects.

INTRODUCTION

One of the major challenges of forecasting is dealing with uncertainty (Hüllermeier 1997) - the broad range of variability of likely outcomes of any event. One approach to uncertainty analysis that allows for some degree of flexibility is the fuzzy sets framework. To a reasonable extent, fuzzy sets basically imply the inclusion of degree of belonging in evaluating variables (Zadeh, 2008). They help to capture irreducible uncertainty as well as model vagueness in human reasoning abilities. Fuzzy relations are special cases of fuzzy sets. Fuzzy relations can be defined as a vague relationship between some fixed numbers of variables (Chan *et al.*, 2009; Zimmerman, 2001). Relations in this case are normative structures that help to interpret the attributes of fuzzy systems. The composition operation is however one class of similarity relation that seeks to establish the relationship between similar

elements in different universe of discourse (Zimmermann, 2001). Two common forms of composition operations are the max–product and max–min compositions. Zimmerman (2001) opines that the max – min composition is the most frequently used and that the operations of fuzzy relations can be well defined using the Extension principle. This paper provides an evaluation of the max-min and max-product composition operator in neuro-fuzzy hybrid cost models. The paper briefly discusses construction cost estimation and neuro-fuzzy modelling before detailing the mapping strategies in neuro-fuzzy hybrid cost models. The paper then proceeds to evaluate the relational efficiencies of two composition operators in a neuro-fuzzy hybrid cost estimation model and concluding with results achieved and their implications for research using the two mapping strategies.

COST ESTIMATION

Effective cost estimation relates the design of 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 limit of expenditure (Kirkham and Brandon 2007). This stage in a project life-cycle is particularly crucial as decisions made during the early stages of the development process carry more far-reaching economic consequences than the relatively limited decisions which can be made later in the process. As noted by Hegazy (2002), 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. Cost estimation is so vital; it can seal a project's financial fate (Nicholas 2004). Rightly, or wrongly, cost estimates produced at the beginning of a project are used by the client to build their budget which often becomes 'the baseline' on which actual project performance may be measured and compared.

Cost estimation techniques range from model-based methods to model-free methods. In between these spectra, lies a variety of techniques available to estimate the cost of a project including traditional bills of quantity, activity schedule and detailed estimation. Model-based techniques consist of static sets of relationships which systematically handle inputs and methodologically translate them into output (Smit 2012). In situations where such relationships are analytical, they mimic some form of mathematical function (Ross 2009). Model-free techniques are more dynamic and adaptive and include fuzzy systems and neural networks (Lee & Lin, 1992).

NEURO-FUZZY COST MODELS

Artificial Neural Networks (ANN), henceforth referred to as neural networks (NN) with artificial implied, is an analogy-based, non-parametric information-processing system that has performance characteristics similar to a biological neural network of the brain (Anderson and McNeill 1992). They retain two features of the biological neural network: the ability to learn from experience and make generalisations based on this acquired knowledge (Haykin 1994). Neural networks are structured to provide the capability to solve problems without the benefits of an expert and without recourse to programming (Boussabaine and Elhag 1999)

Neural networks are promising tools when used in conjunction with fuzzy sets for developing adaptive systems (Kosko and Isaka 1993). Adaptive systems can generally identify rule patterns in incoming data. Neural network and fuzzy logic systems are both numeric model-free estimators and dynamic systems (Lee and Lin 1992). Neural networks provide a platform for classifying patterns without having to provide explanations on the possible sophistications employed by the classification machinery (Eklund 1994). The disadvantage in the neural network technique is that they often increase nodes sporadically or swap network structure arbitrarily (Lee and Lin 1992); a variability that puts to question its reliability. Besides, the blackbox-ness of neural networks, more or less consigns it to the realm of magical arts. Fuzzy models, on the other hand deteriorate significantly where data sets used for identification are highly heterogeneous (Pedrycz 1996). Moreso, its procedures do not seem easily understandable to many cost and construction professionals (Tokede and Wamuziri 2012). Synergizing neural network and fuzzy systems therefore provides promising potentials for intelligent hybrid systems (Lee and Lin 1992). Lin and Lee (1992) pointed out that hybrid learning algorithms perform better than supervised learning algorithm alone. In a more recent study by Ahiaga-Dagbui and Smith (2012), it was discovered that the best neural network models for 98 water infrastructure projects had an average underestimation and overestimation of 1.2% and 4.6% respectively. In comparison, the neuro-fuzzy hybrid cost model using the same dataset achieved an average performance of 0.6% and 0.8% (Ahiaga-Dagbui et al. 2013). Neuro-fuzzy techniques are one of the most common hybrid techniques employed in cost estimation problems. According to Chan et al (2009), such techniques are highly competent in handling pattern recognition and automatic learning. Ahiaga-Dagbui *et al.,* (2013) also suggest that fuzzy sets and neural networks both provide excellent mapping interphases which when combined could be invaluable in pattern recognition.

Mapping Strategies in Neuro-Fuzzy Cost Models

Fuzzy sets are useful in mapping non-empty sets to partially ordered sets (Sanchez 1976). They can be used to bridge the gap between mathematical models and their associated physical reality (Demicco and Klir 2003). This is mainly achieved by representing the vagueness associated with the linguistic description.. Fuzzy relations are essentially the means of modelling the intensity between elements of a fuzzy set. Fuzzy relations emerge from Cartesian representation of two or more sets on a universal scale (Belohlavek and Klir 2011).

A composition is a common mathematical operation that seeks to establish the relationships between similar elements in different universe of discourse (Zimmermann 2001). The compositionality assumption is a sort of logical generalization presupposing that the degree of membership of a compound fuzzy set is a function of the membership degrees of each component. Effectively, this implies the whole is summarily a sum and/or product of its parts (Belohlavek and Klir 2011). There have been contention on the possibility of a single non-parametric operator to appropriately model the meaning of ‘AND’ or ‘OR’ context independently. The composition method is commonly used in applications of artificial neural network for mapping between parallel layers in a multi-layer network.

According to Ross (2009), the fuzzy relation, $\tilde{T}$ of two sets, $\tilde{R}$and $\tilde{S}$ can be defined by the set-theoretic and membership function-theoretic, mathematically expressed as:

$$\tilde{T}=\tilde{R}\circ\tilde{S} \qquad \text{Eqn. 1}$$

Where R is a fuzzy relation on the Cartesian space X x Y. S is a fuzzy relation on Y x Z, and T is fuzzy relation on X x Z. In this cost estimation problem, R represents the set of cost predictors and S refers to the set of standard values of tolerance for linguistic descriptors of project attribute

Max-min Composition

The max-min composition is commonly used when a system requires a conservative solution. Loetamonphong and Fang (2001, pp6) explains this approach as when the "goodness of one value cannot compensate the badness of another value". Figure 1 shows a graphical illustration of the max-min composition. Ross (2009) pointed out the max-min composition is analogous to approximate reasoning using the IF-THEN rules.

Mathematically, the max-min composition can be represented as:

$$\tilde{\mu}_{T(x,z)} = \bigvee_{y\in Y}[\tilde{\mu}_{R(x,y)} \wedge \tilde{\mu}_{S(y,z)}] \qquad \text{Eqn. 2}$$

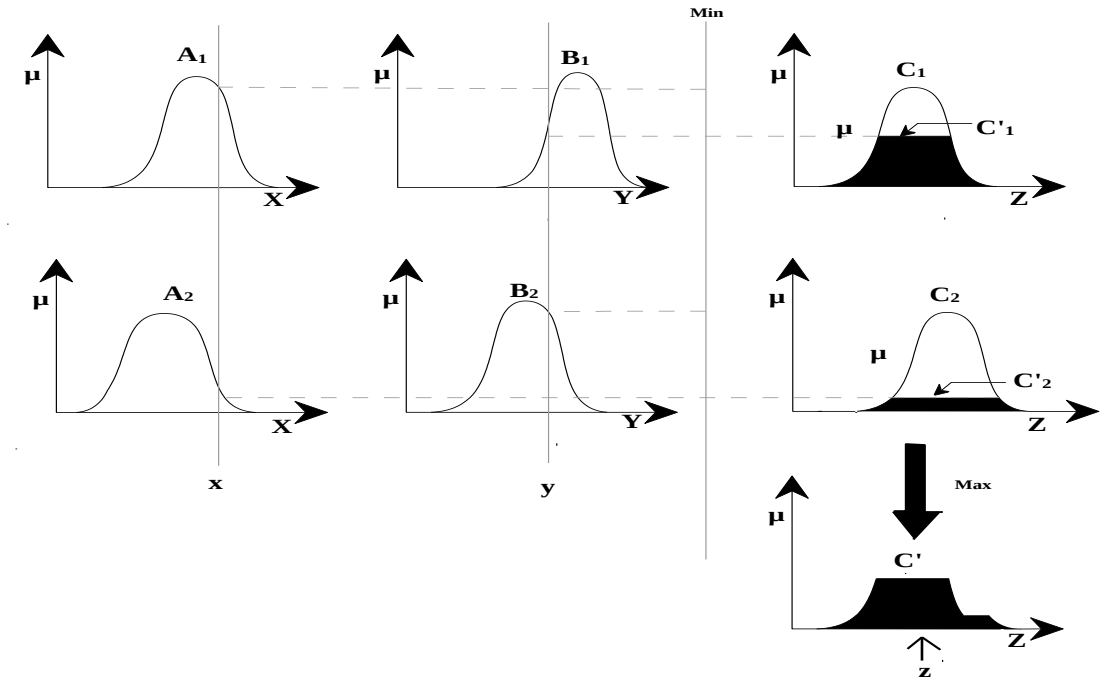


Figure 1 – Graphical illustration of the max-min composition (Dubois & Prade, 2000)

Max-Product Composition

The max-product composition is touted by some researchers as yielding better equivalent results (Loetamonphong and Fang 2001; Ross 2009). One possible

explanation is that conventional risk calculus is presumed to have a combinatorial character.

Mathematically, the max-product composition can be represented as:

$$\tilde{\mu}_{T(x,z)} = \bigvee_{y \in Y}[\tilde{\mu}_{R(x,y)} \bullet \tilde{\mu}_{S(y,z)}] \qquad \text{Eqn. 3}$$

The max-product composition is a fuzzy calculus that expresses the relationship between similar elements. Figure 2 shows a graphical illustration of the max-product composition .Ross (2009) illustrated the max-product composition to relate the rain gauge prediction of large storms to the actual pond performance during rain events.

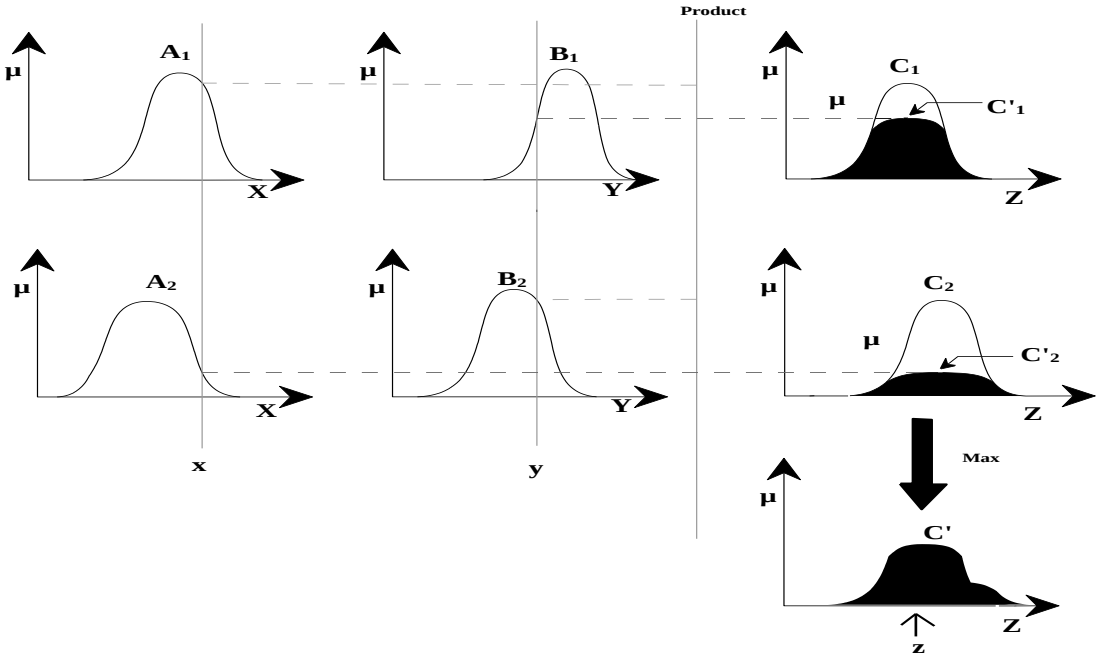


Figure 2 – Graphical illustration of the max-product composition (Dubois & Prade, 2000)

Other possible variants of composition include the max-max, min-min, max-average and sum-product (Ross 2009). Essentially, the composition involves employing hybrid formulations of min, max, average and product to arrive at some relationship formation; thereby specifying a range of mathematical values that could be tolerated by a category (Carpenter *et al.* 1992). Yager and Filev (1994) mentioned that the MAX operator ignores reinforcement inherent in the overlapping in the output fuzzy sets. Carpenter *et al.*, (1992) also stated that the MIN operator helps highlight features that are critically present, whilst the MAX operator flags-off features that are critically absent.

RESEARCH METHODOLOGY

The findings reported in this experimental paper were achieved using the following steps. Approximately 1600 projects completed between 2004 and 2012, with cost range of between £4000 to £15 million, comprising newly built, upgrade, repair or refurbishment projects were used for the study. One hundred cases were

selected using stratified random sampling to be used for independent testing of the final models. The remaining data were then split in an 80:20% ratio for training and testing of the neural network model. All cost values were normalized to a 2012 baseline with base year 2000 using the infrastructure resources cost indices by the Building Cost Information Services (BCIS 2012). The nature of the projects ranged from construction of water mains, water treatment plants, Combined Sewer Overflows (CSOs), installation of manholes or water pumps and upgrades and repairs to sewers.

The data was then pre-processed to structure and present the data to the model in the most suitable way. For this research, extreme values and outliers were either re-coded or deleted from the sample set and missing values replaced with the mean or mode. Input errors were corrected and all cost values were normalized to 2011 with the base year 1995 using the infrastructure resources cost indices by the Building Cost Information Services (BCIS 2012). Invariant variables, such as procurement option, payment method, fluctuation measure and type of client, were removed from the variable set as they would only increase the model complexity whiles offering little to no useful information for model's performance. Categorical variables such as type of project, need for project, etc. were coded using a binary coding (0, 1) format. Data screening using scree test and optimal binning allowed for the selection of five initial predictors (primary purpose of project, project scope, project delivery partners, estimated target cost and project duration) to be used for the actual ANN modelling. Several neural network models were then developed with the 20 best models used to estimate the relative contribution to model performance of each factor used. These values, as shown in Table 1 were then standardized into fuzzy sets in the next phase of the study to establish a consistent effect of each variable on the overall target cost.

Fuzzy Sets Modelling

Fuzzy set theory is applied at this stage of the modelling exercise to evaluate the subjective measures for each of the cost predictors in order to predict final cost. Using Eqn.4 the average weighted ranking for each of the variables from Table 1was normalized to unity in order to generate a standardised index for the subsequent fuzzy set computations (see Table 2)

$$\sum Normalized\ ranking = \frac{w_i}{\sum w} = 1 \qquad \text{Eqn. 4}$$

Where w_i is the average relative weighting of the ith predictor
$\sum W$ is the sum of relative weighting of all predictors

Table 1 - Normalized weighted values of the cost predictors from the neural network analysis

Factors	Project Scope	Primary Purpose	Delivery Partner	Duration	Target Cost
Normalized ranking	0.22	0.11	0.02	0.02	0.63

With mean target cost to predictor plots, all predictors were fuzzified using the range set below:

$outturn\ cost \geq$ £600,000,	Influence is Rather High
£400,000 $\geq$ $outturn\ cost$ $\geq$ £600,000	Influence is High
£100,000 $\geq$ $outturn\ cost$ $\geq$ £400,000	Influence is Medium
$outurn\ cost \leq$ £100,000,	Influence is Low

The next stage of the fuzzy modelling involved developing membership functions. In developing these, the tolerance index is particularly relevant in evaluating and constraining the range of possibilities subject to a complex set of influencing variables, quantitatively and/or qualitatively defined. The tolerance index is vital in order to model the uncertainty in the cost values within a realistic continuum as opposed to a single figure-of-merit. For this study, the tolerances, β, were adapted to follow those indicated by Ayyub (1997) and reported in theTable 2

Table 2: Values of tolerance. Adapted from Ayyub (1997)

β	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Poor/Low	1.0	0.9	0.7	0.4	0	0	0	0	0	0	0
Median	0	0	0.4	0.7	0.9	1.0	0.9	0.7	0.4	0	0
High	0	0	0	0	0	0	0	0.4	0.7	0.9	0
Rather High	0	0	0	0	0	0.4	0.7	0.9	1.0	0.9	0.7

Each of the project variables in the validation set was converted into fuzzy set variables using Table 2

ANALYSIS AND DISCUSSION

Table 3 reports the performance of the NF hybrid models in predicting the final cost for 5 of the 99 different projects used in the validation set. The tolerance of each of the cost values in the validation set was computed using Eqn.4 and defuzzified to obtain a 3-point estimate representing the fuzzy mean, fuzzy upper and fuzzy lower values as illustrated in Table 4. These three values provided a range of likely final cost rather than the customary single value estimate.. The overall results for the performance of the validation cases have been represented in Figure 3.

Table 3: Logarithmic Cost values for both composition operators

Project Validation cases	Max-Product Mean value	Max-min Mean Value	Actual Out-turn Cost value
Project Case 9	6.685	6.672	6.691
Project Case 204	5.592	5.572	5.670
Project Case 901	5.262	5.279	5.385
Project Case 505	5.877	5.934	5.980
Project Case 824	5.575	5.633	5.674

Based on statistical correlations, the max-product composition operator achieved on average a deviation of 1.71%; while the max-mean composition had an average

deviation of 1.86%. The Max-Product composition performed consistently better in both the fuzzy mean and fuzzy lower values but did not show any significant advantage in the fuzzy upper cost values. This might indicate that the benefit of the max-product operator is situated within the fuzzy mean and lower cost target predictions.

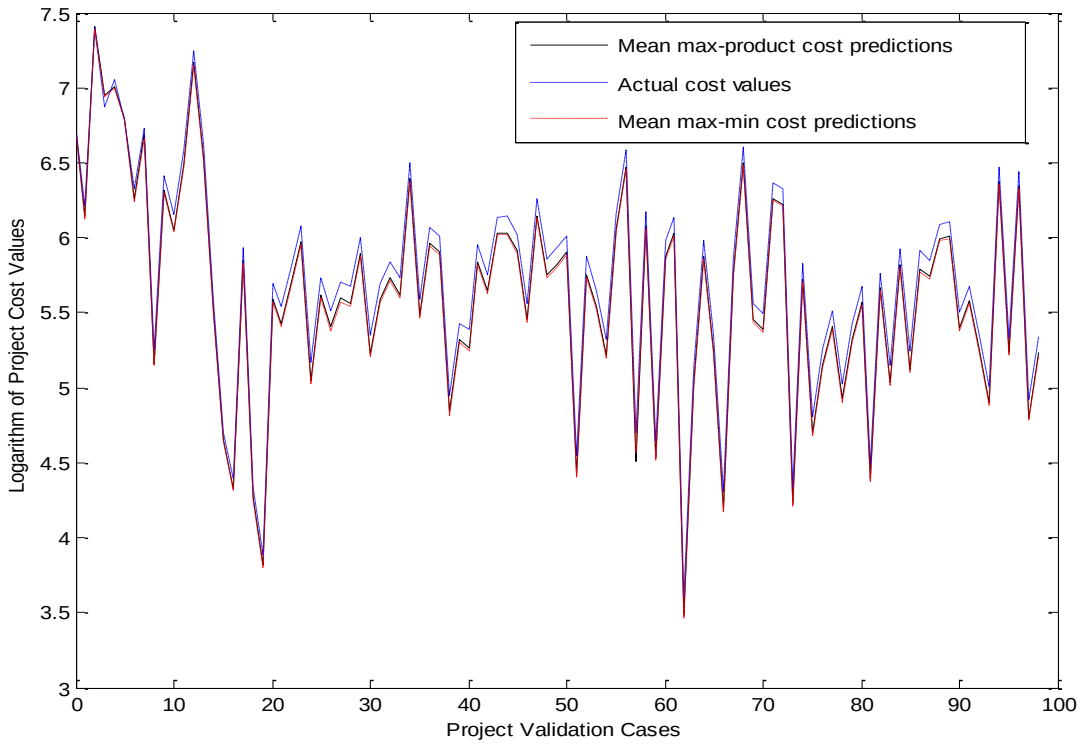


Figure 3 – Graphical plot of the project validation cases and the relational efficiency of composition operators

The corresponding percentage differences in the cost target were also estimated for all the 99 project validation cases. Table 4 provides a summary of the overall result obtained for all the validation cases.

Table 4: Summary of Results from Neuro-fuzzy Model Validation

Cost Category	Fuzzy Upper Value	Fuzzy Mean Value	Fuzzy Lower Value
Max-Min Operator	2.59%	2.07%	0.94%
Max-Product Operator	2.59%	1.74%	0.78%

The volatility measures considered for the range of values for the composition operators were fairly consistent. The standard deviation of the cost values of the max-product was £161,715, while that of the max-min was £188,506. This implies that the range of fluctuation in the max-min composition measure was higher than those obtained from the max-product composition predictions.

CONCLUSION

The research reported in this paper combines the learning and generalization capabilities of artificial neural networks with fuzzy logic's ability to formalise human reasoning and decision making within an environment of uncertainty and incomplete information. This paper develop a neuro-fuzzy hybrid cost model for predicting the final cost of small water infrastructure project and then evaluates the efficiency of the max-product and max-min composition operators in predicting the final target cost. Based on 99 project validation cases, it was found that the max-product composition operator achieved an average a deviation of 1.71% while the max-mean composition had an average deviation of 1.86%.

It is however noteworthy that this two composition operators are not an exhaustive treatment of the relational capabilities of fuzzy sets. However, they currently represent the most popular calculi employed in fuzzy set evaluations. There might be need to improve on the framework of the existing mathematical formulations of fuzzy sets in order to fully realize the potentials of fuzzy sets in modelling the vagueness in human reasoning and capturing irreducible uncertainties in water infrastructure projects. Improvements in the relational efficiency of neuro-fuzzy hybrid cost models will in no little way assist in developing a robust framework for realistic cost targets in water infrastructure projects.

REFERENCES

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

Ahiaga-Dagbui DD, Tokede O, Smith SD and Wamuziri S (2013) A neuro-fuzzy hybrid model for predicting final cost of water infrastructure projects *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, 181-190.

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

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

Belohlavek, R. and G. J. Klir (2011). Concepts and fuzzy logic, MIT Press.

Boussabaine, A. and R. Kirkham (2008). Whole life-cycle costing: risk and risk responses, Wiley-Blackwell.

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

Carpenter, G. A., S. Grossberg, et al. (1992). "Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps." Neural Networks, IEEE Transactions on 3(5): 698-713.

Chan, A. P., D. W. Chan, et al. (2009). "Overview of the application of "fuzzy techniques" in construction management research." Journal of Construction Engineering and Management 135(11): 1241-1252.

Demicco, R. V. and G. J. Klir (2003). Fuzzy logic in geology, Academic Press.

Dubois, D. and H. M. Prade (2000). Fundamentals of fuzzy sets, Kluwer Academic Pub.

Eklund, P. (1994). "Network size versus preprocessing." Fuzzy Sets, Neural Networks and Soft Computing: 250-264.

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

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

Hüllermeier, E. (1997) An approach to modelling and simulation of uncertain dynamical systems. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, **5**(02), pp.117-137.

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

Kosko, B. and S. Isaka (1993). "Fuzzy logic." Scientific American 269(1): 62-67.

Lee, C. G. and C. Lin (1992). Supervised and unsupervised learning with fuzzy similarity for neural-network-based fuzzy logic control systems. Systems, Man and Cybernetics, 1992., IEEE International Conference on, IEEE.

Lee, S. H., F. Peña-Mora, et al. (2006). "Dynamic planning and control methodology for strategic and operational construction project management." Automation in construction 15(1): 84-97.

Loetamonphong, J. and S.-C. Fang (2001). "Optimization of fuzzy relation equations with max-product composition." Fuzzy Sets and Systems 118(3): 509-517.

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

Pedrycz, W. (1984). "Identification in fuzzy systems." Systems, Man and Cybernetics, IEEE Transactions on(2): 361-366.

Pedrycz, W. (1996). Fuzzy modelling: paradigms and practice, Kluwer Academic Pub.

Ross, T. J. (2009). Fuzzy logic with engineering applications, Wiley.

Sanchez, E. (1976). "Resolution of composite fuzzy relation equations." Information and control 30(1): 38-48.

Smit, M. C. (2012). "A North Atlantic Treaty Organisation framework for life cycle costing." International Journal of Computer Integrated Manufacturing 25(4-5): 444-456.

Tokede, O. and S. Wamuziri (2012). Perceptions Of Fuzzy Set Theory In Construction Risk Analysis. Procs 28th Annual ARCOM Conference. S. Smith. Edinburgh. 2: 1197-1207.

Yager, R. R. and D. P. Filev (1994). "Essentials of fuzzy modeling and control." New York.

Zadeh, L. A. (2008). "Is there a need for fuzzy logic?" Information Sciences 178(13): 2751-2779.

Zimmermann, H. J. (2001). Fuzzy set theory-and its applications, Springer.

Appendix B

APPENDIX B

DATA FORMS

APPENDIX B1 - WATER TYPE PROJECTS

		Category			
	Type of data	**1**	**2**	**3**	**4**
	Project Information				
1	Project Start Date				
2	Tender Price (£)				
3	Cost at Completion (£)				
4	Number of Estimates, Nr				
5	Estimated Duration (Years, Months)				
6	Actual Duration (Years, Months)				
	Site Information				
1	Ground Condition	Contaminated	Non-contaminated	Made-Up	
2	Type of Soil[1]	Good	Moderate	Poor	
3	Site Access	Unrestricted	Restricted	Highly Restricted	
4	Type of Location	Remote	City Outskirts	City Centre	
	Other Information				
1	Type of Client	Public	Private	-	
2	Contractor's Need for the project	Low	Medium	High	
3	Frequency of Project[2]	One-off	Repetitive	-	
4	Type of Deadline	Normal	Strict	Fast-track	
5	Fluctuation Measure	Fixed	Formula	-	
6	Type of Project	Repair	Upgrade	New Order	
	Procurement/Contract Information				
1	Tendering Strategy	Open Competitive	Selective Competitive	Negotiated	Serial
2	Procurement Option	Design-bid-build	Design and Build	Management types	Partnering
3	Payment Method	Lump Sum	Ad-measurement	Cost Reimbursement	Target Cost

APPENDIX B2 - BUILDING-TYPE PROJECTS

	PROJECT INFORMATION SHEET (Building Type Projects)	
	Project Name/Code	
	Project Information	
	PART A	*(Please enter the required information in the cells below)*
1	Year of Completion	
2	Tender Price (£)	
3	Cost at completion (£)	
4	Project Start Date	
5	Estimated Duration (Years, Months)	
6	Actual Duration (Years, Months)	
7	Number of Storeys (Nr):	
	a) Below ground	
	b) Above ground	
8	Ground Floor Area (m^2)	
9	Typical GIFA (m^2)[3]	
10	Number of Elevators, Nr	

	PART B: Building-Type Projects						
		Category					
	Building Information						
1	Type of Structure	Steel	concrete	Masonry	Timber	Other	
2	Intended function of project	Hospital	Residential	Office	Educational	Leisure	Industrial
3	Type of Project	New Order	Renovation[4]	Demolition			
4	Type of foundation	Pad	Strip	Raft	Pile	Combined	N/A
	Site Information						
5	Ground Condition	Non-Contaminated	Contaminated	Made-Up	N/A		
6	Type of Soil[1]	Good	Moderate	Poor	N/A		
7	Site Access	Unrestricted	Restricted	Highly Restricted			
8	Type of Location	Remote	City Outskirts	Urban Area			
	Other Information						
9	Type of Client	Public	Private				
10	Frequency of Project	One-off	Repetitive				
11	Type of Deadline[2]	Normal	Strict	Fast-track			
	Contract Information						
12	Tendering Strategy[5]	Open Competitive	Selective Competitive	Negotiated	Serial		
13	Procurement Option	Design-Bid-Build	Design and Build	Management type	Partnering[6]		

[1] **Type of Soil** is classified according to the load bearing capacity as indicated in the table below

Type of Soil	Good	Medium	Poor
Bearing Capacity (KN/m^2)	>600	200-600	<200

[2] **Type of Deadline -** Speed of construction was noted as one of the possible causes of cost escalation in the literature. Three classifications, i.e Normal, Strict and Fast-track have been used here to try and capture the effect of type of deadline used for the project. Strict deadlines usually have an immovable deadline due to a scheduled use of the built facility for an event like the World Cup or Olympic Games.

Fast track projects, like the Scottish Parliament project, are usually scheduled to be built within a shorter time-frame than would normally be required. In case of the Scottish Parliament, design and construction occurred concurrently due to the fast track nature of the project.

[3] **Gross Internal Floor Area** (GIFA): Area between enclosing walls without deduction for internal partitions or openings

[6] **Tendering Strategy** - *Open tendering* is open to all contractors and requires that the contract is advertised publicly by the client. This may involve some level of prequalification.

In *selective competitive*, contractors are invited from a pre-selection list based on known track record and suitability for the contract in terms of its size, nature and complexity. The bid is usually awarded based on lowest evaluated tender cost.

Negotiated tenders are usually used for specialist works, to extend the scope of an already existing contract or for emergency works. It usually involves negotiations with a single contractors.

Serial tenders usually are based on either a schedule of works or bill of quantities. The accepted rates are then used to value works over a series of similar projects, often for a fixed period of time following which the tendering procedure may be repeated.

[5] **Partnering** includes all relationship procurement methods such as alliances, joint venture, PFI, PPPs, etc.

Appendix C

APPENDIX C

Choosing the software

The software package to be used for the data modelling was chosen using the following criteria:

1. Ease of use, preferably with a graphical user interface.
2. Cost (annual, initial licence cost, academic licence)
3. Availability of learning support/tutorials
4. Model deployment options (e.g. Html, C++, excel, PMML, SAS)
5. Flexibility (ability to try different alternative network topologies)
   A fully functional "demo" of the following neural network programs were evaluated using sample data that was to be used for the actual modelling later in the thesis.

a) **Tiberius, Version 7.0.3**

This package was a relatively easy to learn software, available for free on an academic license. It had a wide range of deployment options including Excel, Visual Basic and HMTL. Tiberius can read data from a wide variety of sources including Excel, SPSS, Access, text files, SQL Server and Oracle.

Its major drawback was the inflexibility in dealing with categorical factors as its neural network engine could only cope with numerical predictors.

b) Matlab's Neural Network Toolbox, Version R2011b

Matlab's Neural Network Toolbox runs within the framework of the Matlab computing environment. It is widely available within the University of Edinburgh computing systems, therefore would be at no extra cost. It also comes with the benefit of being a powerful data modelling platform with an abundance of learning resources online and in text books.

However, using Matlab requires learning of the Matlab programming language to be able to write scripts. With very little programming skills, this was not a plausible option within the three year timescale of the PhD.

Also, the Neural Network Toolbox was not easily customizable to accommodate different types of learning algorithms and activations functions.

APPENDIX C

c) Alyuda NeuroIntelligence by Alyuda Research Company

NeuroIntelligence (NI) boasts of a sophisticated automated 'best' network search as well as highly customisable design of neural network models. It can train an infinite number of networks, algorithms, topologies, at the same time and retain the best performing networks after testing. Retained networks can then be retrained with tweaks to their topology to produce even more superior results. NI includes the ability to perform classification, regression, time series analysis, and clustering problems.

It allows model saving options either within the NI framework or as XML codes output. At a single user cost of $497, NI has a very pleasing graphical user interface and offers enhanced features to visualise the neural network architecture search as well plot graphs of network error, error distribution and network comparisons.

d) Statistica by Statsoft. Inc, Version 10

Statistica 10 is a comprehensive data processing and visualisation software by Statsoft Inc. With advanced quality graphical user interface and a rather extensive choice of analysis available from classical statistics to fraud detection in insurance companies as well as six sigma and quality control in manufacturing industry. Statistica provides the framework for running neural network models, support vector machines, genetic algorithm, multiple regression, etc.

Like NI, its neural network engine allows for the development of highly customisable models using a comprehensive list of available network algorithms, activations functions and network architectures. It also incorporates advanced features for data bootstrapping, ensemble modelling, network pruning and weight decay regularisation and training momentum.

The complete single user license however costs £25,000 and allows exporting of developed models in different forms including SAS, HTML, C++, etc. Statsoft however offered the academic licence for this research at only £65. This however came with the limitation of not being able to export the final models out of Statistica after training.

APPENDIX C

Statistica 10 was however chosen for the modelling in this research because of its power, the extensive list of options available to experiment with, the relatively low cost of the academic license, its intuitive graphical user interface, ease of integration with Microsoft Suite and the readiness of the software developers to organise a free three-hour 'getting started' webinar session for this researcher.

Appendix D

Sample Model

(C# script)

```
/**C# deployment code of Neural Networks Model**/

/**================================================================================
Before running the C# deployment code please read the following.

 STATISTICA variable names will be exported as-is into the C# deployment script;
please verify the resulting script to ensure that the variable names follow the C#
naming conventions and modify the names if necessary.

================================================================================**/

using System;



public class Predict

{

   public static void Cost_Model_MLP_12_7_1( double[] ContInputs, string[] CatInputs )

   {

     //"Input Variable" comment is added besides Input(Response) variables.


     int Cont_idx=0;

     int Cat_idx=0;

     double zscores (TC) = ContInputs[Cont_idx++]; //Input Variable

     double zScoreD = ContInputs[Cont_idx++]; //Input Variable

     string Delivery Partner = CatInputs[Cat_idx++]; //Input Variable

     string Primary Purpose = CatInputs[Cat_idx++]; //Input Variable

     string Project Scope = CatInputs[Cat_idx++]; //Input Variable

    double[] __statist_max_input = new double[2];

    __statist_max_input[0]= 8.11505087458973e+000;

    __statist_max_input[1]= 3.85401568326401e+000;


    double[] __statist_min_input = new double[2];

    __statist_min_input[0]= -3.67929873414288e-001;

    __statist_min_input[1]= -1.47002847543739e+000;


    double[] __statist_max_target = new double[1];

    __statist_max_target[0]= 7.81817078097809e+000;


    double[] __statist_min_target = new double[1];

    __statist_min_target[0]= -3.66057073069406e-001;



    double[,] __statist_i_h_wts = new double[7,12];
```

```
    __statist_i_h_wts[0,0]=6.10055491663783e-001;
    __statist_i_h_wts[0,1]=-9.00972477126112e-001;
    __statist_i_h_wts[0,2]=-3.33284518737462e-002;
    __statist_i_h_wts[0,3]=1.33521888728242e-001;
    __statist_i_h_wts[0,4]=-7.97095096756099e-002;
    __statist_i_h_wts[0,5]=7.74460245945113e-003;
    __statist_i_h_wts[0,6]=1.38225161760593e-002;
    __statist_i_h_wts[0,7]=2.58369627399478e-002;
    __statist_i_h_wts[0,8]=-4.60467768382143e-002;
    __statist_i_h_wts[0,9]=-1.26530985813679e-001;
    __statist_i_h_wts[0,10]=4.39882491380093e-002;
    __statist_i_h_wts[0,11]=1.54680076500557e-001;

    __statist_i_h_wts[1,0]=6.27440133242177e-001;
    __statist_i_h_wts[1,1]=6.83694167293472e-001;
    __statist_i_h_wts[1,2]=3.88968243134368e-001;
    __statist_i_h_wts[1,3]=4.04882370941243e-001;
    __statist_i_h_wts[1,4]=-7.15758652960895e-001;
    __statist_i_h_wts[1,5]=4.74529478973639e-003;
    __statist_i_h_wts[1,6]=8.43968467935806e-002;
    __statist_i_h_wts[1,7]=4.03982710658986e-002;
    __statist_i_h_wts[1,8]=-3.42393892604500e-002;
    __statist_i_h_wts[1,9]=1.45222033503073e-001;
    __statist_i_h_wts[1,10]=-3.32770527874177e-002;
    __statist_i_h_wts[1,11]=5.04437710826594e-002;

    __statist_i_h_wts[2,0]=2.49868449644580e-001;
    __statist_i_h_wts[2,1]=1.10664796381584e+000;
    __statist_i_h_wts[2,2]=2.77178941395611e-001;
    __statist_i_h_wts[2,3]=-2.96398302172497e-001;
    __statist_i_h_wts[2,4]=-3.79932921247878e-002;
    __statist_i_h_wts[2,5]=-4.72844806928609e-002;
    __statist_i_h_wts[2,6]=-4.95942542229320e-002;
    __statist_i_h_wts[2,7]=7.47916931474163e-002;
    __statist_i_h_wts[2,8]=4.55203571106139e-002;
    __statist_i_h_wts[2,9]=1.90816788710656e-001;
```

```
    __statist_i_h_wts[2,10]=-5.52385545144320e-002;
    __statist_i_h_wts[2,11]=-2.68756551841026e-001;

    __statist_i_h_wts[3,0]=3.02266802032769e-001;
    __statist_i_h_wts[3,1]=-6.98225789278918e-001;
    __statist_i_h_wts[3,2]=-5.04665053972154e-001;
    __statist_i_h_wts[3,3]=3.54414687155054e-001;
    __statist_i_h_wts[3,4]=2.83227398405789e-001;
    __statist_i_h_wts[3,5]=1.93515508871481e-001;
    __statist_i_h_wts[3,6]=-1.86803602618070e-002;
    __statist_i_h_wts[3,7]=-1.19402899041314e-001;
    __statist_i_h_wts[3,8]=-2.17788004608143e-002;
    __statist_i_h_wts[3,9]=-3.99544482827499e-002;
    __statist_i_h_wts[3,10]=1.41799945313141e-001;
    __statist_i_h_wts[3,11]=2.98152008015934e-002;

    __statist_i_h_wts[4,0]=1.70332262296414e+000;
    __statist_i_h_wts[4,1]=-4.45278570986640e-001;
    __statist_i_h_wts[4,2]=4.71980700692331e-001;
    __statist_i_h_wts[4,3]=-5.61471080283562e-001;
    __statist_i_h_wts[4,4]=-1.43967443000545e-001;
    __statist_i_h_wts[4,5]=4.22522889623698e-002;
    __statist_i_h_wts[4,6]=1.61571557721163e-001;
    __statist_i_h_wts[4,7]=-3.98401318134162e-001;
    __statist_i_h_wts[4,8]=2.07133097825445e-002;
    __statist_i_h_wts[4,9]=5.11421161308927e-002;
    __statist_i_h_wts[4,10]=-8.80934161539714e-002;
    __statist_i_h_wts[4,11]=-8.58878454445226e-002;

    __statist_i_h_wts[5,0]=-2.99001956571505e-001;
    __statist_i_h_wts[5,1]=-7.58160322104295e-001;
    __statist_i_h_wts[5,2]=2.45732161474987e-001;
    __statist_i_h_wts[5,3]=2.63716263798594e-001;
    __statist_i_h_wts[5,4]=-4.52811644304255e-001;
    __statist_i_h_wts[5,5]=1.06761349175363e-002;
    __statist_i_h_wts[5,6]=3.85985335722588e-003;
    __statist_i_h_wts[5,7]=4.29562594409285e-002;
```

```
    __statist_i_h_wts[5,8]=-7.33927726395723e-002;
    __statist_i_h_wts[5,9]=-5.58912058926934e-002;
    __statist_i_h_wts[5,10]=4.01575249383282e-002;
    __statist_i_h_wts[5,11]=1.30642597104837e-001;

    __statist_i_h_wts[6,0]=7.00450325945085e-001;
    __statist_i_h_wts[6,1]=-4.17094331074114e-001;
    __statist_i_h_wts[6,2]=4.17885284201890e-001;
    __statist_i_h_wts[6,3]=-3.87113634738378e-001;
    __statist_i_h_wts[6,4]=-1.39886083479381e-001;
    __statist_i_h_wts[6,5]=-1.13066359051156e-001;
    __statist_i_h_wts[6,6]=-1.35192016624967e-001;
    __statist_i_h_wts[6,7]=1.18273653382723e-001;
    __statist_i_h_wts[6,8]=-4.54577204692083e-002;
    __statist_i_h_wts[6,9]=-8.09074573683310e-002;
    __statist_i_h_wts[6,10]=-4.82791368359781e-002;
    __statist_i_h_wts[6,11]=3.61311055362202e-002;

    double[,] __statist_h_o_wts = new double[1,7];

    __statist_h_o_wts[0,0]=7.56888846091610e-001;
    __statist_h_o_wts[0,1]=3.32137071440306e-001;
    __statist_h_o_wts[0,2]=2.41848324122223e-001;
    __statist_h_o_wts[0,3]=1.53626552809784e-001;
    __statist_h_o_wts[0,4]=3.86585621330931e-002;
    __statist_h_o_wts[0,5]=-6.02101745772497e-001;
    __statist_h_o_wts[0,6]=1.08390122143369e-001;

    double[] __statist_hidden_bias = new double[7];
    __statist_hidden_bias[0]=6.26313425097254e-002;
    __statist_hidden_bias[1]=8.52788624793290e-002;
    __statist_hidden_bias[2]=-8.93609753819305e-002;
    __statist_hidden_bias[3]=1.02705157732035e-001;
    __statist_hidden_bias[4]=-2.09456901774407e-001;
    __statist_hidden_bias[5]=3.46305548973789e-002;
    __statist_hidden_bias[6]=-1.30136072417104e-001;
```

```
    double[] __statist_output_bias = new double[1];
    __statist_output_bias[0]=-3.76993539230292e-002;

    double[] __statist_inputs = new double[12];

    double[] __statist_hidden = new double[7];

    double[] __statist_outputs = new double[1];
    __statist_outputs[0] = -1.0e+307;

    __statist_inputs[0]=zscores (TC);
    __statist_inputs[1]=zScoreD;

    if( Delivery Partner=="CID")
    {
     __statist_inputs[2]=1;
    }
    else
    {
     __statist_inputs[2]=0;
    }

    if( Delivery Partner=="SWD")
    {
     __statist_inputs[3]=1;
    }
    else
    {
     __statist_inputs[3]=0;
    }

    if( Delivery Partner=="SWS")
    {
     __statist_inputs[4]=1;
    }
    else
    {
```

```
 __statist_inputs[4]=0;
}

if( Primary Purpose=="GENERAL")
{
 __statist_inputs[5]=1;
}
else
{
 __statist_inputs[5]=0;
}

if( Primary Purpose=="WASTEWATER")
{
 __statist_inputs[6]=1;
}
else
{
 __statist_inputs[6]=0;
}

if( Primary Purpose=="WATER")
{
 __statist_inputs[7]=1;
}
else
{
 __statist_inputs[7]=0;
}

if( Project Scope=="NEWBUILD")
{
 __statist_inputs[8]=1;
}
else
{
 __statist_inputs[8]=0;
```

```
    }

    if( Project Scope=="REFURB")
    {
     __statist_inputs[9]=1;
    }
    else
    {
     __statist_inputs[9]=0;
    }

    if( Project Scope=="REPLACE")
    {
     __statist_inputs[10]=1;
    }
    else
    {
     __statist_inputs[10]=0;
    }

    if( Project Scope=="UPGRADE")
    {
     __statist_inputs[11]=1;
    }
    else
    {
     __statist_inputs[11]=0;
    }

    double __statist_delta=0;
    double __statist_maximum=1;
    double __statist_minimum=0;
    int __statist_ncont_inputs=2;

    /*scale continuous inputs*/
    for(int __statist_i=0;__statist_i < __statist_ncont_inputs;__statist_i++)
    {
```

```
 __statist_delta = (__statist_maximum-__statist_minimum)/(__statist_max_input
[__statist_i]-__statist_min_input[__statist_i]);

 __statist_inputs[__statist_i] = __statist_minimum - __statist_delta*
__statist_min_input[__statist_i]+ __statist_delta*__statist_inputs[__statist_i];

}


int __statist_ninputs=12;

int __statist_nhidden=7;


/*Compute feed forward signals from Input layer to hidden layer*/

for(int __statist_row=0;__statist_row < __statist_nhidden;__statist_row++)

{

  __statist_hidden[__statist_row]=0.0;

  for(int __statist_col=0;__statist_col < __statist_ninputs;__statist_col++)

  {

   __statist_hidden[__statist_row]= __statist_hidden[__statist_row] +
(__statist_i_h_wts[__statist_row,__statist_col]*__statist_inputs[__statist_col]);

  }

 __statist_hidden[__statist_row]=__statist_hidden[__statist_row]+__statist_hidden_bias
[__statist_row];

}


for(int __statist_row=0;__statist_row < __statist_nhidden;__statist_row++)

{

  if(__statist_hidden[__statist_row]>100.0)

  {

   __statist_hidden[__statist_row] = 1.0;

  }

  else

  {

   if(__statist_hidden[__statist_row]<-100.0)

   {

    __statist_hidden[__statist_row] = -1.0;

   }

   else

   {

    __statist_hidden[__statist_row] = Math.Tanh(__statist_hidden[__statist_row]);

   }

  }
```

```
    }


    int __statist_noutputs=1;


    /*Compute feed forward signals from hidden layer to output layer*/

    for(int __statist_row2=0;__statist_row2 < __statist_noutputs;__statist_row2++)

    {

     __statist_outputs[__statist_row2]=0.0;

    for(int __statist_col2=0;__statist_col2 < __statist_nhidden;__statist_col2++)

      {

       __statist_outputs[__statist_row2]= __statist_outputs[__statist_row2] +
    (__statist_h_o_wts[__statist_row2,__statist_col2]*__statist_hidden[__statist_col2]);

      }

     __statist_outputs[__statist_row2]=__statist_outputs[__statist_row2]+
    __statist_output_bias[__statist_row2];

    }







    /*Unscale continuous targets*/

    __statist_delta=0;

    for(int __statist_i=0;__statist_i < __statist_noutputs;__statist_i++)

    {

     __statist_delta = (__statist_maximum-__statist_minimum)/(__statist_max_target
    [__statist_i]-__statist_min_target[__statist_i]);

     __statist_outputs[__statist_i] = (__statist_outputs[__statist_i] - __statist_minimum +
     __statist_delta*__statist_min_target[__statist_i])/__statist_delta;

    }




      for(int __statist_ii=0; __statist_ii < __statist_noutputs; __statist_ii++)

      {

        Console.WriteLine(" Prediction{0} = {1}", __statist_ii+1, __statist_outputs
    [__statist_ii]);

      }



   }
```

```
  public static void Main (string[] args) {
    int argID = 0;
    double[] ContInputs = new double[2];
    int contID = 0;
    string[] CatInputs = new string[3];
    int catID = 0;

    if (args.Length >= 5)
    {
      ContInputs[contID++] = Double.Parse(args[argID++]);
      ContInputs[contID++] = Double.Parse(args[argID++]);
      CatInputs[catID++] = args[argID++];
      CatInputs[catID++] = args[argID++];
      CatInputs[catID++] = args[argID++];
    }
    else
    {
      string Comment = "";
      string Comment1 = "*****************************************************************
*********\n";
      Comment += Comment1;
      string Comment2 = "Please enter at least 5 command line parameters in the following
 order for \nthe program to Predict.\n";
      Comment += Comment2;
      Comment += Comment1;
      string Comment3 = "zscores (TC)  Type - double (or) integer\n";
      Comment += Comment3;
      string Comment4 = "zScoreD  Type - double (or) integer\n";
      Comment += Comment4;
      string Comment5 = "Delivery Partner  Type - String (categories are { \"CID\"  \"SWD\
"  \"SWS\" } )\n";
      Comment += Comment5;
      string Comment6 = "Primary Purpose  Type - String (categories are { \"GENERAL\"  \
"WASTEWATER\"  \"WATER\" } )\n";
      Comment += Comment6;
      string Comment7 = "Project Scope  Type - String (categories are { \"NEWBUILD\"  \
"REFURB\"  \"REPLACE\"  \"UPGRADE\" } )\n";
      Comment += Comment7;
      Comment += Comment1;
      System.Console.WriteLine(Comment);
```

```
        System.Environment.Exit(1);
      }
      Cost_Model_MLP_12_7_1( ContInputs, CatInputs );
    }

}
```