A conceptual cost estimation model for the pre-design stage of road projects in New Zealand

A thesis presented in partial fulfilment of

the requirements for the degree of

Doctor of Philosophy (PhD)

in

Construction

School of Built Environment

Massey University

Auckland, New Zealand.

Chinthaka Niroshan Atapattu

2023

Copyright

Statement of Originality

I declare that this thesis is my own work, except where due acknowledgement is made, and that it has not been previously included in a thesis, dissertation or report submitted to this University or any other institution for the degree or any other qualification.

…….

Chinthaka Atapattu

Abstract

Cost overruns in construction projects have been a pervasive and challenging issue within the construction industry. Over the years, the researchers explored the causes, effects, and mitigation strategies for the cost overrun issue. However, there has been no major improvement, and projects still experience significant cost overruns. Due to the long project duration, scope, higher coverage of ground, and involvement of public funds, road construction faces more uncertainty compared to building projects. Thus, the cost overrun in road projects has become a crucial issue. Therefore, this study focused on mitigating the cost overrun issue by minimising the errors in conceptual cost estimation by developing a new cost model for the pre-design stage of road projects.

This research followed a secondary data collection through systematic and bibliometric literature review and primary data collection from the New Zealand (NZ) Road construction industry to fill the above research gap and achieve the research aim. The multi-method quantitative research approach, including a questionnaire survey and multiple case study data, was used in this research. Initial cost overrun evaluation used one hundred and six cases from NZ road projects, while fifty-nine detailed cases were utilised to develop and validate the cost model.

Firstly, the research identified ten crucial factors that affect the cost of NZ road projects: frequent design changes, poor planning and scheduling, poor and incomplete tender documentation, delays in design, mistakes/ errors in design and drawings, unforeseen ground conditions, inaccurate cost estimates, poor site management and supervision, poor project management, and inaccurate quantity take-off.

Secondly, the research investigated the severity of the cost overruns in NZ road projects. According to the findings, NZ road projects experience approximately 20% of cost overruns, while the project size and duration significantly impact the magnitude of the cost overruns. Thirdly, using the case study data, the research developed three models to improve the conceptual cost

estimation. Since the conceptual estimate is prepared during the pre-design stage, the variables should be calculated using less information. The performance of the models was evaluated as an error percentage comparing the estimated cost from the model with the actual project cost. The measure was the Mean Absolute Percentage Error (MAPE). The first model was developed using regression analysis (RA), which showed a MAPE of 21.35%. Then, another model was developed using artificial neural networks (ANN) with a MAPE of 11.82%. However, both models considered only the technical aspects of the projects. Therefore, the final model combined ANN with Monte Carlo simulation (MCS). The hybrid model demonstrated a minimum MAPE of 3.53%, significantly improved results compared to the continuous cost overruns experienced in the NZ road construction section.

The findings of this research greatly support the NZ road construction sector to be aware of the severity of the cost overrun issue. Further, the model introduced in the research can be used in the industry to improve the current estimation practices.

The study was conducted using the case study data from NZ road projects. However, the variables considered for the models were selected with the possibility to generalise the mode to the other countries. Further, the findings and conclusions were limited to the modelling techniques identified in this research.

Although other hybrid models are available, very little research was conducted on developing cost estimations for road projects. Most of the developed cost models considered technical characteristics in a project but did not consider any project risk factors. There were a few models considered both using only one modelling technique but failed to produce reliable output. Therefore, hybrid models can be developed by combining several techniques without mixing the technical and risk variables. This research developed the first ANN and MCS hybrid model, considering technical characteristics and risk factors with a significantly lower error.

Ethics Approval

Massey University Human Ethics Committee (MUHEC) granted 'Low-Risk Notification" to this research project. Such approval was granted on 29th July 2020, under Ethics Notification Number 4000023063, for the study. The Approval letter is attached under Appendix A.

Dedication

This thesis is dedicated to

my wife, ***Yasodha Ranathunga,***

my son, ***Thivain Atapattu,***

and my parents, ***Nanda Premathilaka*** *and* ***Yasapala Atapattu***

Acknowledgements

I would like to express my gratitude to my main supervisor, Dr Niluka Domingo, Senior Lecturer of the School of Built Environment (SBE), for her continued support during my PhD journey. It has been an absolute honour for me to learn from her. I am so delighted to acknowledge her steadfast support, patience, constructive feedback, and continuous progress meetings to support me in staying focused. In addition, I would like to thank my co-supervisor, Professor Monty Sutrisna, Head of SBE, for his comments, motivations, and academic advice. His excellent knowledge and experiences have always been highly encouraging, and I appreciate his mentorship approach through discussions. High insights and timely communications from both my supervisors always encouraged me to step ahead, resulting in this achievement. Thank you for all your support.

My profound respect and gratitude also extended to the professionals who supported me in my data collection. Mainly, the participants of my questionnaire survey and the professionals who helped me collect the cost data of the completed New Zealand Road projects should be highly acknowledged. Also, I would like to acknowledge the guidance and support of Dr. Heshani Edirisinghe to make my research a successful one.

I am deeply grateful to my wonderful family for their continuous encouragement, support, and patience throughout my PhD journey. I am grateful for having a very kind and lovely son, Thivain Atapattu, for being so joyful, kind and patient with me. Above all, my lovely wife, Yasodha Ranathunga, for her endless love, continuous support, and sacrifice in making this journey a success. I also acknowledge the support of my lovely mum, Nanda Premathilaka, and my dad, Yasapala Atapattu, for their love and affection throughout my life and for guiding me in becoming the man I am today. I also acknowledge the love and support given by my wife's parents, Anura Ranathunga and Rathnakanthi Weerakoon.

I would also like to acknowledge the continuous support and encouragement given by the Academic staff of the SBE in my research and teaching. Also, I would like to recognise the past and present administrative staff of SBE for their kind support towards ensuring a smooth administrative running of my PhD. Finally, my sincere gratitude is extended to my lovely colleagues Ravindu Kahandawa, Vishan Weerasekara, Nuwan Sampath, Achini Weerasinghe, Nishadi Sooriyamudalige, An Le, and Alice Bui for being there with me throughout the last three and half years of my PhD journey and whose kind words always encouraged me when I felt weary.

Table of Content

List of Tables

List of Figures

Abbreviations and Acronyms

AACE	American Association of Cost Engineers - International
AC	Actual Cost
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AT	Auckland Transport
ATAP	Auckland Transport Alignment Project
CBA	Cost Benefit Analysis
CBR	Case-based Reasoning
CCI	Construction Cost Index
COV	Coefficient of Variation
LGWM	Let's Get Wellington Moving
MAPE	Mean Absolute Percentage Error
MCS	Monte Carlo simulation
MLP	Multiple Perceptron
MLRA	Multiple Linear Regression Analysis
MPCO	Mean Percentage of Cost Overrun
MPE	Mean Percentage Error
NZ	New Zealand
NZIQS	New Zealand Institute of Quantity Surveyors
NZTA	New Zealand Transport Agency
PC	Predicted Cost
RA	Regression Analysis
RBF	Radial Basis Function
RCF	Reference Class Forecasting
RMSE	Root Mean Squared Error
SD	Standard Deviation
SI	Severity Index
SLR	Systematic Literature Review
SVM	Support Vector Machine
TI	Transportation Infrastructure
VIF	Variance Inflation Factors

List of Peer-reviewed Publications

Chapter number	Publication reference	Publication type	Status
3	Atapattu, C. N., Domingo, N. and Sutrisna, M. (2023). Significant factors affecting the New Zealand transportation infrastructure project cost – quantity surveyors' perception, *Built Environment Project and Asset Management*, Vol. 13 No. 5, pp. 756-777. DOI: https://doi.org/10.1108/BEPAM-07-2022-0105	Journal paper	Published
	Atapattu, C. N., Domingo, N. and Sutrisna, M. (2022). Causes and effects of cost overruns in construction projects. The 45th Australasian Universities Building Education Association Conference (AUBEA 2022), Sydney, Australia, 23-25 Nov. DOI: https://doi.org/10.26183/a6pq-mg06	Conference paper	Published
4	Atapattu, C. N., Domingo, N. and Sutrisna, M. (2023). How significant is the cost overrun in road projects in New Zealand? *Developments in the Built Environment*.	Journal paper	Under review
5	Atapattu, C. N., Domingo, N. and Sutrisna, M. (2023). A bibliometric review of statistical modelling techniques for cost estimation of infrastructure projects. *Smart and Sustainable Built Environment*, (Ahead of print) DOI: https://doi.org/10.1108/SASBE-01-2023-0005	Journal paper	Published
	Atapattu, C. N., Domingo, N. and Sutrisna, M. (2022). Statistical cost modelling for preliminary stage cost estimation of infrastructure projects. IOP Conference Series: Earth and Environmental Science, Vol. 1101, pp. 052031. DOI: https://doi.org/10.1088/1755-1315/1101/5/052031	Conference paper	Published
6	Atapattu, C. N., Domingo, N. D. and Sutrisna, M. (2023). A conceptual cost estimation model for pre-design stage of road projects in New Zealand using multiple regression analysis. *Journal of Financial Management of Property and Construction*. Manuscript ID: JFMPC-08-2023-0052R1	Journal paper	Under review
7	Atapattu, C. N., Domingo, N. D. and Sutrisna, M. (2023). A conceptual cost estimation model for the pre-design stage of road projects in New Zealand using Artificial Neural Networks. *Journal of Construction, Engineering and Management*. Manuscript ID: COENG-14398R1	Journal paper	Under review
8	Atapattu, C. N., Domingo, N. D. and Sutrisna, M. (2023). A Hybrid model for conceptual cost estimation at pre-design stage of road projects – A hybrid between artificial neural networks, and Monte Carlo simulation, *Sustainable Cities and Society*.	Journal paper	Manuscript is ready

1 Introduction

1.1 Prologue

This chapter gives the introduction to the thesis. Firstly, the background of the research is provided, this is followed by a research gap analysis. After that, the following sections discuss the research problem and the research rationale, developed in accordance with the thesis aim. The thesis aim is later subdivided into respective research questions and research objectives. Finally, the chapter briefly explains the thesis structure and how the publication outputs connect and answer the research questions.

1.2 Background

The construction industry in New Zealand (NZ) is crucial to the country's economy in terms of Gross Domestic Product (GDP) and employment (PWC, 2022). The construction sector contributed to 6.7% of the GDP, approximately 18.1 billion NZ Dollars (NZD) in the year ending March 2022 (MBIE, 2022). In addition, the report also stated that 295,200 workers (1.5% out of the total workforce) and 70,629 enterprises (12.6%) contributed to the construction sector.

NZTA (2020) reported that 17 road projects in NZ delivered between 2017 and 2020 faced overall cost overruns of 1.1 billion NZD. The NZTA report further stated that out of these 17 projects, the minimum cost overrun experienced by a project was 8%, while the maximum cost overrun was 147%. NZTA emphasised that the cost overrun value is approximately equal to the entire annual budget for state highway improvements. According to recent reports, the Auckland City Link rail project was significantly over the budget (www.citylink.co.nz). The City Link website reported that the estimate in 2014 was $3.4 billion, which increased to $4.419 billion by 2019 (www.citylink.co.nz). Therefore, the cost overrun in this project was more than $1 billion. The same website also reported that the total infrastructure work in Australasia in 2018 was $80 billion.

However, by 2019, the total cost overruns experienced by all the above projects were $230 billion (www.citylink.co.nz).

Subsequently, the cost overruns can severely impact other budgetary plans. If the additional risks and cost overruns can be minimised or forecasted during the conceptual estimate, the decision-making process of the project initiation would be more fruitful, and the project would be planned effectively to sustain any unexpected financial blowouts.

According to the New Zealand Transport Agency (NZTA), for the decade from 2021-2031, there are four significant investment projects in motion, namely, 'The Auckland Transport Alignment' project (ATAP), 'Let's get Wellington moving' (LGWM), the 'road to zero' strategy, and the NZ new rail plan (NZTA, 2020). According to reports, all four projects above were allocated approximately 32 billion NZD (NZTA, 2020). Further, more than 2700 transportation infrastructure projects worth a total of 65.8 billion NZD approximately have been planned for the next decade by the government (Infrastructure Commission, 2022)

Therefore, the NZ government has allocated significant investment into road development. However, if the initial investment is poorly calculated, the funding allocation and decision-making will also be affected. However, according to a project management survey by KPMG-NZ (2017), 71% of New Zealand projects face cost overruns, significantly impacting the NZ economy. Furthermore, KPMG NZ published some critical insights into the projects completed in NZ in 2010, 2013 and 2017 (refer to Table 1.1).

Table 1.1. *Project management survey findings (KPMG, 2010, 2013, 2017)*

	2010	**2013**	**2017**
Projects delivered on budget	48%	33%	29%
Project delivered on time	36%	29%	31%
Projects delivered on scope	59%	35%	33%

According to the findings in Table 1.1, in NZ, the number of projects delivered within the budget has decreased significantly over the years. Furthermore, as mentioned above, NZ transportation

infrastructure is planned to substantially contribute to the national economy due to significant investment in road construction, with the vital involvement of NZTA and Auckland Transport (AT). Nevertheless, according to the findings in Table 1.1, it seems inevitable that the projects will experience cost overruns. However, this assumption must be tested with actual data before making conclusions. Therefore, it is essential to ensure the funding is well-spent as it is mainly public funds. Based on the above discussion, the following section confers the research gap analysis of this thesis.

1.3 Problem statement

According to the American Association of Cost Engineering – International (AACE), conceptual estimates belong to class 5 of the estimate classification (AACE, 2020). The project definition is at 0% to 2% level at this stage. That means the availability of information is deficient. Therefore, based on the traditional estimate practices, the accuracy of the conceptual estimate will vary between a lower margin of -20% to -50% and an upper margin of +30% to 100%. Consequently, the traditional estimation practices will cause more cost overruns due to the higher percentage of uncertainty in the estimate. Additionally, the conventional estimation process requires a significant amount of manual work and expert judgement, which leads to less accuracy and reliability (Wang et al., 2021).

Therefore, the cost estimation needs automation to improve accuracy because the automated method will investigate the project's nature and characteristics, comparing them with the completed projects (Choi, et al., 2015; Kim et al., 2012). Unlike manual methods, automated estimation methods will have fewer errors due to human judgement (Choi, et al., 2015; Salmi et al., 2016). Therefore, cost modelling has become an emerging and highly demanded research area.

On the other hand, the manual and current cost estimation process primarily focuses on the technical variables of the project. However, there are innumerable uncertain factors that a project

faces throughout the design, tendering, and construction phases (Choi, et al., 2015; Kim et al., 2012; Salmi et al., 2016). That is even more crucial for infrastructure projects than building projects, considering their geographical spread, longer durations, and higher interaction with the public (Grimsey and Lewis, 2002; Siraj and Fayek, 2019; Vickerman, 2007).

Even though infrastructure projects incur massive funding, a high level of activities, and significantly longer durations, the existing empirical knowledge on the costs and risks involved is considerably lacking (Flyvbjerg et al., 2003; Tan and Zhao, 2019). Therefore, considering the background and the problem statement, the following section analyses in-depth the research gap relevant to cost estimation issues in infrastructure projects.

1.4 Research gap

With the increase in urbanisation and industrialisation, transportation infrastructure has become a fundamental requirement in recent years. Therefore, in emerging economies, transportation infrastructure development plays a significant role (Andrić et al. 2019; Daniel and Kumar, 2023). Andrić et al. (2019) explained that due to various reasons such as complexity, large scale, longer construction durations, higher investments, geographical spread over larger site areas and varying site conditions, transportation infrastructure projects face much higher uncertainties than vertical sector projects. However, finishing the project within the allocated budget is a crucial factor in defining the project's success (Kim et al., 2004; Ika and Pinto, 2022).

Lee (2008) investigated cost overruns using 138 road projects in Korea. The researcher observed that 82% of the projects face a maximum of 50% cost overruns. Further, Flyvbjerg et al. (2003) investigation found that nine out of ten projects face significant cost overruns. In addition to that, road projects globally would meet 20% of average cost overruns (Flyvbjerg et al., 2003). Since roads and other transportation infrastructure projects gain large amounts of funding, the cost overruns, as mentioned above, can be crucial for the funding parties. Most of the infrastructure

projects are primarily funded by the public sector. Hence, if the budget was not initially planned accurately, cost overruns will draw unexpected additional funds from the public sector. Nevertheless, Flyvbjerg et al. (2003) also found that the cost overrun issue has remained unchanged over the past 70 years with no improvement. Consequently, the estimates used for critical decision-making and budget allocation process of infrastructure projects are mostly unreliable.

Looking at construction projects worldwide, it is evident that only a little research has been done to study cost overruns in road projects. Further, Ismail et al. (2021) conducted a systematic literature review of cost overrun research in construction projects in developed and developing countries. According to their findings, during the 2010 – 2020 period, out of 152 research studies into cost overruns, only 9% related to road projects in developed countries. Further, out of the 9%, no research was done to study the cost overruns in any NZ construction project types. Table 1.2 summarises the most recent research done on cost overruns. According to Table 1.2, all the research was mainly on developing countries, and the majority were from African countries (Amadi and Higham, 2019; Ammar et al., 2022; Akinradewo et al., 2022; Mahmud et al., 2021; Melaku Belay et al., 2021). Therefore, the findings of recent studies cannot apply to the NZ context, as the authors mentioned that the findings can only be generalised to similar developing countries. Much research was done to study the factors affecting cost overruns (Al Hosani et al., 2020; Ammar et al., 2022; Andrić et al., 2019; Anish et al., 2019; Akinradewo et al., 2022; Huo et al., 2018; Mahmud et al., 2021; Sohu et al., 2017). However, only a few research studies have been conducted on finding solutions to cost overruns through addressing the issues of cost estimating. Various models were developed to address the cost estimating issues (Cirilovic et al., 2014; El-Kholy, 2019; Jaafari et al., 2021; Mahalakshmi and Rajasekaran, 2018). Nevertheless, such models only focused on the tender stage or post-contract stages. To overcome the cost overruns, the

conceptual estimate must be accurate and reliable as it is used for budget allocations and decision-making.

Table 1.2. *Recent research conducted on cost overruns*

Reference	Project type	Location	Scope of the paper	Research gap/ limitations
Ammar et al. (2022)	Roads	Egypt	To identify the factors that significantly contribute to cost overrun for road projects in Egypt during the implementation phase.	Only 75 projects were used to analyse the cost overruns. The focus is only on factors affecting the implementation phase.
Akinradewo et al. (2022)	Roads	Ghana	To identify the crucial variables of cost overruns	The study was carried out only in a specific region in Ghana. Hence, the findings were not generalised to other contexts
Heravi and Mohammadian (2021)	Roads and buildings	Iran	To investigate how the delay and cost overruns depend on planned time, cost, type, and nature of the projects	To generalise the findings the paper suggested conducting more case studies.
Mahmud et al. (2021)	Roads	Nigeria	To explore the driving factors of cost overrun.	The sample size is small – 16. The findings can only be generalised to the other Nigerian context or similar developing countries.
Melaku Belay et al. (2021)	Roads and buildings	Ethiopia	To explore the extent of cost overrun and schedule delays	The cost overrun effect on the project size was not considered since all the contract values were in the same range.
Al Hosani et al. (2020)	Roads	UAE	to generate an understanding of the causes and possible solutions to cost overruns as a form of project failure in the UAE	The study only focussed on the identification of the causes.
Calahorra-Jimenez et al. (2020)	Roads	Chile	To propose a structured approach to provide measures based on design-build practices that can help design-bid-built road projects to mitigate the principal reasons that lead to cost overruns	The results were not validated properly. The international research findings were used to triangulate the findings and suggest conducting more case studies to analyse the topic in-depth.
Paraskevopoulou and Boutsis (2020).	Tunnel	UK	To investigate the impact of unforeseen geological conditions due to the limited geological investigation that may lead to poor and unsound designs.	The dataset was not strong due to the reluctance of the parties to give the details.
Amadi and Higham (2019).	Roads	Nigeria	To investigate the lack of adherence to geotechnical best practices from a logical theoretical perspective, it is necessary to understand the cost overruns experienced in delivering highway projects in the Niger Delta.	The study emphasised the necessity of further studies on conceptual cost estimation of highway projects. The authors noted that the findings can be generalised to developing countries only.
Andrić et al. (2019)	Infrastructure	Asia	To determine the cost performance of infrastructure projects, to examine the impact of project type, size and	102 cases were selected covering roads, railways, and energy projects through the

			implementation period on cost overrun and to identify the causes of cost overruns	Asian Development Bank (ADB)
Anish et al. (2019)	Roads	India	To analyse factors leading to time delay and cost overrun	The factors were identified. However, the way forward to overcome cost overruns was not explored.
Huo et al. (2018)	Infrastructure	Hong Kong	to examine the characteristics of the cost overruns in mega-infrastructure projects	The study only focussed on the impact of cost overruns based on project type, implementation period and project size. Further, the findings were based on a case study from Hong Kong. The number of cases was only 57, covering roads, rails, bridges, and tunnels.
Sohu et al. (2017)	Roads	Pakistan	To find the major causes of cost overrun and to determine possible mitigation measures from experienced respondents for the identified causes.	Mitigation measures are identified for the causes. However, if the cause cannot be mitigated, then there will be a financial impact on the project. That gap is not covered.

Table 1.3 shows the fundamental research conducted in NZ to address issues related to cost estimation and overruns. However, none of these were focused on infrastructure projects.

Table 1.3. *Research conducted in NZ related to construction cost*

	Citation	Topic	Focus
1	Ji et al. (2014)	Factors influencing the accuracy of pre-contract stage estimation of final contract price in New Zealand	This study focuses on identifying the risk factors surrounding the tender estimation.
2	Adafin et al. (2020)	An evaluation of risk factors impacting project budget performance in New Zealand	Although the study focusses on identifying the risk factors affecting the project budget the focus is on general construction project. But to develop a cost model, the factors must be specific to the transportation infrastructure projects.
3	Zhao et al. (2019)	Forecasting residential building costs in New Zealand using a univariate approach	The research was done on residential projects. The model cannot be directly applied to road projects sue to differences between the two project types.

In light of this lack of suitable research, this thesis investigates project cost issues in NZ transportation infrastructure projects to fill the research gap discussed above. The following section discusses the scope and limitations to narrow the broad research gap.

1.5 Research scope and limitations

According to the research gap analysis, no research has been done to study and find solutions for cost overruns in NZ transportation infrastructure. However, it is too broad to cover all the TI project types in one research. Therefore, this research focussed on NZ road projects only. First, the research explores the causes of the cost overruns. Then, this study further analysed how significant the cost overruns are in NZ road projects. Later, based on the findings, a solution was provided to overcome the cost overrun problem in NZ road projects. The conclusions chapter concludes the overall limitations in detail (Chapter 10).

The research scope is described in detail in the chapter 2. The aim of the research, as specified above, is to improve the current estimation process, and the developed model is to be applied to NZ road projects. To achieve the aim, this research adopted the quantitative research methodology (questionnaire survey and project cost document analysis). The road projects in NZ and the selected quantitative research approach form the unit of analysis and observations in this study. Considering the research scope, the collected data for this project covers various locations and regions throughout NZ. Therefore, the model developed in this research applies to any geographical location within NZ. Further, the study is limited to new constructions and significant alteration projects and data collected from projects completed between 2002 and 2022. For the model development, the study used forty-three cases and sixteen more cases for model validation. Subsequently, the study was conducted using the primary data collected within the NZ context only.

Since the research aims to provide a solution to the conceptual estimation, it is necessary to briefly understand the estimation stages and their accuracy levels. This understanding enables us to connect the cost overruns with the estimation issues.

The following sub-section discusses the estimation classes in a construction project.

1.5.1 Construction project estimation classification

The American Association of Cost Engineering: International classifies construction project estimation into five classes (AACE, 2020). First, the class 01 estimate is a detailed estimate prepared as a check estimate. At this stage, most of the information is available. This estimate may be prepared for a part of the project to compare with the tender estimate or the contract value. This estimation is used for subcontracting, claims, variations, and dispute resolution. Thus, at this stage, the estimation method is a detailed unit rate method based on detailed quantity take-off. However, the estimation cannot be 100% accurate. According to AACE (2020), an estimate would be between -3% to -10% of underestimation and +3% to +15% over estimation.

Secondly, the class 2 estimate is the final tender estimate that would become the contract value of the project. At this stage, based on the procurement strategy and the contract type, the availability of information would be between 30% and 70%. Also, at this stage, the estimation method is detailed unit rate build-up using the quantity take-off. In this class, the estimation would have underestimation from -5% to -15% or overestimation from +5% to +20%.

Next, the class 3 estimate is prepared for budget authorisation and funding. At this stage, the project is defined up to 10% to 40%. Therefore, the detailed unit rate is not feasible for the entire project. Hence, the estimation method is semi-detailed unit cost. Consequently, this estimate class shows an underestimation between -10% and -20% and overestimation between +10% to +30%.

Class 4, the conceptual estimate is prepared for feasibility studies of the project when the project is defined only between 1% and 15%. Therefore, equipment factored or parametric models can be used to calculate the cost at this stage.

Finally, the class 5 estimate is the conceptual estimate prepared at the beginning of the project with 0% to 2% project information. The client's requirements are the only information available at this stage. Therefore, the commonly used estimation methods at this stage are capacity factoring, parametric models, judgement, or analogy methods. Due to the lack of information, this estimate may have the highest error compared to other classes. Further, the estimate can be -20% to -50% underestimated to +30% to +100% overestimated. Therefore, it is evidenced that current conceptual estimation methods need upgrading to overcome these massive errors, which is the focus of this research. Within the research scope, it is clear which estimation class this research targets. The following sub-section explains the road classification in NZ.

1.5.2 Road classifications

New Zealand Transport Agency (NZTA) has a road classification system based on the road's function (NZTA, 2018). The first category is the 'national roads', which make the most significant contribution to the NZ economy. These roads connect the major cities with high populations, major ports, and international airports and have a significant volume of traffic with general and commercial vehicles. The second category is 'regional roads', which connect the regionally significant places and contribute to the regional economy. Next, the third category is 'arterial roads', which also contribute substantially to social and economic well-being and link significant places within the region. This road type may have heavy traffic movement within the urban areas. The fourth category is the 'primary collector'. These roads connect the urban areas within a particular region with moderate traffic. The fifth category is 'secondary collector', which consists of roads that connect local areas that contain daily traffic and moderate traffic of commercial vehicles. The final road type is 'access roads', which provide access and connections to day-to-day travelling for houses, shops, and schools.

In addition to the above, two more road types appear in the One Network Road Classification Map produced by NZTA (https://nzta.maps.archgis.com). The first type is the high-volume roads, which generally comprise state highways. The second type is the low-volume roads, which are below the access road category.

In contrast, Auckland Transport (AT) provides a different classification. However, there are similarities between the two classifications. According to AT, there are two major categories: 'arterial' and 'non-arterial' roads. Arterial roads are further categorised into four types: motorways, strategic routes, primary arterials, and secondary arterials. Non-arterials are subdivided into collector roads, local streets, lanes and service lanes, and shared spaces/ zones.

Figure 1.1 shows the NZTA road classification more clearly. As specified in the above discussion, this research considers only arterial roads, those that make a significant contribution to the national economy, compared to non-arterial roads. Therefore, regarding NZTA classification, this research collected data from high-volume, national, regional, and arterial road types. On the other hand, based on AT classification, this research considered motorways, strategic, primary, and secondary arterial road types. The following section discusses the problem statement of this research.

***Figure 1.1** - NZ Road Classification (Source: NZTA, 2018)*

***Legend for Figure 1.1:** red – national; brown – regional; dark blue – arterial; light blue – primary collector; dark green – secondary collector; and light green – access roads*

1.6 Research rationale

Appraisal of the research background and problem statement indicates the necessity of addressing the risk factors associated with cost estimation, requiring a proper cost model comprised of reliable statistical analysis techniques for early design stage conceptual cost estimation of road projects in NZ. This research thesis addresses the contextual background and theoretical concepts surrounding the cost overruns, cost-related risk factors affecting the project cost, and cost modelling for NZ road projects by addressing the following needs:

1. The need to identify the major risk factors affecting the project cost to prioritise the factors to be incorporated in the cost estimation to overcome the cost overrun issue.
2. The need to investigate the severity of cost overruns in NZ road projects to identify if the issue is significant and needs an immediate solution.
3. The need to identify the statistical techniques that are being used to predict construction project costs.
4. The need to investigate existing cost models developed for road projects to identify the statistical techniques most suitable for road project estimation.
5. The need to develop a cost model specific to NZ road projects using identified cost modelling techniques to improve the conceptual cost estimation practices.
6. The need to validate the developed model using additional case data to test the performance of the project and make recommendations to the cost estimation practises accordingly.

The enlisted research rationale forms the motivation for this thesis. The research rationale specifically forms the knowledge gaps in conceptual cost estimation in the NZ road projects domain. The following section discusses the primary aim of the research and how the aim is achieved through a series of research questions under several research objectives.

1.7 Research aim, research questions, and research objectives

This study aims to improve the accuracy of the current early-design stage cost estimation practice used in NZ road projects. The research proposes an improved cost estimation model based on statistical analysis-based techniques to improve performance to achieve the aforesaid aim. Therefore, the aim has been split into four research questions, then further split into ten research objectives as summarised in Table 1.4.

Table 1.4. *Research questions and objectives*

Research Aim	Research questions		Research Objectives	
To improve the accuracy of the conceptual cost estimation practice at the pre-design stage in NZ road projects.	RQ1	What are the primary risk factors affecting the final cost of NZ road projects?	RO1	To explore the risk factors affecting the cost of construction projects.
			RO2	To distinguish the significant risk factors affecting the cost of NZ road projects.
	RQ2	Does the cost overrun a significant issue in NZ road projects?	RO3	To identify the severity of cost overrun issues in NZ road projects.
			RO4	To investigate the relationship between the project size and time with the cost overrun.
	RQ3	Which modelling techniques can be used to develop a prediction model for cost estimation of transportation infrastructure projects?	RO5	To examine the modelling techniques used for developing cost models for construction projects.
			RO6	To investigate appropriate conceptual cost modelling techniques to use the pre-design stage for ensuring more accurate out-turn cost prediction in transportation infrastructure projects.
	RQ4	Can modelling techniques be used to develop a reliable cost estimation model for NZ road projects?	RO7	To develop and validate the best cost model for conceptual estimation for pre-design stage of road projects in NZ.
			RO8	To examine if the model performance can be improved by combining several methods.

1.8 Ethical approval

Since this research is human-based, only human ethics are considered here. The researcher must always ensure that the research is conducted in an environment that ensures the safety, protection, and rights of the researcher and other human participants. Potential ethical issues are identified and discussed in the research methodology chapter. The thesis was assessed as low-risk research under the Massey University Human Ethics Committee. Upon the ethical approval of the committee, the research topic was assigned to an ethics code compliance number 4000023063 in 2020 for three years, starting 29th July 2020.

1.9 Thesis outline

The official document is a doctoral thesis with publications. It comprises ten chapters, including an introduction, methodology, literature review, data findings and analysis, discussion, and a conclusion. The chapters based on publications have been developed to cover all the research objectives. The chapters are arranged chronologically to show the proper flow of information.

Chapter 1 addresses the background of the study the knowledge gap, targeted objectives, the significance of the research, and other required outlines.

Chapter 2 discusses the research design, approach and research methodology adopted in this study, based on the general research philosophy.

Chapter 3 identifies the significant risk factors that affect the final cost or the cost overruns of NZ road projects.

Chapter 4 investigates the significance of cost overruns in NZ road projects to identify the necessity of improvement in the current early design stage cost estimation practice.

Chapter 5 explores the possible modelling techniques that can be used as cost forecasting models. At the end of the chapter, it concludes which techniques showed higher performance levels in infrastructure project cost estimation.

Using the findings of Chapter 5, **Chapter 6** investigates the reliability of multiple linear regression as a cost estimation technique for the early design stage of road projects in NZ.

Furthermore, **Chapter 7** investigates the possibility of using the artificial neural network as a reliable technique in cost estimation for road projects in NZ at the early design stage.

Finally, **Chapter 8** discusses the findings of Chapters 3, 4, 5, 6, and 7 and investigates the idea of having a hybrid model, combining several models to improve the performance of mono-technique-based models developed in this research.

Subsequently, **Chapter 9** discusses and connects all the chapters as a whole towards achieving one aim. Finally, **Chapter 10** concludes the research by presenting the theoretical and industrial impact and provides recommendations for further research. The chapter also highlights the limitations of the research.

At the end of the chapters, the references and apendices are provided. The appendices include: the ethics approval letter, sample questionnaire, and statements of contribution to the publications. Table 1.5 summarises the thesis outline of this research.

Table 1.5. *Summary of thesis outline*

Chapter number	Chapter title	Publication number
1	Introduction	-
2	Research Methodology	-
3	Risk factors affect the final cost of New Zealand Transportation Infrastructure projects	1
4	Significance of cost overruns in New Zealand Road projects	2
5	Statistical cost modelling techniques for conceptual stage cost estimation of infrastructure projects	3
6	Conceptual cost estimation at the pre-design stage of road projects in New Zealand using regression analysis	4
7	Conceptual cost estimation at the pre-design stage of road projects in New Zealand using artificial neural network	5
8	A Hybrid model for conceptual cost estimation at the pre-design stage of road projects in New Zealand based on neural networks and Monte Carlo simulation.	6
9	Discussion	-
10	Conclusions and recommendations for further research	-
	Reference	-
	Appendices	-

1.10 Literature review

This thesis follows the guidelines of the thesis by publication provided by Massey University. Hence, the chapters do not follow the traditional thesis structure comprising an introduction, literature review, research methodology, results, discussion, and conclusion. Under the thesis by

publication guideline, the publications are considered as chapters. Therefore, literature reviews are also scattered into several chapters. The following Table 1.6 summarises the literature review sections with reference to their locations of the thesis.

Table 1.6. *Literature review summary*

	Description	Chapter number
1	Background study	Chapter 01
2	Literature review – Causes of cost overruns in transportation infrastructure projects	Chapter 03
3	Literature review – Significance of cost overruns in NZ road projects	Chapter 04
4	Literature review – Cost modelling techniques	Chapter 05
5	Literature review – Regression analysis-based models developed for construction projects' estimation	Chapter 06
6	Literature review – Artificial neural network-based models developed for construction projects' estimation	Chapter 07
7	Literature review – Risk-based estimation and Monte Carlo Simulation as a technique for risk-based estimation	Chapter 08

1.11 Epilogue

This chapter discussed the background related to the selected research topic problem statement and identified the research gap based on that. Later, the aim of the research was to answer the research problem within the research gap. Next, the research questions were defined to address the primary research aim. Based on the research questions, relevant objectives were set up to be achieved. The next chapter discusses the research methodology of this research.

2 Research methodology

2.1 Prologue

In pursuing of knowledge and understanding, the methodology employed in a research study plays a crucial role in shaping the outcome and credibility of the findings. This chapter serves as a guideline that discusses the methodological framework used to address the research questions identified in the first chapter. This chapter explains research design to justify the approach to uncovering insights, testing hypotheses, and drawing meaningful conclusions. It begins with introducing to the existing research perspectives, and then the adopted philosophical stance. Based on the philosophical stance, the research approach, research strategy, data collection and analysis methods will be justified. This chapter also encompasses the ethical considerations that have safeguarded the rights of the researcher, the participants of the data collection process, and all the human participants who supported this research. The following section discusses the research purpose.

2.2 Research purpose

This study aims to develop an early-design stage cost estimation model based on the factors affecting cost overruns for road projects in NZ. Research studies differ based on the purpose of the research questions or the intended answers being looked for at the end of the research. Therefore, depending on the intentions, research studies could be categorised into exploratory, descriptive, explanatory, and Evaluative (Cooper and Schindler, 2008; Saunders et al., 2016).

Saunders et al. (2016) emphasised that exploratory studies focus on ‘What’ or ‘How’ related research questions. Therefore, these studies can be used to gain an in-depth understanding of an area of interest by asking open-ended questions to develop hypotheses and propositions for further

inquiry (Saunders et al., 2016; Yin, 2009). In addition, exploratory research starts with a broader focus and then narrows it to the primary research problem through the research progress. On the other hand, Ghauri and Gronhaug (2010), Gill and Johnson (2010), and Saunders et al. (2016) explained that descriptive studies focus on questions based on 'Who', 'What', 'Where', or 'How'. Therefore, these studies aim to gain an accurate profile of happening, people, or situations. Conversely, explanatory studies can be used to understand the causal relationships among the variables (Ghauri and Gronhaug, 2010; Saunders et al., 2016). Subsequently, the authors clarified that evaluative studies aim to find how well something works or to evaluate the performance or effectiveness of a strategy, policy, programme, initiative, or process.

Saunders et al. (2016) confirmed that it is possible for several purposes to co-exist in one research design. Similarly, through the literature review, this research intends to find the causes of cost overruns and the statistical cost modelling techniques that can be used for cost estimation at the early design stage of construction projects. Therefore, the research starts with a broader perspective, and then, towards the end, the focus is narrowed down to the road projects in NZ. Using the broader information, a cost estimation model is developed for road projects in NZ. Hence, this part of the research is designed as an exploratory study. In addition, it is necessary to explore and understand the relationships and correlations among the model's variables to develop the model. Consequently, this research must accommodate both exploratory and explanatory aspects. The following section discusses the research design.

2.3 Research design

Before conducting the research, it is crucial to have a proper research design to justify the research process. The research design includes research philosophy, approach, methodology, strategy, time horizon, techniques, and procedures. Several authors have introduced research designs but with varying terminologies. Tan (2002) noted six research design types: case studies, surveys,

experiments, correlational research, causal-comparative research, and historical research. Yin (2003) introduced experiments, surveys, archival analysis, history, and case studies as research design types.

Further, Bryman (2004) identified five types of research designs: experimental, cross-sectional, longitudinal, case study, and comparative study. Saunders et al. (2016) introduced the research onion, which comprises six layers. According to his view, research design is one of the six layers of research strategy. This layer introduced seven strategies: experiment, survey, archival research, case study, ethnography, action research, grounded theory, and narrative inquiry. In contrast to Bryman (2004), cross-sectional and longitudinal were identified as the separate layers of time horizon in Saunders et al.'s (2016) research onion. Selection of research design through research onion needs careful step-by-step consideration of each layer. Since the process is more straightforward than other research designs, this research adopted the research onion to design the research process.

2.4 Research Philosophy

Research is undertaken to develop new knowledge or theories to address an issue or a problem that cannot be solved with the existing knowledge. That is done through a systematic, in-depth study of the problem, its surroundings, and its nature (Cavana et al., 2001; Fellows and Liu, 2015; Sekaran and Bougie, 2016). It was further emphasised that research must be testable, replicable, objective, able to be generalised, purposive, precise, rigorous, and parsimonious (Sekaran and Bougie, 2016). Similarly, in their studies on research methodology, Esterby-Smith et al. (2002) and Saunders et al. (2016) explained that a belief system and a set of assumptions regarding the world, its reality, the nature of the reality and existing knowledge are required to develop new knowledge based on the nature of research. That was defined as the research philosophy. Hence, in designing the research, the first step is to identify the philosophical stance of the research.

The appropriate philosophical stance will depend on the researcher's ontological, epistemological, and axiological assumptions. In addition to these three philosophical assumptions, Creswell and Poth (2016) identified two more assumptions: rhetorical and methodological. On the other hand, the two extreme stances in philosophical assumptions are objectivism and subjectivism. Table 2.1 summarises the views of these two extremes. Understanding the assumptions of using these terminologies helps to understand the research question, methods, and findings.

***Table 2.1**. Comparison between objectivism and constructivism*
Source: (Elander and Johannes, 2016; Johannes, 2006)

Category	Objectivism	Constructivism
The real world…	has characteristics that can be grouped based on their properties and relations.	is designed based on human interactions and individual minds
Reality is…	common despite the difference in human thoughts of the knowers. So, it can be mapped and shared with others.	personal to individuals. Individual realities are mapped based on individual perception.
Symbols are…	representations of reality are only meaningful to the degree they correspond to reality.	products of a culture that are used to construct reality.
The human mind…	processes abstract symbols and fashions them so that they mirror nature.	observes and interprets the world by creating symbols.
Human thought is…	symbol-manipulation and is independent of the human organism.	is imaginative, and develops out of awareness, sensory experiences, and social interaction.
Meaning…	exists objectively, does not depend on the human mind, external to the apprehender.	is an outcome of an interpretation process which highly depends on the expertise level, experience and understanding of the apprehender.

Ontological assumptions are based on how we see the world's social reality (Bhattacherjee, 2012; Bryman and Bell, 2015; Saunders et al., 2016). Therefore, it helps to outline the means of understanding or seeing the research objects by the researcher. Ontological objectivism embraces realism, or materialism, and regards the world as a tangible, physical entity. Objectivism considers that objects, procedures, and social phenomena exist without the individuals who interact with them, and only one reality is experienced by all. On the other hand, ontological subjectivism,

known as Nominalism, considers multiple realities and does not regard any reality beyond the people who constructed it. A less extreme form of subjectivism is constructivism, which regards the world as a collection of shared meanings and realities as a product of the interaction between individuals (Bryman & Bell, 2015; Johannes, 2006; Saunders et al., 2016).

Epistemology considers the information that is valid and acceptable, focusing on the origins, nature, method, and limit of human knowledge (Fellows and Liu, 2015; Knight and Ruddock, 2008). Epistemological assumptions can also be divided into two extremes: positivism and interpretivism. Positivism, also known as epistemological objectivism, denotes that the knowledge confirmed by our senses can be true knowledge. It uses experience and observation to justify new knowledge and tries to generate universally generalised social realities through these observable and measurable facts (Drake et al., 1998; Saunders et al., 2016; Tan, 2002).

On the other hand, interpretivism, or epistemological subjectivism, uses reasoning and logical argument to gain new knowledge (Bryman and Bell, 2015; Heal and Perry, 2000; Silverman, 1998). It considers that social phenomena are in constant flux and change due to social interaction between the actors. To understand the current reality in detail, interpretivism focuses on rationalisation and logical justification used to observe and justify idealistic views. Thus, it focuses on different opinions and narratives to understand different realities.

The ideology of every researcher plays a role in the research output. Axiology refers to the researcher's views on the study's values, ethics, and scope (Creswell and Creswell, 2018). Even though there are many research philosophies and methodologies, the researcher's view plays a significant role in the credibility of the output. In the objectivist stance, researchers believe that social entities and actors are independent. Researchers keep themselves detached from their values and beliefs to make them unbiased and value-free. In axiological subjectivism, people and social

entities survive dependently, continuously changing and value bound. Therefore, the researcher should engage with the people to understand their views of the people (Saunders et al., 2016).

2.5 The philosophical stance adopted for this research

Researchers conduct research based on assumptions about what they will learn and how they will learn from it (Creswell, 2003; Creswell and Creswell, 2018). This research has focussed on what can be learnt from the existing knowledge. Furthermore, based on the facts from the already completed projects, what more knowledge can be generated, and how can that knowledge be used for other projects? According to Sounders (2016), assumptions based on acceptable, valid, and legitimate knowledge and how to communicate that knowledge to others are known as epistemological assumptions. Similarly, in this research, the primary source of knowledge is the numerical data gained from already completed road projects in NZ. Therefore, the knowledge is acceptable, valid, and legitimate. Consequently, it can be confirmed that this research is based on the epistemological assumption. Next, deciding which philosophy to adopt under this assumption is required.

The epistemological assumption could be on the two extremes of positivism or interpretivism or otherwise be neutral. This research believes that the actual numerical data from the projects gives a better perspective on the project than observing the perceptions of humans, such as industry experts or the members of the project teams. That is because the ideas can be biased or sometimes affected by the level of people's knowledge of the area or understanding of the questions. Therefore, generating new knowledge to develop a cost estimation model depending on human perceptions is challenging. Subsequently, this research focuses on observable and measurable facts to analyse the data and generate a model based on its trends and behaviours for future projects' applications. Hence, the philosophical view of this research can be concluded as epistemological positivism.

2.6 Approach to theory development

Sounders et al. (2016) explained that epistemological positivism is rather deductive and uses a larger sample size to collect data in a well-structured manner. According to Fellows and Liu (2015), the deductive method provides accurate interpretations, provided the premises from which the interpretation is made are true. Therefore, this type of research occurs within the limitation of the existing knowledge and may reinforce the boundaries. Blaikie (2010) explained the deductive method in six sequential steps. He explained that the first step is establishing a hypothesis or an idea to form a theory. Then, a literature review is used to establish a testable proposition or a framework. Thirdly, that proposition or the conceptual framework must be tested with the existing theories to determine if it is still valid. If valid, the required data can be collected and analysed to measure or examine the concept. If the test fails, which means the results are inconsistent, the theory or the concept is false for the examined population. In that case, the theory must be rejected or modified and re-tested. In contrast, if the results are constant, as expected, then the theory is valid and accepted.

In this research, it was necessary to establish three assumptions at the beginning. Then, through the in-depth study, these assumptions were tested to create new knowledge to fill the research gap.

1. The factors that affect the cost of general construction projects can be the same factors that affect the cost of road projects in NZ.

2. Cost overrun is a significant issue in road projects in NZ; therefore, the traditional conceptual estimation practice needs improvement.

3. The statistical analysis techniques in estimating other construction projects can be adapted to develop a model for NZ road project cost estimation.

Thus, the literature gave a broader perspective of the research problem. Then, through the literature, it was further established that there is a need for reliable cost estimation models due to the constant issue of cost overruns. The next step is to identify the causes of cost overruns or, in other words, factors affecting the project cost. Fifty-three factors were identified in the literature, which should be tested to identify the NZ-specific factors. The factors were presented to NZ experts with experience in transportation infrastructure projects in a questionnaire survey. It was identified that the factors were still valid with different severity levels.

Subsequently, the first data collection was done to identify the significance of cost overruns in NZ road projects. If the issue is not significant, there is no need to continue the research to develop a new model. However, the data analysis revealed that the cost overrun is significant in NZ road projects.

The third part of the study was to develop the cost model. The literature identified several cost modelling techniques and narrowed them down to the ones with higher performance levels. Then, using the data from completed road projects in NZ, several models were developed using the best techniques. Based on the model validation results, a theory and conclusion were developed to strengthen the boundaries of the existing knowledge.

Therefore, this research was conducted following the deductive approach in theory development. Generally, deductive research in epistemological positivism is either mono or multi-method quantitative.

2.7 Research methodology

Once the theory development method is identified, it is necessary to identify what method or methods should be followed. Since this research required three areas of propositions to be tested, the multi-method quantitative approach was selected. However, Fellows and Liu (2015) stated that

there should be either no or minimum influence possible by the researcher on the data collected. That is because the quantitative methods are objective and value-free of the researcher's influence. Therefore, the strategy or strategies must be selected accordingly.

2.8 Research strategy

2.8.1 Survey

One of the objectives of this research was to identify the factors affecting the cost of road projects in New Zealand. As discussed under theory development, the literature identified the broader perspective of the factors affecting the cost of construction projects. The findings should be tested for applicability to the NZ road projects. The questions associated with this study were as follows.

1. What are the factors affecting the cost of NZ road projects?

2. How much severity or significant impact does each factor make on the cost?

In the deductive approach, the above types of questions can be answered through the survey. According to Tan (2002), a survey systematically collects primary data based on a sample. Furthermore, Saunders et al. (2016) emphasised that the survey can be used to answer the questions of 'who', 'what', 'where', 'how much', and 'how many'. Therefore, this research used the survey strategy to identify and rank the factors. According to Bougie and Sekaran (2020), the survey can collect information from or about people to understand, compare and explain their knowledge, understanding, behaviour and attitudes. Under the survey, several methods can be followed in data collection and analysis, which depend on the intention of the research problem. Fellows and Liu (2015) state that the surveys can be qualitative or quantitative. However, this research expects to measure the severity to rank the factors affecting the cost. Hence, the quantitative survey method was selected. Survey design will be discussed in detail under the data collection methods.

The following section discusses the case study strategy.

2.8.2 Case study

According to Yin (2014), a case study is an in-depth investigation of a topic or phenomenon in a real-life context. The author further highlighted that case studies are helpful when the boundaries between the phenomenon being studied and the context within which it is being studied are not clear. According to Tan (2002), case studies can be used to test theories guided by hypothesis. Case studies answer the 'why', 'how', and 'what' research questions (Gerring, 2007; Saunders et al., 2016). Several data collection methods are within the case study method, such as interviews, observations, documentary evidence, and questionnaires (Yin, 2014). A distinctive feature of the case study, or between several case studies, is the use of multiple sources of evidence to examine the case holistically and in-depth (Gering, 2007; Yin, 2014). Yin (2014) further added that triangulation is also possible within case studies.

This research conducted case studies with two objectives.

1. To study the significance of the cost overruns in NZ road projects.

2. To collect relevant project information for the independent variables for the model development.

In both situations, the case was considered the completed road projects in NZ. For the first objective, cases were collected from the past twenty years in NZ to understand the trend and nature of cost overruns. One hundred and six projects with various sizes and scales were collected for the first objective. Further, the information required for the analysis was minimal in this objective. (The required data and analysis methods are discussed in detail in Chapter 4). Auckland Transport (AT) and New Zealand Transport Agency (NZTA) were approached to obtain the required data. However, only one hundred and six cases from the past twenty years were collected. Although more projects were completed within this timeline, the data required for the analysis were not

available in other projects due to various reasons. One reason was the poor record keeping. For example, some projects completed more than ten years before were recorded only on paper. Therefore, some detailed records were not available. Secondly, for some of the projects, the person responsible left the agency and did not correctly transfer the records to the others.

On the other hand, as described in Chapter 1, the scope was limited to new construction or significant alterations only. Therefore, all the medium to low maintenance and alteration work-related cases were eliminated. Further, from the Road Classification of NZ, only the major road types, such as high volume, national, regional, and arterial, were considered for the case study, as explained in Chapter 1.

However, more in-depth information is required from the projects for the second objective. That is because, in this step, the model is developed using the independent variables. Therefore, independent variables should have information from all the cases selected for the model development. Nevertheless, getting all the required data from all one hundred and six cases was impossible. Subsequently, based on the data availability, fifty-nine cases were considered for the second objective. Out of fifty-nine, forty-three cases were used for model development, while another sixteen were used for model validation. Meanwhile, seven cases were eliminated as outliers.

Sampling

According to Saunders et al. (2016), several sampling methods exist. Figure 2.1 summarises the sampling approaches. A simple random sampling approach was used to collect the fifty-nine cases for the model development and one-hundred and six cases for the cost overrun analysis. Chapters 6, 7, and 8 further explain the case selection process. The following section discusses the time horizon of the research study.

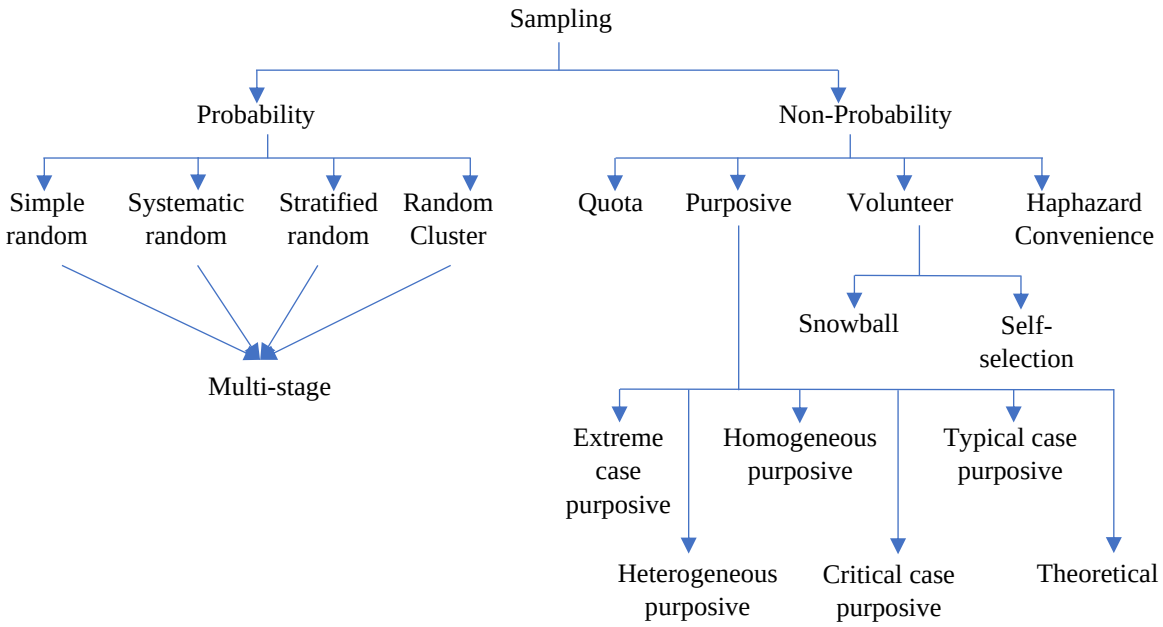

Figure 2.1. *Sampling methods (Source: Saunders et al., 2016)*

2.8.3 Time horizon

This research was conducted as a partial fulfilment of the degree of Doctor of Philosophy. Therefore, the duration was limited to three years. Hence, according to Saunders et al. (2016), the time horizon applied for this research is cross-sectional. Subsequently, based on the above discussion, the research design of this research can be summarised as shown in Figure 2.2.

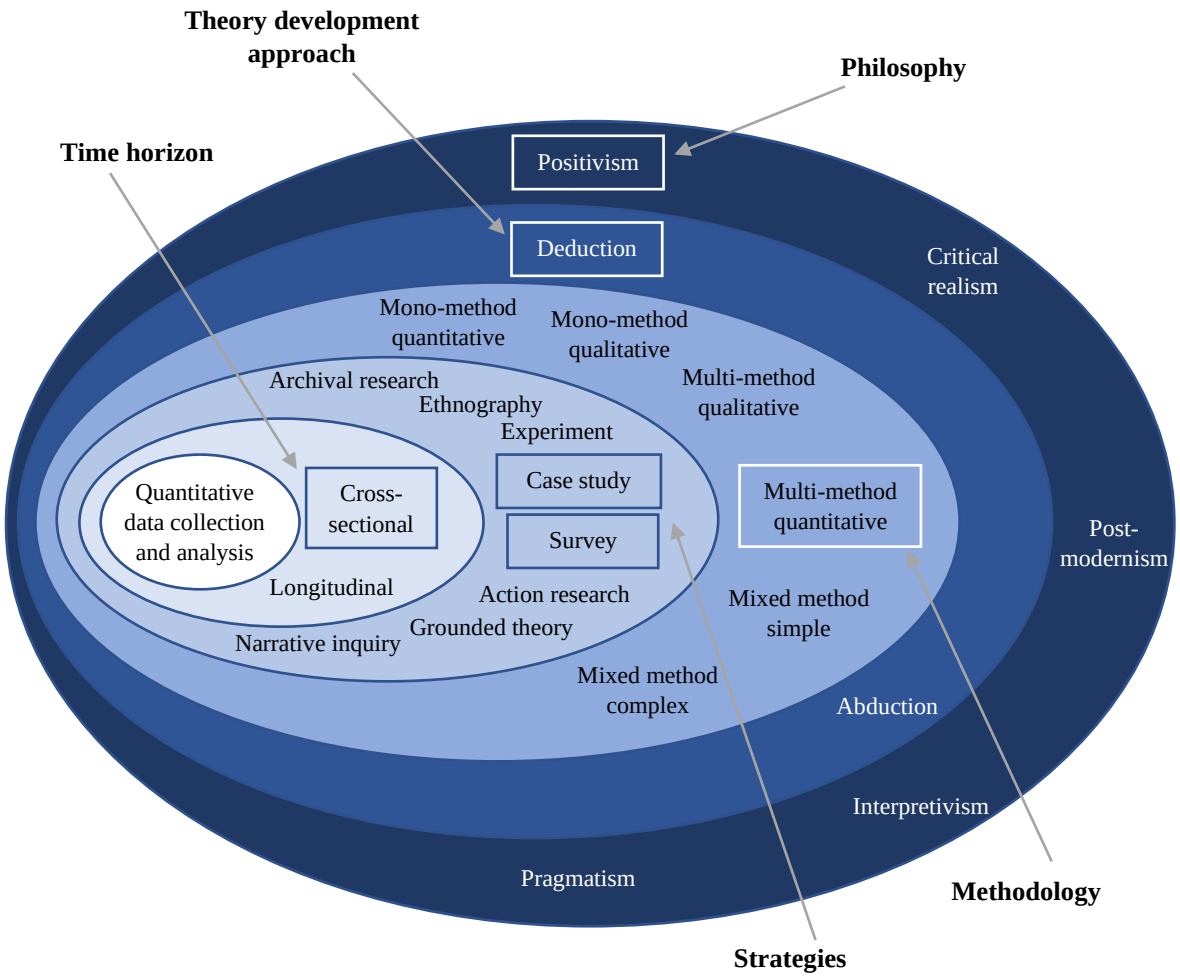

Figure 2.2. *The research design of this research: Source: Saunders et al. (2016)*

2.9 Data collection

Figure 2.3 explains the overall picture of the research process adopted in this research. Based on the process, the data collection steps are discussed in the following sections.

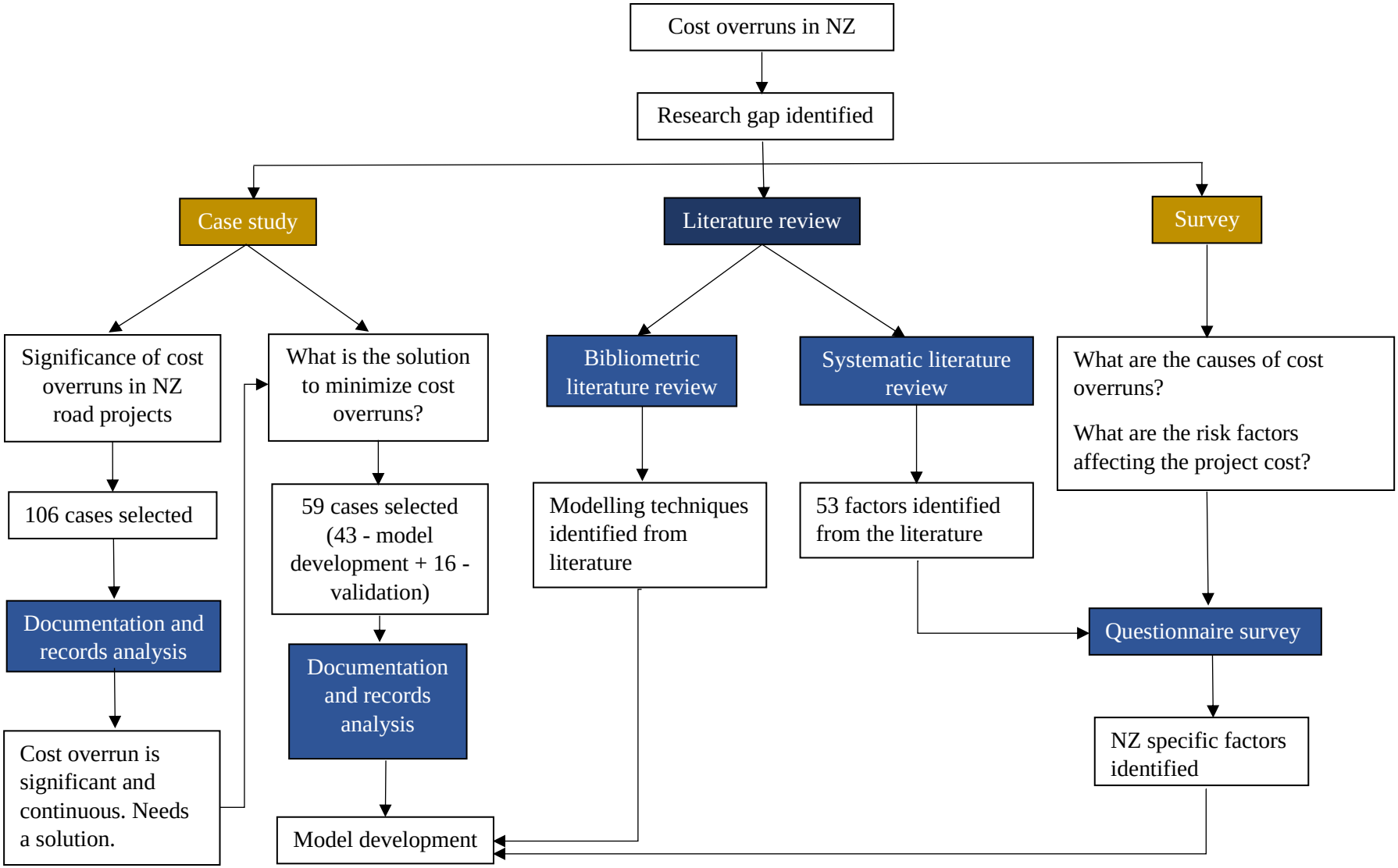

Figure 2.3. *Research process and data collection*

2.9.1 Literature review

A literature review is a selection of available documents (both published and unpublished) on the topic, which contains information, ideas, data, and evidence written from a particular standpoint to fulfil specific aims or express certain views on the nature of the topic and how it is to be investigated, and the practical evaluation of these documents in relation to the research being proposed (Hart, 2018). Sekaran and Bougie (2016) point out that the functions of the literature review are to prevent the rediscovery of known knowledge, be familiar with the terminologies, have a better structure to the research, replicate methods of other researchers and include a contextual and theoretical background for the research providing a broader debate. Saunders et al. (2016) added that a literature review would generate and refine your ideas and place the research within the context of the broader body of knowledge. Therefore, the literature review process is an essential part of the research.

Two significant areas needed the propositions to be identified from the literature before conducting the data collection.

1. What are the factors affecting the cost of construction projects?

2. What cost modelling techniques were already used as cost estimation models in construction projects?

The literature sources reviewed were relevant journal articles, conference papers, books, reports from acceptable entities and government reports. The literature search was conducted on local and online databases, including the Massey University library database, Discover, Google Scholar, Google, Scopus, and Web of Science.

Systematic literature review

The first part of the literature survey identified the factors affecting construction project cost overruns. A systematic literature review is critical to examine all the literature related to the topic. The literature review search started with keywords such as 'cost overruns', 'cost increase', 'cost escalation', 'budget overrun', 'construction', 'infrastructure', 'factor', 'determinant', and 'predictor'. The search was extended using the 'snowball' method by referring to the references and critical authors. The literature search was also conducted on research methodologies to identify the current knowledge on philosophical stances, research methods, research designs, data analysis methods and the most appropriate research design for this study. The primary aspiration was to identify the factors affecting the cost of construction projects. Therefore, other types of systematic reviews were not conducted, such as publications, country, and author reviews. However, the review also comprised a summary of primary findings from the literature. The systematic literature review process and justification are further explained in Chapter 03. The following section explains the bibliometric literature review process.

Bibliometric literature review

Initially, the literature identified seven cost modelling techniques used for cost modelling in construction projects. Conducting a systematic literature review was not easy, considering the number of sources to be reviewed. Another primary reason for selecting a bibliometric review over a systematic review was that it reduced the time and effort of reviewing literature for different modelling techniques. Systematic review would be lengthy and time-consuming because of several different techniques. For instance, in this study, seven modelling techniques were identified. Therefore, the literature review must be conducted in-depth for all seven techniques. If the selected method were systematic, seven different systematic reviews would have to be conducted. However, using a bibliometric review, it is possible to conduct the review for all the techniques

together using VOS-Viewer software. Therefore, a bibliometric literature review was conducted to review the significance of these techniques and the performance of the models developed for infrastructure projects.

As explained in Chapter 5, the Scopus database was used to gather the literature because it is the journal database supported by VOS-Viewer, which was used for the bibliometric analysis. The search terms used were 'cost estimation model', 'cost model', statistical cost model', 'estimation model', 'construction', 'artificial neural network', 'regression analysis', 'support vector machine', 'case-based reasoning', 'reference class forecasting', 'Monte Carlo simulation', and 'fuzzy'. Once the bibliometric review was carried out, the findings were visualised using the VOS-Viewer mapping tools. Several reviews were carried out under the bibliometric analysis. The primary study was the keyword co-occurrence analysis, where the occurrences and the relationship between the keywords were related to the above search terms. This primary concern for the analysis was identifying the major statistical techniques and their significance in cost estimation. The analysis was further extended to identify the sources of publications, significant countries, and authors involved in the relevant research areas. Subsequently, the key publications from the primary authors involved in the research area were reviewed, and the findings were summarised.

Based on the information acquired, the literature review is structured logically. A background review identifies the necessity for an early-design stage cost estimation model for road projects; this was further explored in the literature review.

The bibliometric literature review process and justification are further explained in Chapter 05. The following section discusses the questionnaire survey as a data collection tool.

2.9.2 Questionnaire survey

The literature presents neither a precise evaluation nor many research studies into the impact of the factors affecting the cost of road projects. Instead, it has emphasised a need for research in this field. As a part of the exploratory phase of the research, the survey research design was selected to capture a broad view of the research issues. A survey is a systematic method of collecting primary data based on a broader population using economic data collection methods, such as questionnaires and structured interviews (Fellows and Liu, 2015). Furthermore, the research questions were to be investigated in the form of "what", and views were to be captured at once from many respondents. Thus, a cross-sectional survey design was considered an appropriate research design. The cross-sectional design best suits studies that determine the prevalence of a phenomenon, situation, problem, attitude, or issue by taking a cross-section at once (Saunders et al., 2016).

Additionally, as the philosophical stance of the research was based on epistemological interpretivism, it was expected to collect only quantitative data. Furthermore, data collected through surveys provides the opportunity to use statistical analysis. However, the type of data collected through surveys depends on the data collection method. Therefore, the data collection method of this research was an online questionnaire survey. It is an economical method and facilitates the collection of data from respondents scattered over a large geographical area during a specific period (Bougie and Sekaran, 2020).

Furthermore, the questionnaire survey was carried out during the initial phase of COVID-19. Therefore, meeting the participants in person made it impossible to distribute the questionnaire. Further, the selection was appropriate as the questionnaires could be used in descriptive and explanatory research.

Questionnaire design and development

The questions were designed to capture "opinions" (i.e. variables record how respondents feel about something or what they think or believe is true or false), "behaviour", i.e. what respondents do – concrete experience (did/do now/will do) and "attributes" – respondent characteristics (exploring how opinion and behaviour differ between respondents/to check that data collected are representative of the total population) (Saunders et al., 2016) from the survey respondents. Fellows and Liu (2015) explained that questionnaires can be open or closed based on the type of answers the researcher expects. The delivery mode was set to online, as the survey was conducted during the initial stage of COVID-19. Therefore, participants may have generally been reluctant to answer open-ended questions that need time if there is no personal connection with the researcher during the survey. Moreover, Saunders et al. (2016) explained that questionnaires delivered over the web or email should not be too complex. Hence, it was decided to have only closed-ended questions. The questionnaire was divided into seven sections, including background (8 questions) and rating of factors (10 categories - 53 questions). Refer to Annexure 02 for the final version of the questionnaire. The questionnaire design and sampling method are explained and justified in detail in Chapter 04.

Questionnaire sampling method

The main aim of the questionnaire was to identify the risk factors affecting the project cost. Therefore, the sample should contain the experts who mainly handle the project cost management. Therefore, the selected sample was the Quantity Surveyors, Estimators and other professionals who work closely with the project cost of transportation infrastructure projects. However, it is also recommended to have the opinions of other construction professionals working on the project. Nevertheless, the study was conducted during the alert level 3 and 4 lockdowns of COVID-19. Therefore, there were restrictions in meeting people to gather questionnaire answers.

Subsequently, the questionnaire was distributed through the New Zealand Institute of Quantity Surveyors (NZIQS) to its members in all grades. Therefore, the other professionals' opinions are recommended for further research.

Questionnaire sample size

The calculation method of the questionnaire sample size is described in detail in Chapter 3.

The following section discusses the documents and record analysis as a data collection tool.

2.9.3 Documents and record analysis

Under the case study research strategy, the adopted data collection method was documents and record analysis. Documentation prepared for more formal use is called records, and information recorded for personal use is called 'documents' (Hodder, 1994). The author added that building contracts and legal agreements are records, whereas memos, letters, and field notes are documents. Further, Bowen Glenn (2009) and O'leary (2017) emphasised that there are three types of primary sources for document analysis. First, organisations officially publish public records such as annual reports, strategic plans and mission statements. The second source is personal documents containing personal experiences or individual events, such as blogs, newspapers, journals, and Facebook posts. The third source is physical evidence, such as training materials, posters, and flyers. Document analysis can provide background information within a specified context, monitoring growth or decline trends, and complement findings from an existing source of information (Angrosino and Mays de Perez, 2000; Bowen Glenn, 2009)

Acquiring cost data from projects is difficult; due to the sensitivity of the details, acquiring cost and time details will depend on participants' willingness to supply these. Therefore, the sampling method to acquire such data will be the snowball method (Saunders et al., 2016). Cost data from databases do not have such limitations, so that the total data population can be acquired for

research. The limitations of not having data from specific projects could be minimised by using information from databases.

This research requires the analysis of records such as government reports, contract documents, and final accounts documents. The document analysis collected two primary data types: actual cost data from completed road projects and generic cost data from databases such as cost indices. The required cost data for the models' parameters were collected from the final accounts, contract documents, government reports, and relevant standard methods of measurement. Contract documents specify the exact time and cost estimate for the project and the type of payment method. In addition, the final accounts provide the actual cost and time spent on the various elements of the project by the end of the contract completion, as well as the time that has elapsed. These documents give the cost and time deviation between estimation and actual. Cost databases provide indices that can be utilised to calculate cost fluctuations over time. Therefore, document analysis was the primary data collection method for model development and validation. Data collection for document analysis is explained in detail under Chapters 3, 4, 5, 6, 7, and 8.

2.10 Quantitative data analysis

This research contains three types of quantitative data to be analysed. The first type is the questionnaire data analysis, and the second type is the non-parametric analysis of the cost data to identify the significance of the cost overruns. The last is the model development. The following sections discuss these three types of data analysis.

2.10.1 Non-parametric analysis

Before developing the model or identifying the factors affecting the cost of road projects, it was crucial to identify if the cost overrun is a crucial issue in NZ road projects. Therefore, the data were collected from 106 road projects. Before conducting any quantitative analysis, it is necessary

to determine the normality of the distribution because the test type will depend on the distribution. If the distribution is normal, then parametric tests can be conducted. In contrast, the tests must be non-parametric if the data is not from a normal distribution.

There are two commonly used normality tests supported by SPSS. They are Kolmogorov-Smirnov and Shapiro-Wilk tests. However, the Kolmogorov-Smirnov test could perform well with larger data sets (Grech and Calleja, 2018), while the Shapiro-Wilk test could be used for smaller data sets (Villasenor and Estrada, 2009). Therefore, the Shapiro-Wilk test was used for the normality test as the data set was only 106 projects. In addition to this, the normal Q-Q plot was also examined for normality. According to Augustin et al. (2012), the data points would fall on a 45-degree reference line if the data distribution is normal. However, the data set did not meet the normality requirement. Therefore, the data were analysed using non-parametric tests.

Kruskal-Wallis H test

The Kruskal-Wallis (KW) test is a non-parametric test similar to the Analysis of Variance (ANOVA). To apply the KW test, there should be a defined hypothesis. The null hypothesis was that there is no effect on the cost overrun by the project size or the time. In other terms, if there is no effect, the mean cost overrun of all projects with different sizes or completed any year should be equal. The alternate hypothesis was that the project size and the time affect the magnitude of cost overruns. To conduct the KW test, all one hundred and six projects were categorised into small, medium, and large-scale projects based on the contract price. If the *p*-value of the test is higher than 0.05 alpha level, the null hypothesis can be accepted (Tan, 2002; Bryman and Cramer, 2005; Ilozor, 2009). The test is further explained in Chapter 4.

2.10.2 Questionnaire survey data analysis

The second part of the data collection was the questionnaire survey. The data analysis techniques depend on the data type and their measurement scales, such as nominal, ordinal, interval, and ratio.

It is possible to conduct quantitative analysis manually or using Microsoft Excel because all the tests are mathematical calculations. However, it will take a considerably long to analyse the data, and the results may be influenced by the researcher's capacity to handle all the functions and equations. Using computer-aided statistical software, the data can easily be manipulated, interpreted, and displayed differently (Robson, 2002). Therefore, SPSS (Statistical Product and Service Solutions) version 28.0.1.1(15) was used for the analysis.

The primary objective of the questionnaire was to identify the significant factors that affect the cost of transportation infrastructure projects. The Likert scale is the most used rating scale. Jacoby (1971) emphasised that the Likert scale can be used in measuring attitude and image and is often considered an interval estimate. The data collected through the questionnaire were tested for reliability using Cronbach's alpha. Data reliability is explained in further sections of this chapter.

Three major measurements were taken: severity index (SI), coefficient of variance (COV), and coefficient of correlation (COC). SI was used to identify the significance of the factors, while COV was used to measure the agreement of the level of significance of the factor among the participants. Based on the SI and COV, the factors were ranked to identify the significant factors affecting the cost of transportation infrastructure projects in NZ. The justification and further details of these measures are explained in Chapter 3.

2.10.3 Parametric Analysis

Through the literature review, seven cost modelling techniques were identified, and out of those seven, three techniques performed well with the infrastructure projects. These were namely: regression analysis (RA), artificial neural network (ANN), and support vector machine (SVM). Of those three, RA and ANN were common and performed well for road projects, and the industry experts were very familiar with the techniques. Therefore, SVM was not considered for model development. However, none of the three abovementioned techniques work well with risk

analysis. However, Monte Carlo simulation (MCS) is an excellent technique identified from the literature that works well with risk-based estimation. Therefore, in this research, three models were developed using the data collected from road projects in NZ using RA, ANN, and MCS. The following subsections discuss the overview of the modelling techniques used.

Multiple linear regression analysis (MLRA)

Table 2.2. *Regression assumptions and tests*

Assumption	Description	Hypotheses	To accept the H_0
Linearity	Independent variables and dependent variables have a linear relationship	H_0: coefficients of the regression equations are equal to zero H_a: any of the coefficients or at least one coefficient is not equal to zero	Regression model statistics *F*-value > *F* (critical value) And *p*-value > 0.05
Normality	The residuals of the regression are normally distributed	H_0: the errors follow a normal distribution H_a: the errors do not follow a normal distribution	Kolmogorov-Smirnov test, Shapiro-Wilk test *p*-value < 0.05
Independence	Independent variables do not correlate.	H_0: the independent variables do not have any correlation H_a: the independent variables are not independent of each other	Correlation analysis and multicollinearity The correlation coefficient should be lower. VIF < 10
Homoscedasticity	Data points have a constant variance.	H_0: the data points have homoscedasticity behaviour H_a: the data points have heteroscedasticity behaviour	The scatter plot of the predicted value against the residuals must not indicate any pattern but scattered randomly.

The first model was developed based on MLRA using SPSS software. Hair et al. (2007) stated that regression analysis is the most widely used analytical technique in measuring linear relationships. The model was developed to predict the final cost of road projects based on the details available at the early design stage of the project. The ANOVA test is used to analyse the model's fitness, validity, and significance in predicting the cost of road projects. The regression model depends on the validity of four assumptions. Once the model is developed, the assumptions must be tested to validate the model. The assumptions are linearity, normality, independence, and homoscedasticity.

These assumptions were tested using a hypothesis. Table 2.2 elaborates on the assumptions and hypotheses used in this research. Chapter 6 discusses the regression analysis process in detail.

Artificial neural network (ANN)

The ANN has simply adopted the architecture and functions of human brain cells and the nervous system that learn from previous experience and respond to complex problems (Tijanic et al., 2020). Silva et al. (2017) state that ANN can be used in universal functional approximation, pattern recognition, process identification and control, predictions, and system optimisation.

ANN models comprise three primary layers: input, hidden and output. The input layer receives the information/ data to the model. The number of neurons in this layer depends on the number of independent variables selected for the model. Each independent variable represents one neuron in the input layer. Thenceforth, the centre part of the model contains the hidden layer. Depending on the model architecture, an ANN model can have one or more hidden layers. This layer extracts patterns associated with the process or system being analysed (Silva et al., 2017). The number of neurons in the hidden layer can be decided based on the model's performance. The model can be re-run, adjusting the number of neurons in the hidden layer, and the model with the highest performance will be comprised of the ideal number of neurons. Ultimately, the output layer is composed of the neurons that produce the result or the outcome based on the analysis and calculations performed in the model. The number of neurons is based on the model's expected output. Usually, the output layer consists of one neuron because most estimation models are developed to forecast the project's final cost.

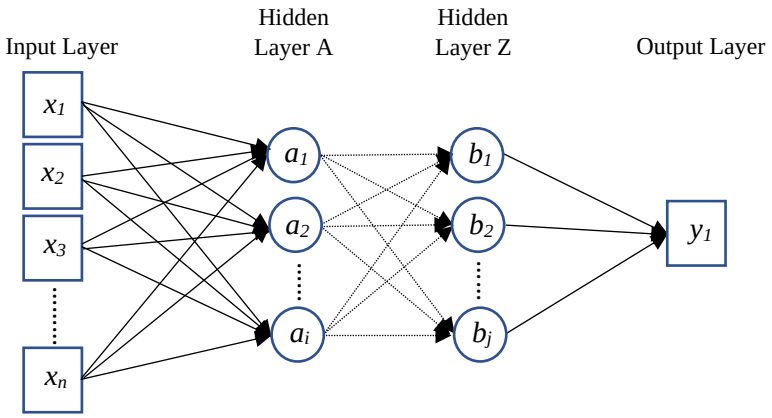

Figure 2.4. *ANN Architecture (Source: Chapter 7)*

The multiple-layer feedforward method is the most used architecture in ANN models because the single-layer feedforward models do not contain any hidden layers (Silva et al., 2017). Figure 2.4 shows a sample multiple-layer feedforward network with an '*n'* number of independent variables (*x*) and two hidden layers with '*i'* and '*j'* numbers of neurons, respectively. The output layer has only '*m'* neurons representing the respective output of the problem being analysed by the model. The Multiple Perceptron (MLP) and the Radial Basis Function (RBF) are the most widely recognised multiple-layer feedforward architectures (Ahiaga-Dagbui and Smith, 2012; Silva et al., 2017; Tijanic et al., 2020). These two models utilise the generalised delta and competitive/delta rules, respectively, as the learning algorithm in the training phase. However, Ahiaga-Dahbui and Smith (2012) emphasised that the MLP model is superior to the RBF model. That is because the MLP model focuses on identifying the relationships between the inputs and the output, whereas the RBF model performs in two stages. The first stage performs a probability distribution of the inputs and identifies the relationships in the second stage. Therefore, this study only focuses on developing the ANN model adopting the MLP principles. ANN model concept and development is further explained in Chapter 7.

Monte Carlo simulation

In Chapter 5, it was noticed that statistical techniques such as regression analysis, artificial neural networks, and support vector machines do not perform well with risk-related variables. Therefore,

these techniques cannot correctly predict the risk involved in construction projects. However, risk estimation is a significant component of the project cost. Without proper risk analysis and estimation, the project may experience cost overruns. Therefore, this chapter investigates the risk estimation of construction projects.

Construction risk cannot be estimated with exact figures because the risk is an uncertain future situation, which could either happen or not happen. Sometimes, the risk may occur but at a different magnitude than estimated. Therefore, the best way to estimate the risk is through its probability. However, Peleskei et al. (2015) stated that proper risk analysis is often challenging during the estimation stage, yet it is very significant for the reliability of the estimation. Over the years, researchers have developed several models and techniques to analyse the project risk. Monte Carlo simulation (MCS) is a well-known technique commonly used in the construction industry to estimate risk probability. Therefore, this study analyses the road project data used in chapters 6 and 7 to analyse the risk probability using MCS.

According to the estimation classification system introduced by AACE (2020), at the conceptual stage, usually, the estimate should be expected to have -20% to -50% of a lower range to +30% to +100% of a higher range of variation. Therefore, the project risk level is higher at this stage due to the limited information availability. However, the conceptual estimate will lead to significant project decisions and budget establishment. Hence, it is essential to have a reliable estimate. Consequently, risk estimation is required to ascertain the possible percentage of variation from the primary estimate. AACE (2020) defined contingency as an amount added to the project estimate to allow for items, conditions, or events for which the state, occurrence, or effect is uncertain and that occurrence will possibly demand additional cost. Generally, the contingency is estimated using statistical analysis of experts' judgement based on past experience or lessons learnt (Maronati and Petrovic, 2019). However, it was observed that some of the projects allow a general

contingency sum as a percentage of the estimated value (Gbajobi et al., 2018; Peleskei et al., 2015). However, allowing the contingency without a proper risk analysis creates a riskier situation as it is uncertain what to expect during construction. Further, if the project faces unexpected risk with a higher financial impact, the allowed contingency may not be sufficient (Gbajobi et al., 2018). Therefore, this study investigates MCS as a technique for risk estimation that combines with the ANN-based model developed in Chapter 7 to improve its performance and reliability.

2.11 Data Validity and reliability

Evaluating the quality of the research is required to decide upon the dependability of the research outcome. The significant concerns about the quality of the research focus on its validity and reliability. These measures focus more on qualitative research, but some concerns relate to quantitative and mixed-method research. Thus, these measures must be discussed in this research.

2.11.1 Research bias

Bias is a crucial concern in both quantitative and qualitative approaches. That is because if the survey participants are biased toward a person or a group, then the answers provided may not be valid. Therefore, survey findings may deviate from the factual findings (Shuttleworth, 2009). However, in quantitative research, bias can be eliminated through statistical analysis (Bougie and Sekaran, 2020). Therefore, under the quantitative analysis, how to use the analytical tools to test the validity and reliability of the data and findings will be discussed.

2.11.2 Data triangulation

The need for triangulation arose from the ethical need to confirm the validity of the processes. Denzin (1984) identified four (04) types of triangulation: (a) Data source triangulation when the researcher looks for the data to remain the same in different contexts; (b) Investigator triangulation, when several investigators examine the same phenomenon; (c) Theory triangulation, when

investigators with different viewpoints interpreted the same results; and (d) Methodological triangulation, when one approach is followed by another, to increase confidence in the interpretation. This research study used data source triangulation and methodological triangulation.

Firstly, under data source triangulation, several data sources in NZ were approached to collect data, assuming that the data from all sources react to the tests similarly to ensure the reliability of the data. Under methodological triangulation, this research followed several approaches to the same study to ensure the validity and reliability of the outcome.

2.11.3 Data validity

The measure used to check the correctness of the research is called the research's validity (Creswell and Poth, 2016). There are many definitions and subcategories of validity. Validity is the process of verifying collected data, data analysis and interpretation of the data to verify its validity, authenticity, and credibility (Saunders et al., 2016). Bougie and Sekaran (2020) defined validity as the Evidence that the instrument, techniques, or process used to measure a concept indeed measures the intended concept. Thus, the research validity is an essential factor in the research. There are three approaches to assessing the validity of the research. They are content validity, construct validity and criterion validity (Bougie and Sekaran, 2020; Hair et al., 2007; Saunders et al., 2016).

Researchers explain that content validity, or face validity, measures the ability or the extent of the coverage of the investigation related to the research question addressed in the data collection (Bougie and Sekaran, 2020; Hair et al., 2007; Saunders et al., 2016). Construct validity is comprised of two approaches: convergent and discriminant validity. Convergent validity is established if the two different instruments that measure the same concept receive highly correlated scores (Bougie and Sekaran, 2020). In contrast, discriminant validity is established

when two variables show low or no correlation when the expectation is to have two distinctive variables (Bougie and Sekaran, 2020; Saunders et al., 2016).

Finally, criterion validity also has two approaches: predictive and concurrent. Predictive validity concerns the ability to predict the outcome using the measures or the questions used in the instrument. Conversely, concurrent validity establishes when the participant/variable's different perspectives/ criteria should be measured and identified differently by the measurement. For example, while the review identifies the statistical analytical tools, it is also necessary that the tools can be used for developing prediction models. That is because the findings cannot be valid if the review does not differentiate the tools based on their ability to be used as a predictive model. Therefore, the results should discuss the tools that satisfy both criteria. Table 2.3 explains in detail the above three approaches, with solutions.

Table 2.3. *Research validity measures (Source: Bougie and Sekaran, 2020; Hair et al., 2007; Saunders et al., 2016)*

Approach		Description	Method of achievement
Content validity	Face validity	Does the questionnaire cover the required scope? If the review considered all the models developed for	The systematic literature review and bibliometric review were aimed to cover the entire area within the expected scope of the questionnaire or model development. Once the questionnaire was developed, it was presented to several industry experts to get their opinions on its validity. Finally, the PhD supervisor approved the questionnaire before commencing the data collection.
Construct Validity	Convergent validity	Do the two instruments measuring the same concept highly correlate?	Search terms of the literature review were chosen to satisfy the required criteria of the research. Statistical analytical tests such as correlation analysis and factor analysis were used to measure the validity
	Discriminant validity	Does the measure have a low correlation with a variable that is supposed to be uncorrelated to this variable?	
Criterion validity	Predictive validity	Does the measure differentiate individuals in a manner that helps predict a future criterion?	
	Concurrent validity	Does the measure differentiate in a manner that helps to predict a criterion variable currently?	

2.11.4 Data reliability

Reliability refers to the consistency of the data collected. Researchers have identified three ways of achieving data reliability. They are test re-test, internal consistency, and alternative form (Bougie and Sekaran, 2020; Hair et al., 2007; Saunders et al., 2016). However, in this research, the test re-test was not adopted as it required administering the questionnaire to the same participants twice. Moreover, re-distributing the same questionnaire may be interpreted negatively by the participants. On the other hand, identifying the position held by the participant who completed the questionnaire is another way of assuring the data's reliability (Oppenheim, 1992; Love, 2002). Also, the questionnaire ensured that only professionals with experience in transportation infrastructure projects completed it.

Similar to the first method, alternative forms measure the same concept in two instances with the same participants (Hair et al., 2007). However, by achieving this, the questionnaire may be lengthier than expected. Hence, the participants may lose interest, notice similar measures in different areas, and refer to the previous response. Therefore, this method was also not tested as the survey was conducted online, and it was crucial to minimise it as much as possible.

Table 2.4. *Data reliability measures*

Threat	Description	Mitigation method
Participant error	Factors affecting the performance of the participants	The questionnaire survey was designed to complete within short period of time to keep participants interest and to make it easier to read, understand and answers the questions easily.
Participant bias	Factors affecting the false response	Participants identity was secured and not disclosed to others. Participant bias also minimised using triangulation methods.
Researcher error	Factors altering the researcher interpretation of the data	The data analysed were checked by other researchers such as main supervisor and co-supervisor.
Researcher bias	Factors that induce bias of the researcher	The final output of the research was validated with actual case study data by statistical validation. (Refer model validation in chapter 6, 7, and 8)
Source reliability	To ensure the data reliable it must be ensured that the data	In the questionnaire survey, the position and the experience of the participants were recorded and analysed to enhance the reliability.

	is gathered from a reliable source.	For the case study, the data was collected from the responsible personnel for completed projects from the NZTA and AT.
Internal consistency	The extent to which the items in the questionnaire or items in a question are related to each other.	Before analysing the survey data Cronbach's alpha test is conducted. The test identifies problem items that should be excluded from the scale.

In contrast, internal consistency is used to measure the consistency of responses across the subgroup of questions. That can be measured statistically. This test determines the extent to which the questions or items in the same subgroup or category are related to each other. Cronbach's alpha is the most used internal consistency test (Bougie and Sekaran, 2020; Hair et al., 2007; Saunders et al., 2016). Therefore, the quantitative data analysis also used this test for the reliability of the questionnaire. Based on the above discussion, Table 2.4 summarises the reliability methods adopted in this research.

The following section discusses the achievement of research questions through the adopted research methodology discussed above.

2.12 Research questions versus research methodology

Figure 2.5 illustrates how the research methodology adopted in this research is used to achieve the research objectives and answer the research questions.

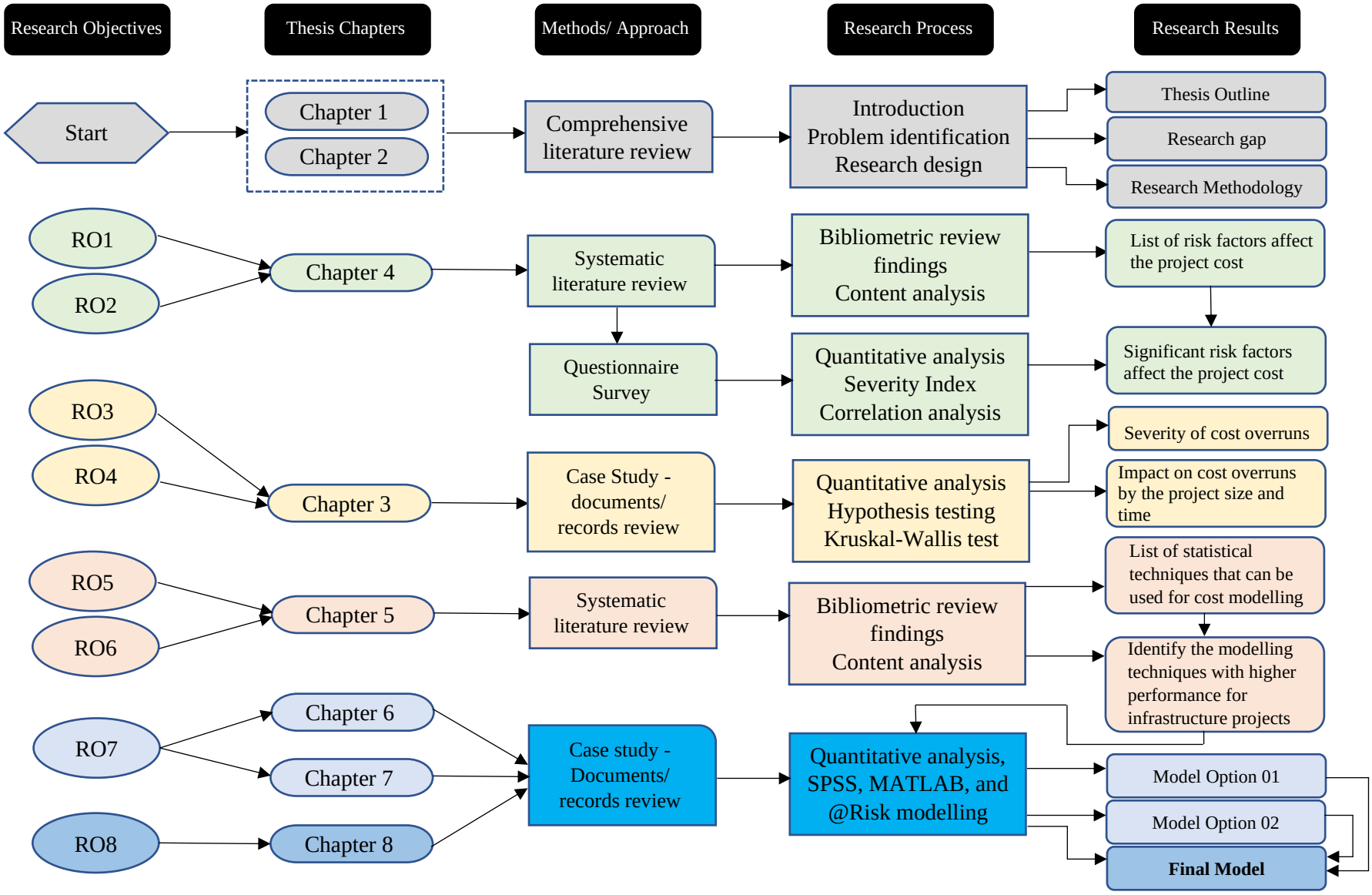

Figure 2.5. *Research process through research methodology*

2.13 Ethical considerations

This thesis was assessed as low-risk research under the Massey University Human Ethics Committee. Upon the ethical approval of the committee, the research topic was assigned to an ethics code compliance number 4000023063 in 2020 for three years from 29th July 2020.

This aims to ensure that the research is not undertaken for harmful purposes, that no harm comes to anyone involved, and to protect the participants' safety, protection, rights, and confidentiality. It is required to take the participants' consent to take their details while collecting data, and the information must not be accessed by anyone else but the researcher. The issues concerned with gaining ethical approval are gaining access to particular subjects, ensuring the information they give is safeguarded and not spread to others beyond the research, ensuring that participants are treated with respect, as well as the information they provide and that the information they provide is only that which is needed in the research, not just gathering information for the sake of it, ensuring that there are no risks in the gathering of information, either to the participants (reputation, health, security) or to the researcher (emotional upset, danger to the person) (Wisker, 2007).

The research was peer-reviewed by the supervisors for ethical issues such as the anonymity of participants and organisations, participants' consent, the confidentiality of information collected and the conflict of interest. Precautions were taken throughout the research to manage the participants' sensitive information.

- The anonymity of participants: Only the personal views of study participants were collected. Identifying information such as name, organisation, and business was not collected on either person or organisation, and that information does not appear in any part of the research.

- Participants' consent: Information on the study was presented to the participants, and their consent was duly obtained in writing before participating. Participants could withdraw their participation at any time during and after the study.

- Confidentiality of information: All responses were kept confidential in line with Massey University's code of ethical conduct for research, teaching and evaluations involving human participants (Revised code 2017).

- Conflict of interest: The research is generic, and no conflict of interest is engaged with any party or organisation.

2.14 Research scope

This thesis focuses on improving the reliability of conceptual cost estimation at the early design stage of NZ road projects. The research scope covers the research domain of investigation, geographical coverage, and unit of analysis and observation.

2.14.1 Geographical coverage

Considering the above research scope, the collected data for this project covers various locations and regions throughout NZ. Therefore, the model developed in this research applies to any geographical location within NZ.

2.14.2 Domain of investigation

This research considers cost overruns of the NZ road projects, their severity, risk factors affecting the project cost, and their impact. Subsequently, the research develops three models using several modelling techniques to identify the ways to improve the current estimation practice for NZ road projects.

2.14.3 Unit of analysis and observation

This research conducted a questionnaire survey to prioritise and validate the identified risk factors for Quantity Surveyors and Estimators in NZ with experience in road projects. As specified above, the research aims to improve the current estimation process, and the developed model is to be applied to NZ road projects. This research adopted a quantitative method (questionnaire survey and project cost documents analysis) to achieve the overall research aim. Therefore, the NZ road projects and the selected quantitative research approach form this study's unit of analysis and observations.

2.15 Epilogue

This chapter elaborated on the research methodology adopted in this research. The background and research gap analysis identified the cost overrun as a significant issue commonly faced by construction projects. Therefore, as the first step, the next chapter will investigate the significant risk factors affecting the cost of transportation infrastructure projects. That is because, ultimately, such causes lead the projects to cost overruns or budgetary failure.

3 Risk factors affect the final cost of New Zealand transportation infrastructure projects

3.1 Prologue

Understanding the underlying causes of cost overruns is critical to mitigating their occurrence and improving the efficiency and effectiveness of the project. Factors contributing to cost overruns can be attributed to a combination of technical complexities, personnel, and external environment influencers. As construction projects become more intricate and demanding, the need for accurate cost estimation and effective project management becomes paramount. Therefore, this chapter investigates the factors that affect the project cost of the road projects in New Zealand. It provides the background information for further investigation of cost model development because the cost estimation process should be able to address the significant risk factors. The chapter investigate in-depth literature to identify the factors. Subsequently, the factors are validated and prioritise based on the severity using questionnaire survey. The chapter identifies some knowledge gaps in relation to the risk factors and cost estimation in NZ road projects. Finally, there will be some recommendations provided that will lead to the next stage of this research[1].

[1] This chapter is based on the following publications.

Atapattu, C.N., Domingo, N., and Sutrisna, M. (2023). What significant risk factors affect the final cost of road projects in NZ – Quantity Surveyor's perception, *Built Environment Project and Asset Management*, Vol. 13 No. 5, pp. 756-777. DOI: https://doi.org/10.1108/BEPAM-07-2022-0105.

Atapattu, C. N., Domingo, N. and Sutrisna, M. (2022). Causes and effects of cost overruns in construction projects. The 45th Australasian Universities Building Education Association Conference (AUBEA 2022), Sydney, Australia, 23-25 Nov.

3.2 Abstract

Cost overrun is one of the critical issues faced in construction projects, as nine out of ten projects will likely go over the budget. In particular, transportation infrastructure (TI) projects, such as roads and bridges, are vastly affected by cost overruns, which can delay the entire project. This research intends to identify the factors affecting the cost overruns in New Zealand (NZ) TI projects. The research was conducted using a questionnaire survey involving ninety-two participants experienced in infrastructure project estimation in NZ. Quantitative methods were used to analyse the data. Fifty-three factors were identified through literature under ten categories. Based on the survey, ten significant factors were identified with a high grade of importance. The three most critical factors were 'frequent design changes', 'poor planning and scheduling', and 'inadequate tender documentation'. It was found that the cost overrun is primarily affected by the pre-contract stage causes. Although much research is done to identify these factors, they are only considered in a few statistical cost models. These new statistical models mainly focused on technical variable factors similar to the current standard estimation process. However, the results of this research, qualitative and quantitative factors, will be used for the future cost model. The results will improve the current estimation practice by developing a new statistical model considering all the significant variables for NZTI projects. The data were collected from professionals involved in NZTI projects. Therefore, the implications may be different for other contexts.

3.3 Introduction

Construction cost is an important performance indicator of a project's success. Thus, construction projects that cannot finish within the budget have been considered projects delivered with poor performance (Abdul *et al*., 2013; Ameh *et al.,* 2010). Therefore, the accuracy of the project's budget estimate can be partially responsible for the cost overruns as it heavily informs the decision-

making to go ahead. Accordingly, a better understanding of the behaviour of project cost during the various project stages from inception to completion would be beneficial in producing more realistic budget estimates (Creedy *et al.*, 2010). The estimation method used in preliminary estimation should consider all the critical and possible factors affecting the cost overrun. For example, the study by Lind and Brunes (2015) on Swedish infrastructure projects reported that most cost overruns took place at the initial design stages and planning until the design was finalised due to technical and administrative issues.

Much research has been done to identify the factors affecting the cost overrun in the construction industry of various countries, as shown in Table 3.1. However, the construction industry heavily depends upon geographical changes, weather changes, economic capacity, and changes in construction methodology. Therefore, the significance or the severity of the impact of these factors can be different to the New Zealand (NZ) context. However, recent research has not been studied on the NZ transportation infrastructure (TI) projects. Mbachu (2011) researched payment-related risks in the NZ construction industry and observed that cost estimations do not consider the most crucial price-related risk factors.

Nevertheless, O'Brien *et al*. (2014) investigated the challenges the NZ Quantity Surveyors (QS) faced. They identified that the cost estimation needs to be more reliable and accurate due to the unavailability of information. Further, the research suggested finding the significant factors affecting the project cost, which should be considered in estimations.

According to the sources identified through the literature review, TI, such as roads/ highways, bridges, tunnels, and rail projects, face considerable cost overruns in NZ. For example, Auckland City link rail has faced a one-billion-dollar cost overrun. Also, from 2010 to 2017, the number of projects delivered within the budget decreased from 48% to 29% (KPMG, 2017). Therefore, it is necessary to identify the factors critical to NZTI projects, even if it is impossible to eliminate the

causes. However, identifying and addressing the essential variables through a cost model is the best way to mitigate them (Adam *et al.*, 2017).

Furthermore, TI projects are mainly funded by the public sector. Therefore, facing significant cost overruns will affect public funds as the project's budget was not planned for the additional funds at the decision-making stage. That will lead to substantial delays in the projects. Moreover, it will affect the public as TI projects impact traffic controls, route diversions, and road closures. Therefore, having an accurate and reliable budget at the decision-making stage of a project will allow the project team to be prepared for all eventualities. This research aims to identify the factors that can be used as variables to develop a cost model to mitigate these issues in pre-contract estimation.

The following section details the research methodology adopted in this study.

3.4 Research methodology

3.4.1 Systematic literature review (SLR)

A two-stage research methodology was implemented for this study. The first stage involved an SLR of relevant literature to identify the factors previously found to affect cost overruns. The SLR provides an evidence-based approach to identify relevant data for particular research while eliminating bias issues and errors arising due to its transparency, inclusivity, explanatory, and heuristic nature (Denyer and Tranfield, 2009). An SLR is vital as the researchers need up-to-date information, reviews, and discussions on a particular research area to investigate current issues and problems before conducting an actual field investigation (Lipsey and Wilson, 2001).

The literature search was based on keywords related to the cost overruns of TI projects. The databases were selected to ensure a wide range of publications were retrieved. Scopus, Web of

Science, and Science Direct are the most used databases. Figure 3.1 explains the criteria used to gather literature for the SLR in detail.

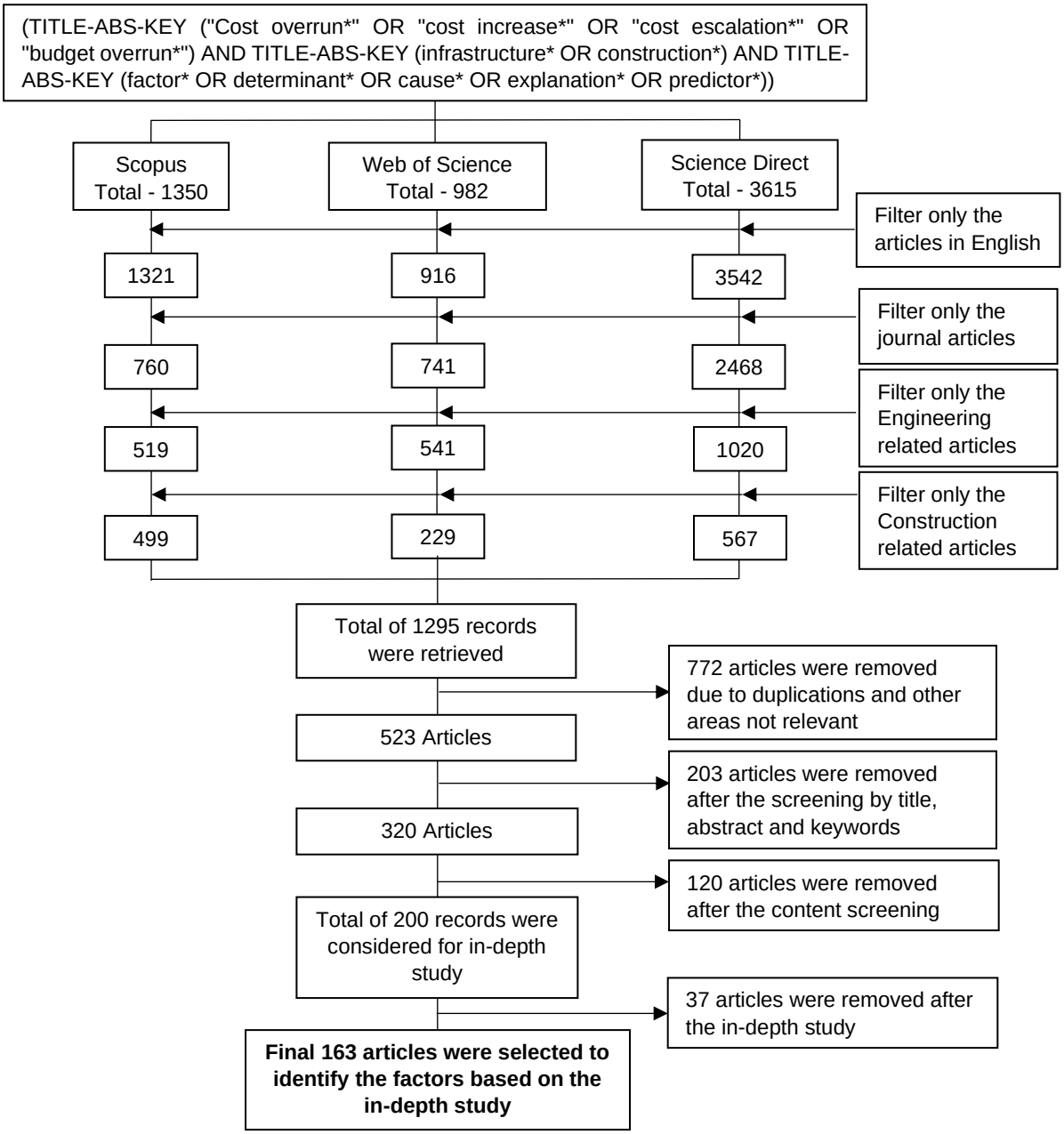

Figure 3.1. *Systematic literature review process*

3.4.2 Questionnaire survey

The second stage of the methodology was to identify and prioritise the most significant factors found through the literature survey, particularly concerning NZTI projects. Due to the COVID-19 restrictions, an online questionnaire survey was carried out. Since the research requires a

knowledge of estimation procedures, it was decided to select professional QSs or other professionals involved in estimating TI projects. Therefore, this research's results indicate the Quantity Surveyor's perception of the TI project cost.

Sample size.

Considering the infinite population, the following equation was used to calculate the sample size (Israel, 1992).

$$n_0 = \frac{t^2 pq}{d^2}$$

Where;

- n_0 : Sample size
- t^2 : abscissa of the normal curve that cuts off an area α at the tail (1.96 for 95% confidence level)
- p : estimated proportion of an attribute that is present in the population (0.5 is considered)
- q : (1 - p)
- d^2 : the desired level of precision

Therefore, when considering a 10% margin of error, the equation would look as follows

$$n_0 = \frac{1.96^2 \text{x } 0.5 \text{ x } (1-0.5)}{0.1^2} = 96.04 \approx 97 \text{ QSs}$$

However, the above calculation is for an infinite population. Therefore, the sample size should be readjusted for the finite population using the following equation (Israel, 1992).

$$n = \frac{n_0}{1 + \frac{(n_0 - 1)}{N}}$$

Where,

- n : Sample size
- n_0 : Sample size for infinite population
- N : Population

According to NZIQS, the organisation had 1243 members, including all the membership grades, in 2021. Therefore, according to the above equation;

$$n = 97/ (1+ ((97-1)/1,243)) = 90.05 \approx 91 \text{ QSs}$$

Hence, the minimum sample size was ninety-one. The questionnaire was circulated via email with a link to the survey to all NZIQS members. The Survey allowed the collection of the responses for one month and collected ninety-two completed responses.

In the questionnaire, the respondents were asked to rate factors identified through the literature review (refer to Table 1) on a scale of one to five, where five means the factor is significant. At the same time, one is assigned to a factor of very low importance. Since the questionnaire survey only comprises closed-end questions, it is easier to code each possible response for each question using numbers. For instance, a five-point Likert scale would have six possible codes, one for each response and another for a missing response. In addition, the survey was designed to rank all 53 factors compulsory before submitting the response to ensure all data is present in all responses. Therefore, the collected 92 responses contain answers for all the factors.

Before conducting any statistical tests, it is required to test the normality of the data distribution. The Shapiro-Wilk test was used for that. The data is considered a normal distribution if the *p*-value is higher than 0.05. The mean rating or severity index (SI) representing the average responses for a particular variable was computed using the equation below (Knight and Ruddock, 2008).

$$SI = (\sum_{1}^{5} wi \text{ x } fi) \text{ x} \frac{100\%}{n}$$

Where,

SI : severity index

wi : rating points ranging from 1 to 5 on the Likert scale

fi : frequency of response, i.e., the number of responses associating a cost factor with a particular rating point

n : the total number of respondents rating a particular cost factor in the survey.

The coefficient of variation (COV) indicates the standard deviation (SD) as a percentage of the mean. COV helps compare the relative variability of different responses. The COV results will reflect the divergence or convergence of opinions among the participants in their ratings. A lower COV value means a higher agreement between all participants (Ji *et al.,* 2014).

However, COV cannot be used solely to decide the validity and reliability of the research design, measuring instrument, and findings. Therefore, Cronbach's Alpha is used to measure the reliability of the collected data. Generally, if Cronbach's Alpha of the data set is higher than 0.7, it is understood that the data set is reliable and valid (Bonett and Wright, 2015).

In addition, correlation analysis was conducted to evaluate the factors with the same rank. The correlations of the factors can support deciding the significance of the factors over those with the same rank. The correlation coefficient is varied between +1 and -1. +1 means a perfect positive correlation, while -1 means an ideal negative or no linear correlation between the two factors (Adafin *et al*., 2016).

The following section will analyse the significant findings of the systematic literature review.

3.5 Literature review

A considerable amount of research has been done to study the factors affecting cost overruns and their effects by categorising them differently. One of the most common classifications was technical; economic; contractual; psychological; and political factors (Plebankiewicz, 2018). However, it should be noted that environmental factors have yet to be included in this classification, even though they can affect the project's cost overruns. Nevertheless, most environmental factors are unpredictable, which can cause a considerable amount of cost overruns

for a project (Adam *et al.*, 2017; Allahaim and Liu, 2012; Ameh *et al.*, 2010; Johnson and Babu, 2018; Kumar, 2020). Therefore, those factors will be considered in this analysis. Furthermore, after reviewing various research articles in different countries in different contexts, such as building, high-rise projects, groundwater projects, highways, and infrastructure projects, Ameh *et al*. (2010) identified environmental factors as an essential category among the factor classification. Still, it needs to be clear how these unpredictable causes should be fully addressed or accounted for in a project cost estimate.

Allahaim and Liu (2012) took a different approach to studying cost overrun factors by categorising them into five groups: market volatility, pressure for distorting estimation, novelty, complexity, and time pressure. In contrast to Plebankiewicz's (2018) classification, Abdul *et al*. (2013) divided the factors into seven major fields: contractor related; design documentation related; financial management related; information and communication technology-related; labour management-related; material and machinery-related; and project management and contract administration related. This classification is more detailed than others as it covers every aspect of the construction project. Hence, it would be easy to identify which areas would be affected by the factor. Therefore, the above classification is adopted in this research with three additional and significant areas: environmental, psychological, and political factors. Hence, this research used a classification system with ten important categories to identify the factors. Table 1 below uses these ten categories with the factors identified under each category and how many recently published journal papers identified these factors as significant in cost overrun.

According to Table 1, most articles identified 'poor project management' as a significant cause of cost overrun. In addition, 'frequent design changes', 'unpredictable weather conditions', 'lack of experience in similar projects', and 'shortage of technical personnel' were also identified by a

higher number of articles as they were ranked within the top five in Table 3.1. On the other hand, the lowest two ranks were taken by 'deception' and 'manipulation of forecast'.

Table 3.1. *The factors leading to cost overruns in construction projects*

Category	Factors	Number of articles*	Rank based on the number of articles**
Contractor related	Poor site management and supervision	68	15
	Incompetent sub-contractors	46	23
	Poor planning and scheduling	62	18
	Poor monitoring and controlling	57	19
	Lack of experience in similar projects	**91**	**4**
	Inaccurate cost estimates	**76**	**10**
	Mistakes during construction/ rework	83	28
	Improper construction methods/ techniques	64	17
Project management and contract administration related	Tendering strategy	31	30
	Type of contract	28	33
	Procurement method	25	35
	Type of client	18	39
	Project complexity and scope	71	13
	Project duration	69	14
	Project location	50	21
	Poor risk assessment	23	37
	Poor project management	**93**	**1**
	Delays in decision-making and approvals	48	22
	Inaccurate quantity take-off	9	49
Design and documentation related	**Frequent design changes**	**92**	**2**
	Delays in designs	45	24
	Mistakes/ errors in designs/ drawings	**42**	**7**
	Inadequate tender documentation	75	11
	Variations and extra works	43	26
Financial management related	Financial difficulties of the Contractor	38	29
	Financial difficulties of the client	24	36
	Poor financial planning and control	21	38
	Delayed payments to the Main Contractor	28	33
	Delayed payments to the supplier/ sub-contractor	11	44
	Contractual claims such as the extension of time with cost claims	8	50
Information and communication technology related	**Lack of coordination and communication between parties**	**82**	**8**
	Slow information flow	10	45
	Poor team coordination	10	45
Labour management related	Poor labour productivity	56	20
	Shortage of site workers/ labours	74	12
	Shortage of technical personnel	**91**	**4**
	High cost of labour	43	26
	Lack of incentives	12	41
	Labour absenteeism	10	45
Material management related	Price fluctuation	66	16
	Shortage of materials	**81**	**9**
	Delay in delivery of materials/ equipment	45	24
	Unavailability or failure in equipment	29	32
Environment-related	**Unpredictable weather condition**	**92**	**2**
	Conflicts with neighbours	11	43
	Unforeseen ground conditions	**85**	**6**

Psychological related	Optimism bias among local officials	12	41
	The cognitive bias of people	10	45
	Cautious attitude towards risk	6	51
	Deception	5	52
Political related	Government initiatives	31	30
	Deliberate cost under-estimation	13	40
	Manipulation of forecasts	5	52

** Number of articles – How many journal articles identified each factor as a significant cause of cost overrun*
*** Rank based on the number of articles – rank 1 being the factor with the most cited number of articles and last rank with the lowest number of articles*

3.5.1 Contractor-related.

Eight factors were identified in the literature relevant to this category. Out of those, lack of experience in similar projects and inaccurate cost estimates were cited by more research compared to the other factors. The finding is arguably valid since a contractor without proper experience can lead to unnecessary cost overruns. At the same time, mistakes in the tender estimation will cause issues for the contractor's cash flow and may lead to disputes with clients (Adam *et al*., 2017; Ameh *et al*., 2010; Shah, 2016).

3.5.2 Project management and contract administration related.

Previous research has identified that the project size may also significantly affect the cost estimation because when the size of the project becomes more extensive, the probability of having cost overruns also becomes higher (Adam *et al*., 2017; Allahaim and Liu, 2012; Ameh *et al*., 2010; Flyvbjerg *et al*., 2004; Creedy, 2006; Ji *et al.*, 2014). However, based on the geographical location, the project size and complexity definition parameters can be changed. For instance, a mega-scale project in a developing country may be defined as a medium or small-scale project in a developed country (Adam *et al.*,2017). As the project size or the scope is known during the pre-contract stage, the impact of this factor as a variable on the cost overrun should be minor as it is predictable. However, the research did not identify the cause of why this predictable factor becomes significant in cost overruns (Adam *et al.*, 2017). On the other hand, the study acknowledged that the project's location could be a critical cause of cost overruns, especially for a TI rather than a building project

(Mbachu and Cross, 2015). Most of the factors under this category represent the general project characteristics of the projects. However, based on the literature, these factors only significantly impact project cost if poor project management, which is the highest cited factor in the literature.

3.5.3 Design and documentation-related.

Previous research identified that most of the cost overruns occur during the planning and design stage due to technical and administrative matters of the design team (Lind and Brunes, 2015; Plebankiewicz, 2018). However, Mbachu and Cross (2015); Shanmugam *et al.* (2006) showed that variations are the most significant factor impacting cost overruns. However, according to the literature, variations did not rank significantly. Conversely, frequent design changes and design mistakes were ranked within the top ten based on the literature. Furthermore, Mbachu and Cross (2015) explained that the quality of design documentation and unbalanced bidding at the tendering stage also significantly impact cost overruns.

3.5.4 Financial management-related.

A study by Abdul *et al*. (2013) based on 274 questionnaires done among professionals in Malaysia identified the contractor's and client's financial difficulties and payment delays as the most common causes of cost overruns. However, considering the number of research that identified these factors, it was noticed that none of the factors in this category was significant compared to the top ten factors.

3.5.5 Information and communication technology-related.

Three factors were identified that affect the project cost due to the issues in communication and coordination issues. However, the lack of coordination and communication between the parties was primarily cited in the literature, as this issue will lead to disputes and delays (Adam *et al.*, 2017; Abdul *et al*., 2013).

3.5.6 Labour management-related.

Generally, 40% of projects' essential cost is considered labour cost (Masu *et al.*, 2012). Therefore, it is vital to consider the factors related to the workforce. Nevertheless, the shortage of technical personnel was the significant factor in this category, while the shortage of site labours also achieved rank 12 based on the literature. But the other factors were not that substantial.

3.5.7 Material management-related.

Adam *et al*. (2017) also found that price fluctuation causes a significant impact on cost overrun and further explained that 20 – 25% of the cost overruns were based on price fluctuations. However, price fluctuations should also be predictable based on construction cost indices. Hence, it is unclear why a predictable factor significantly impacts project cost overruns. Further, most of the projects allow additional claims for price fluctuations. Therefore, predicting the cost during the tendering in such projects is not required.

3.5.8 Environment-related.

Unpredictable weather conditions and unforeseen ground conditions are the most significant factors in this category. The other remaining factor is a conflict with neighbours, which is insignificant towards project cost.

3.5.9 Psychology-related.

Psychology is an essential factor that can impact the project cost. According to the researchers, there are four factors under this category. However, they were insignificant according to the literature compared to the other factors. It was widely recognised that strategic misrepresentation and deception cause under-estimation construction projects in common and more critically in construction projects (Adam *et al*., 2017; Allahaim and Liu, 2012; Cantarelli *et al.*, 2012;

Flyvbjerg *et al*., 2003; Flyvbjerg *et al*., 2002). In addition, an optimism-biased decision-making tendency is another side of this failure of proper estimation (Allahaim and Liu, 2012).

3.5.10 Political-related.

Under this category majority of the researchers identified three significant factors government initiatives, deliberate cost under-estimation and manipulation of forecasts. However, these factors ranked less based on the literature. Mainly the manipulation of the estimates was identified only in five articles.

Ji *et al*. (2014), Adafin *et al*. (2016), and Zhao *et al*. (2019) studied the NZ context to identify the factors affecting the accuracy of the pre-contract stage estimation of building projects, which impacts the cost overruns. The identified factors were categorised into six different categories. Those research emphasised that inadequate tender documentation is the most critical factor affecting tender estimation. The other top factors influencing the estimate were 'the complexity of design and construction', 'completeness of project information', 'insufficient estimating time', and 'the client's financial ability'. The study implies that inaccurate pre-contract estimations will likely lead to cost overruns. Although the above results focused on tendering, the above factors may also impact the preliminary estimation (Adafin *et al*., 2016; Zhao *et al*., 2019). However, the above NZ-based studies were all focused on building projects. Considering the differences in type, magnitude, scope, duration, and risk involved in the building versus TI projects, the above results can only be used to develop a cost model for TI projects after an in-depth investigation.

The government reports of NZ stated that the coming years are planned for significant investments in the TI development in the country. For example, $ 16.3 billion is allocated for the "Auckland transport alignment" project, while $ 3.8 billion is earmarked for the "Let us get Wellington moving" project. In addition, the "road to zero" project will be initiated with a budget of $ 10 billion while another $ 1.2 billion is allocated for developing the NZ rail network. Therefore, it is

vital to have a reliable cost estimate at the beginning of the project. Otherwise, these public funds will not be well-spent due to cost overruns. Hence, this study aims to identify the causes of cost overrun of NZTI projects aiming to improve the current estimation practices. Although the literature identified 53 factors that impact the project cost, the severity of the impact cannot be determined based on the literature. Hence, a survey needs to be conducted on NZ professionals to identify the significant factors. The following section will summarise the survey's findings and data analysis.

3.6 Survey results and analysis

Ninety-two participants completed the survey. Of these participants, 93% practice as QSs, while 7% work in project management and other related professions. Further, 43% of the participants have had five years of experience, 21% with 11 to 20 years of experience, and 19% with more than 20 years in construction. 80% have up to five years of experience, and the rest have six to ten years of experience in TI projects in NZ.

Table 3.2 summarises the survey results and contains the normality test *p*-value, Cronbach's Alpha, average SI, and average COV for each sub-category. First, the data were tested for normality using the Shapiro-Wilk test, and the *p*-value was higher than 0.05. Therefore, the data set can be considered normally distributed (refer to Table 3.2). Second, the reliability of the data set was tested using Cronbach's Alpha. According to Bonnet and Wright (2015), the data set is considered reliable if the Alpha value is greater than 0.7. Table 3.2 shows the Cronbach's Alpha of each category, and all the values are greater than 0.7. Therefore, the collected data can be considered reliable.

Table 3.2. *Data analysis*

	Factors	Normality test *p-value*	Cronbach's Alpha	SI	Total Rank	Group Rank	SD	COV	Coefficient of correlation						
									3.2.1.2	3.2.8.1	3.2.6.1	3.2.6.2	3.2.10.1	3.2.1.4	3.2.2.3
3.2.1	**Contractor-related**		0.921	70%				23.77%							
1	Poor planning and scheduling	0.067		84%	**2**	1	0.78	18.79%	.816**	0.330	.554*	0.459	.510*	.568*	0.275
2	Inaccurate cost estimates	0.086		78%	**6**	2	0.89	22.67%	-	-0.027	0.491	0.248	.576*	.607*	-0.083
3	Poor site management and supervision	0.072		76%	**8**	3	0.83	21.94%	.814**	-0.029	.522*	0.414	0.450	0.485	-0.133
4	Incompetent sub-contractors	0.065		69%	15	4	0.81	23.42%	0.491	.683**	-	0.020	.847**	.590*	0.370
5	Lack of experience in similar projects	0.095		68%	17	5	0.73	21.53%	.759**	0.211	.534*	0.456	.517*	0.435	0.107
6	Mistakes during construction/ rework	0.078		59%	34	6	0.79	26.44%	.759**	0.403	.543*	.522*	.566*	0.373	.499*
7	Improper construction methods/ techniques	0.089		58%	35	7	0.92	31.57%	0.489	0.353	0.461	0.321	.497*	0.068	0.466
3.2.2	**Project management and contract administration-related**		0.854	65%				26.14%							
1	Poor project management	0.059		75%	**9**	1	0.81	21.42%	.752**	-0.076	0.236	.610*	0.377	.590*	-0.139
2	Inaccurate quantity take-off	0.065		75%	**10**	2	0.85	22.78%	-0.033	0.288	-0.200	.767**	-0.249	-0.253	.647**
3	Tendering strategy	0.068		69%	15	3	0.78	22.36%	0.248	0.181	0.020	-	-0.054	0.152	0.300
4	Type of contract	0.069		66%	18	4	0.75	22.73%	-0.350	0.126	-0.324	0.070	-0.228	-0.234	0.439
5	Project duration	0.058		65%	19	5	0.83	25.51%	0.334	0.387	0.402	0.234	0.496	0.383	0.236
6	Project complexity and scope	0.078		65%	20	6	0.84	26.03%	0.355	0.251	0.214	0.436	0.326	.730**	0.065
7	Lack of experience in similar projects	0.082		64%	21	7	0.76	23.66%	0.112	0.496	0.222	0.221	0.253	-0.069	.745**
8	Procurement method	0.084		64%	23	8	1.01	31.91%	0.209	.666**	0.411	0.226	.525*	0.132	.690**
9	Project location	0.054		64%	23	8	0.86	26.96%	.751**	0.494	.731**	0.188	.831**	.538*	0.399
10	Delays in decision-making and approval	0.091		63%	25	9	1.00	31.90%	0.102	.702**	0.348	0.007	.532*	0.205	.691**
11	Poor risk assessment	0.084		62%	29	10	0.78	25.29%	.737**	0.494	.683**	0.377	.814**	.821**	0.232
12	Type of client	0.092		51%	42	11	0.85	33.09%	.501*	0.472	.740**	0.439	.514*	.578*	0.263
3.2.3	**Design and documentation-related**		0.890	80%				20.02%							
1	Frequent design changes	0.100		95%	**1**	1	0.78	16.44%	0.472	0.113	0.151	0.387	0.483	.554*	-0.101
2	Poor and incomplete tender documentation	0.123		82%	**3**	2	0.72	17.44%	0.331	0.252	0.199	0.167	.608*	0.422	0.131
3	Delays in design	0.082		79%	**4**	3	0.79	19.78%	0.201	0.268	0.186	0.056	0.400	0.035	0.247
4	Mistakes/ errors in design and drawings	0.059		79%	**5**	4	0.76	19.40%	.508*	-0.141	0.008	0.410	0.306	0.337	-0.185
5	Variations and extra works	0.076		64%	21	5	0.87	27.04%	-0.024	.632**	0.359	0.129	0.494	0.144	0.406

3.2.4	**Financial management-related**		0.912	55%				30.51%						
1	Poor financial planning, such as cash flow forecast	0.095		74%	11	1	0.83	22.58%	0.399	0.491	.665**	-0.124	.524*	0.296	0.486
2	Financial difficulties of the client	0.058		62%	27	2	0.72	23.03%	0.250	.552*	0.453	-0.038	.559*	.538*	0.443
3	Contractual claims such as EOT with cost claims	0.068		62%	29	3	0.70	22.72%	.756**	0.274	.681**	0.256	.578*	.556*	0.253
4	Financial difficulties of the contractor	0.057		47%	47	4	0.84	35.53%	0.152	.680**	.647**	-0.030	0.495	0.174	.679**
5	Delay payments to the sub-contractors/ suppliers	0.082		44%	51	5	0.92	42.39%	0.302	.638**	.652**	-0.114	.586*	0.259	.616*
6	Delay payments to the main contractor	0.102		41%	52	6	0.75	36.81%	0.144	.841**	.656**	0.121	.683**	0.272	.733**
3.2.5	**Information and communication technology-related**		0.934	58%				31.45%						
1	Lack of coordination and communication between parties	0.091		63%	25	1	0.88	27.93%	0.234	.736**	.506*	0.016	.605*	.514*	.588*
2	Slow information flow	0.058		61%	32	2	0.87	28.52%	0.158	.772**	0.452	0.095	.600*	0.344	.704**
3	Poor team working skills	0.061		50%	43	3	0.95	37.89%	0.390	.688**	.737**	-0.172	.832**	.514*	.515*
3.2.6	**Labour management-related**		0.958	64%				24.58%						
1	Shortage of labour	0.105		73%	12	1	0.90	24.78%	.576*	.552*	.847**	-0.054	-	.759**	0.199
2	Shortage of technical personnel	0.089		73%	12	2	0.87	23.87%	.607*	0.305	.590*	0.152	.759**	-	-0.107
3	Poor labour productivity	0.095		62%	29	3	0.82	26.48%	.741**	0.319	.677**	0.223	.795**	.910**	0.037
4	High cost of labour	0.066		60%	33	4	0.77	25.57%	0.488	.604*	.784**	0.065	.897**	.793**	0.192
5	Labour absenteeism	0.081		56%	36	5	0.72	24.91%	.818**	0.410	.673**	0.295	.690**	.703**	0.324
6	Lack of incentives	0.118		58%	38	6	0.66	23.48%	0.471	.518*	.556*	0.177	.755**	.860**	0.291
3.2.7	**Material management-related**		0.964	48%				36.92%						
1	Unavailability or failure of equipment	0.072		49%	44	1	0.95	38.72%	-0.056	.542*	.650**	-0.435	.569*	0.191	0.275
2	Shortage of materials	0.059		48%	45	2	0.84	35.01%	0.099	.649**	.762**	-0.212	.618*	0.328	0.459
3	Delay in delivery of materials/ equipment	0.063		48%	46	3	0.87	36.73%	-0.013	.716**	.664**	-0.073	.587*	0.219	.626**
4	Price fluctuation	0.084		46%	49	4	0.87	37.23%	0.253	.727**	.815**	-0.091	.781**	0.435	.536*
3.2.8	**Environment-related**		0.820	62%				24.65%						
1	Unforeseen ground conditions	0.135		78%	**6**	1	0.61	15.65%	-0.027	-	.683**	0.181	.552*	0.305	.810**
2	Unpredictable weather conditions	0.086		62%	27	2	0.76	24.31%	0.083	.775**	.506*	0.249	0.400	0.014	.912**
3	Conflicts with neighbours	0.075		47%	47	3	0.80	34.00%	0.258	0.343	0.315	.651**	0.453	0.399	0.170
3.2.9	**Psychology-related**		0.913	47%				35.86%						
1	Optimism bias among local officials	0.095		56%	40	1	0.76	27.14%	0.171	.530*	0.318	0.206	0.224	0.098	.768**
2	Cautious attitude towards risk	0.058		56%	40	2	0.58	20.87%	-0.043	0.485	0.088	0.387	0.034	-0.037	.790**

3	The cognitive bias of people	0.106		44%	50	3	0.76	34.38%	0.365	0.205	0.139	0.312	0.289	0.380	0.408
4	Deception	0.085		30%	53	4	0.92	61.04%	0.432	0.414	0.400	0.398	0.432	0.322	.611*
3.2.10	**Political-related**		0.844	62%				28.55%							
1	Deliberate cost underestimation	0.096		73%	12	1	1.05	28.90%	-0.083	.810**	0.370	0.300	0.199	-0.107	-
2	Manipulation of forecasts	0.091		58%	36	2	0.63	21.89%	.519*	0.052	0.096	0.435	0.286	.552*	0.101
3	Government initiatives	0.052		56%	38	3	0.98	34.86%	.501*	0.436	0.383	0.490	0.477	0.485	.531*

Legend: SI – Severity Index; SD – Standard Deviation; COV – Coefficient of Variation
***. Correlation is significant at the 0.01 level (2-tailed).*
**. Correlation is significant at the 0.05 level (2-tailed).*

Ten factors were identified with SI on or above 75%, namely, 'frequent design changes', 'poor planning and scheduling', 'inadequate tender documentation', 'delays in design', 'mistakes/ errors in design and drawings', 'unforeseen ground conditions', 'inaccurate cost estimates', 'poor site management and supervision', 'poor project management', and 'inaccurate quantity take-off'. Out of these ten factors, eight relate to the pre-contract stage. 'Poor site management and supervision'; and 'poor project management' are the only factors associated with the post-contract stage. Except for three factors, all other factors have an SD of less than 1. However, the SD values are considerably high, indicating that the selection spread is higher. A higher SD means a higher COV. Therefore, any factor with higher COV shows high dispersion. So lower COV means the agreement among the participants is higher (Knight and Ruddock, 2008). The following sections will discuss the findings under each category.

3.6.1 Contractor related-factors

Three factors in this category achieved more than 75% of the SI. Those factors were ranked within the top ten (two, six, and eight) of the most highly rated factors overall, which means that professionals believe these three factors significantly impact the cost of the TI projects in NZ and can therefore be responsible for cost overruns. According to the SDs, the average value is spread away from the mean. Conversely, five factors achieved COV of less than 25%. Therefore, the lower COV means industry professionals' agreement on the selection was comparatively similar. Thus, the fact that the respondents agreed about the severity level of the factors suggests that 'poor planning and scheduling', 'inaccurate cost estimates', and 'poor site management and supervision' critically impact the cost of the NZTI projects.

3.6.2 Project management and contract administration-related factors

Two factors achieved 75% of the SI with an overall ranking of 9 and 10, namely 'poor project management' and 'inaccurate quantity take-off'. They both achieved higher SD and COV below

23%. Therefore, although the agreement among the parties is comparatively high, the average deviated considerably from the mean. Even though 'tendering strategy', 'type of contract', 'project duration', and 'project scope and complexity' were essential factors in estimation, they were ranked 15, 18, 19, and 20, respectively, having a SI between 65% - 69%. Then, 'lack of experience in similar projects', 'procurement method', 'delay in decision making and approval', and 'poor risk management' were ranked between 21 to 30 overall, with an SI between 62% to 64%. According to the results, the type of client makes the most negligible impact on the cost under this category. Three factors show COV exceeding 30%. Because even though the factors achieved a higher rank, the experts have different ideas on these factors.

3.6.3 Design and document-related factors

Four factors were ranked among the top five in the overall ranking. All four achieved over 79% of SI, and COVs were below 20%. Therefore, the project's design critically impacts the cost in various ways.

3.6.4 Financial management-related factors

Financial management is a critical aspect of any project. However, as per Table 3.2.4, all six factors in this category were considered insignificant compared to the others. In contrast, 'poor financial planning' achieved the 11^{th} rank with an SI of 74% and COV of 22%.

3.6.5 Information and communication technology-related factors

As listed in Table 3.2.5, 'lack of coordination and communication between parties' and 'slow information flow' were rated as important factors. The agreement among the participants regarding ranking these two factors is moderately higher, with COVs between 27% and 29%. Although poor teamwork can lead to many disputes and issues in any project, the participants ranked this cause as less critical at 43 out of 53 factors.

3.6.6 Labour management-related factors

As captured in Table 3.2.6, 'shortage of labours' and 'shortage of technical personnel' are equally essential as they achieved the same SI and ranked 12 overall. The other four factors were ranked above 29, with SIs between 56% and 62%, which means those factors will have a medium impact on the project cost.

3.6.7 Material management-related factors

Materials and equipment are essential in the construction industry as they are part of the leading project components. However, compared with other factors, the factors related to materials and equipment play a minor role in cost overruns. All the factors (Table 3.2.7) achieved SIs below 50% but COVs above 35%. That means there may be doubt among the experts regarding the significance of the factors.

3.6.8 Environment-related factors

Environment-related factors are not easily foreseen beforehand and are hard to control. As shown in Table 3.2.8, 'unforeseen ground conditions' achieved 78% for its SI, and the level of agreement was considerably high as the COV was 15.65%. Therefore, this factor ranked 6th overall. Also, 'unpredictable weather conditions' can affect the cost in many ways. Nevertheless, compared to the other factors, its level of importance was only 62% and ranked 27.

3.6.9 Psychology-related factors

The factors related to the construction team's psychology are often considered significant, especially in government-related projects. Since TI projects are often public-funded, these factors must be addressed concerning their effects on the cost. Table 3.2.9 shows four factors identified for this category, none of which shows any critical importance. All of the factors ranked 40 and above. 'Deception' was considered to have the lowest relevance among all factors. However, the

COV of ’deception’ is considerably high, as much as 61%. Therefore, there can be changes in this factor's actual level of importance on the TI project cost.

3.6.10 Political-related factors

Whether a project is public or private, the political background can affect a project in many ways. As shown in Table 3.2.10, ‘deliberate cost underestimation’ ranked 12 with a 73% SI. However, the SD of the above was 1.05, while the COV was 28.9%. Therefore, the results may not indicate the real significance of the factor. ‘Manipulation of forecasts’ and ‘government initiatives’ did not achieve significant ranks.

3.6.11 Correlation analysis

Several sets of factors ranked the same due to the similar SI. Therefore, it is required to analyse the correlation coefficient to identify the most important factor out of each set. Table 3.2 shows the correlation matrix for the factors with the same rank within the top 20. Ranks below 20 were not considered table due to the less importance of the factors. The first set was ‘inaccurate cost estimates’ (3.2.1.2) and ‘unforeseen ground conditions’ (3.2.8.1) in rank 6. According to the correlation matrix, 2.8.1 shows more significant correlations with other factors than 3.2.1.2. Therefore, out of these two factors, 3.2.8.1 is considered more important. Similarly, ‘labour and technical personnel shortage’ (3.2.6.1 and 3.2.6.2) and ‘deliberate cost underestimation’ (3.2.10.1) achieved a rank of 12. However, according to the correlation matrix, 3.2.6.1 and 3.2.6.2 show higher correlation coefficients than 3.2.10.1. It indicates that the labour and technical personnel shortage have similar importance and correlation with other factors. Finally, rank 15 was achieved by ‘incompetent sub-contractors’ (3.2.1.4) and ‘tendering strategy’ (3.2.2.3). According to Table 3.2, 3.2.1.4 shows a significant correlation with other factors compared to 3.2.2.3.

The following section will analyse, compare, and discuss the above survey findings and data analysis against the literature review.

3.7 Discussion

According to the survey findings, the top eight factors were related to the pre-contract stage. That contrasts with previous findings for NZ, as the top three factors identified by Abdul *et al*. (2013) were all associated with the post-contract stage. Those three factors were 'price fluctuation', 'financial difficulties of the contractor', and 'poor site management and supervision'. Although 'poor site management and supervision' also featured in the top 10 ranking in this study, it was only positioned in eighth place. To some extent, the findings aligned with Ameh *et al*. (2010), who found that 'lack of contractor experience', 'frequent design changes', and 'economic stability' were the most critical factors in their study. However, compared to the findings of Ameh *et al.* (2010), this analysis showed a lower ranking for 'lack of contractor's experience' while 'frequent design changes' ranked first. Similarly, Brunes and Lind's (2014) study also listed 'frequent design changes' among the most critical factors.

The reason for the above outcome could be due to the selected sample. The survey results are the perception of QSs, who predominantly examine the estimation factors during the pre-contract stage. From QS's perspective, site management may not be vital as one of the top seven factors. But, if the survey is carried out with the site management team, then more weight may be given to the post-contract management factors.

The results of this study contrasted with the literature when considering the factor groups as opposed to individual factors. Thus, conferring to this study, management-related factors were ranked the highest, and the communication and psychology-related factors groups were ranked the lowest in Adam *et al.* (2017) study. However, this study found that the highest-ranking factors are related to the design-related group, which is also confirmed in the overall literature ranking (Table

3.1). On the other hand, the literature identified the most crucial factor as poor project management, whereas this study changed that to poor site management. Other than 'poor site management and supervision', all other factors related to the contractor's expertise were ranked outside the first ten. Generally, before awarding the contract, the consultant carries out a detailed technical evaluation is carried out to find the expert. Therefore, it is rare that these issues can occur. However, insufficient skilled personnel or labours may lead to poor site management because the contractor may be urged to use unskilled staff and labours. Accordingly, our study ranked labour and skill shortage factors higher than the literature. That can be due to the current skill shortage faced by NZ due to the various border restrictions.

On the other hand, the material-related factors were ranked with low importance. That means all the issues related to materials and equipment will not significantly impact the cost overrun compared to the other factors. Finally, although price fluctuation is a persistent problem faced by construction projects, it still needs to be identified as a critical factor, as most projects allow price fluctuation claims. Therefore, it is possible to forecast the trend using indices to a greater extent.

Remarkably, "unforeseen ground condition" achieved rank 6 in survey results and literature. The only financial-related factor ranked with importance was poor financial planning. Although the financial difficulties of the contractor and the client are significant issues in the construction projects, the survey did not rank them with high importance.

Six factors within rank 11 to 20 are significant in defining a project. These factors are: 'tendering strategy', 'procurement strategy', 'contract type', 'project duration', 'project scope', and 'location'. These factors could help the client to decide which method/ type of each factor is most cost-effective. Ji *et al.* (2014) also explored factors affecting costs, but while their study was also based in NZ, the main focus was on the accuracy of the tender price of the building projects. Hence, their findings cannot be compared with our study. However, Ji *et al.* (2014) ranked 'incompetent

tender document', 'design and construction complexity', and 'completeness of the project information' in the top three. Nevertheless, regardless of the project type, 'poor tender documentation' was identified as a significant factor.

According to Table 1, 'lack of experience in similar projects', 'lack of coordination and communication between parties', 'shortage of technical personnel', 'shortage of materials', and 'unpredictable weather conditions' were cited by a higher number of articles making these factors ranked among the top ten factors. However, the survey results did not identify these factors as crucial in project cost. Nevertheless, based on the sample perceptions, experience level may affect these decisions. Hence, it is not suggested to disregard these non-critical factors identified in this study completely.

According to the above comparison of the survey findings with the literature, there were similarities and discrepancies. Although, as Flyvbjerg *et al.* (2003) found, the cost overrun problem has remained the same over the past 70 years all over the world in the construction industry. Therefore, this common global challenge urge for a resolution. Furthermore, most research agreed that incomplete design and documentation heavily contribute to significant cost overruns. Hence, both consultant and the client should pay attention to improving the documentation quality before tendering. In addition, Flyvbjerg *et al.* (2004) and Harrera *et al.* (2020) emphasised that the design team should be more careful in project planning if the project delivery is comparatively long.

Consequently, it can be added that adequate planning is mainly required for larger projects rather than smaller ones (Flyvbjerg *et al.,* 2004; Cantarelli *et al.,* 2012). This remark is primarily compelling for TI projects since they are large-scale projects with extended construction periods. Conversely, Andrić *et al.* (2019) argued that the project size would not affect the magnitude of

cost overrun. Similarly, our study found that the project scope is more significant than the project size.

The primary aim of this study was to improve the traditional estimation practice in NZTI projects. The study identified various factors with quantitative and qualitative impacts on projects cost. Therefore, the results of this study cannot incorporate into the traditional estimation methods as those considering the technical variables of the project predominantly. However, a statistical analytical tool can be developed for cost modelling while combining this study. As the factors mentioned above become the project's risks, combining a risk analysis and estimation tool with a statistical cost model is feasible to improve the performance.

Based on the above discussion, the next section will conclude the research problem against the findings and make recommendations while highlighting the research limitations.

3.8 Conclusions and recommendations

The study investigated the factors affecting the cost overruns of NZTI projects. Nevertheless, using studies from other countries to make decisions on the NZTI projects is only possible with a proper investigation. So instead, a survey was carried out to collect data through an online questionnaire distributed to QSs with experience in NZTI projects. The results were thus based on the participant's perceptions depending on their experience level.

Fifty-three factors were identified that affect cost overruns of TI projects through an SLR. These factors were categorised into ten classification groups. Survey results showed the three most critical factors are ‘frequent design changes’, ‘poor planning and scheduling’, and ‘inadequate and incomplete tender documentation’. At the other end of the spectrum, ‘delay payments’ and ‘deception’ were the lowest-ranking factors. The findings were then compared with literature research findings. Although the ranking orders are slightly different, similarities were found.

Out of the top 10 factors, eight were related to issues encountered in the pre-contract stage. Thus, the findings suggest that the design team should focus on finalising the design and site investigation before the tendering phase to avoid issues in the tender documentation. When the post-contract stage begins, the consultant team should focus more on mitigating design delays, frequent design changes, and project management issues. On the other hand, the contractor's team shall pay attention to the site management and supervision capabilities and planning and scheduling errors. Consequently, the findings support the project team from inception to completion to monitor the project's cost and minimise cost overruns by conducting a thorough risk factor analysis. However, it is required to highlight that this argument is based on QS's perception. Therefore, to have a proper risk factor analysis, gathering information from all the project professional insights is recommended.

This study established that design documentation and site management are the crucial controllers of cost overruns. Furthermore, it can be emphasised that although these factors are well-known, no action has yet to be taken to control them. From the QS perspective, the control can be done through estimation. However, all the traditional cost estimation models use technical variables while ignoring the major non-technical parameters identified in this study. Therefore, it is recommended to incorporate these factors into a cost model to improve the reliability of the current estimation practice. The factors used for the survey were identified through the SLR only. Since the survey was carried out among NZ QSs, the ranking order of these factors may vary in other countries and professionals. Nevertheless, the findings can be used to improve the current estimation practice in NZTI projects.

3.9 Epilogue

This chapter introduced several key information in relation to the cost estimation, construction risk factors and causes of cost overruns for NZTI projects. That includes the main possible risk factors

affect NZTI project cost and the severity of the factors. The chapter also identify the risks that will affect significantly on the project cost factors, because these factors should be addressed in the cost estimation process. It also, emphasised that the traditional estimation process does not consider the risk factors in the NZTI cost estimation. Next step is to identify the severity of the impact of these causes on the project. If e projects face significant cost overruns, then based on that, a solution must be found to mitigate the overruns. Therefore, the next chapter investigate the cost overruns in NZ road projects as a case study.

4 Significance of the cost overruns in road projects in New Zealand

4.1 Prologue

Road infrastructure plays a vital role in the socio-economic development of nations, facilitating the movement of goods, people, and services. However, despite its importance, road construction projects often face challenges that hinder their successful completion within allocated budgets. Cost overruns are among the most persistent and disruptive issues in road construction projects. Cost overruns occur when the actual expenses of a project exceed its initial budgeted estimates. These overruns strain public finances and lead to delayed project delivery, compromised quality and negative impacts on communities and the environment. This chapter investigates the severity of cost overruns in New Zealand (NZ) road projects. The background study in Chapter 1 identified that the cost overrun is an ongoing and common issue in the construction industry. Therefore, this thesis investigates a solution to the problem in NZ Road projects. Before investigating a solution, this chapter explores whether cost overrun is a significant problem in NZ Road projects. Hence, this chapter sets the scene for further research on cost estimation of road projects by identifying the impact of cost overruns on the project size and time. For the study mentioned above, this chapter considers the contract price to decide the size of the project and the year of completion as the time.[2]

[2] This chapter is based on the following journal paper (under review).

Atapattu, C.N., Domingo, N., and Sutrisna, M. (2023). How significant is the cost overrun in road projects in New Zealand? *Developments in the Built Environment*

4.2 Abstract

Cost performance in a construction project is crucial to the project's success. However, construction projects often completed significantly deviated from the budget. This research investigated the significant deviation in New Zealand (NZ) road projects. Data were collected from 106 projects completed between 2002 and 2022. The project cost overruns were compared against the time and the project size. The Results were validated through non-parametric tests. The study found that the cost overruns of NZ road projects will significantly depend on the project size and the duration. The mean percentage of cost overrun (MPCO) of road projects was 20%. Furthermore, the cost performance of road projects has shown no significant improvement over the past 20 years. Although much research identified the importance of finding a solution to the cost overrun matter, primary research has yet to be conducted to study the NZ road projects. Therefore, the findings of this research can be used to investigate further to find a solution to minimise the consequences. Therefore, the findings and conclusions may differ for contexts other than NZ road projects. Subsequently, this study emphasises that the current estimation practice in NZ road projects needs significant improvements.

4.3 Introduction

Various researchers defined the term "cost overrun" in numerous ways. For example, Endut et al. (2009) explained cost overrun as an extra cost beyond the contract price agreed upon during the tender. Conversely, other researchers defined it as the ratio of the additional cost to the estimate made at the decision-making stage of the project to go ahead (Love et al., 2016; Flyvbjerg et al., 2002; Odeck, 2019). Therefore, it compares the initial budget with the actual final cost of the project. In a similar view, Danisworo and Latief (2019) defined the term as the actual costs that exceed the budget. Furthermore, Odeck (2019) emphasised that cost overrun should not be

described as cost escalation, which is used to explain the effects of inflation on project cost. Therefore, based on the widely held views of the researchers, our study also defines the cost overrun as the ratio of extra cost to be expended to accomplish the actual final cost of the project compared to the preliminary stage project budget.

Cost-benefit analysis (CBA) plays a massive part in the decision-making process. The decision to proceed with the project will be taken based on the CBA results, and the initial project estimate is part of the analysis. (Ahiaga-Dagbui et al., 2017; Chambers et al., 1971). Nicolaisen et al. (2012) and Sodikov (2005) identified that CBA was used as a widely used tool in infrastructure projects to evaluate the benefits of the project economically to the country, and it demands a high accuracy level of the initial estimate. Otherwise, it will also affect the public because most infrastructure projects are funded by the public sector (Creedy et al., 2010).

According to the studies, cost overruns happen in construction projects in various magnitudes (Nijkamp and Ubbels, 1999; Odeck, 2004; Ellis et al., 2007; Lee, 2008). Based on their findings, the researchers provided various guidelines and solutions in other contexts. For instance, several cost models were developed to overcome the issues in estimation (Adel et al., 2016; Ahiaga-Dagbui et al., 2013; Shr and Chen, 2006; Sodikov, 2005)

In contrast, no study was carried out for transportation infrastructure projects in New Zealand (NZ) as the researchers mainly focussed on building projects (Adafin et al., 2020; Ji et al., 2014; Zhao et al., 2019). The NZ government reports say that the expected investment in infrastructure is significantly high in the coming years (Ministry of Transport NZ, 2020; NZTA, 2020). Therefore, this study aims to identify the significance of the cost overruns on TI projects in NZ. However, only a few recently completed projects are available for rails and tunnel projects in NZ. Hence, the focus will be narrowed down to NZ road projects. Therefore, this study will enable researchers and industry practitioners to understand the severity of the cost overrun issue. Based on the results,

conclusions will be drawn to evaluate whether the current estimation practice in road projects is reliable or needs improvements for policymaking and planning. The following section outlines the research methodology adopted for this study.

4.4 Literature review

In particular, infrastructure projects face cost overruns frequently and significantly compared to the other sectors (Adam et al., 2017; Ahiaga-Dagbui et al., 2017; Brunes and Lind, 2014; El-Sawalhi, 2015; Flyvbjerg et al., 2003; Narayanan et al., 2019; Touran and Lopez, 2006). Cost overrun in infrastructure projects has been considered one of the significant global issues without any improvement, control, or solution for the last 70 years (Flyvbjerg et al., 2003). That can affect the planning and programming of current and future projects (Creedy et al., 2010). Several researchers have recognised the importance of infrastructure development towards the economic growth of a country, especially transportation infrastructure (TI) projects such as highways, bridges, tunnels, and railways (Flyvbjerg et al., 2003; Love et al., 2016; Odeck, 2019). However, a few research studies were done on the cost of the TI projects, and most were either not recent or smaller-scale case studies (Creedy et al., 2010; Flyvbjerg et al., 2003; Flyvbjerg et al., 2002).

However, according to the facts, the projects are experiencing a mean percentage of cost overrun (MPCO) of 28% (Flyvbjerg et al., 2003; Huo et al., 2018; Mok et al., 2015). Lind and Brunes (2015) studied Swedish infrastructure projects. They identified that cost overruns occur at the initial stages of design and planning until the design is finalised due to technical and administrative issues. Considering the avoidable causes of cost overruns, the project team should pay more attention to these pre-construction issues (Lind and Brunes, 2015).

In a different view, according to Seeley (1996), the vital duty of a Quantity Surveyor is cost management while giving the value for money set against perceived expectations. Cost

management comprises estimating, monitoring, and controlling (Owens et al., 2007). Hence, cost control alone cannot address the issue of cost overruns. Nevertheless, more than focusing only on the pre-contract stage issue, it is required to look at the overall picture. The cost monitoring and controlling steps will also fail if the cost estimate fails (Malkanthi et al., 2017; Owens et al., 2007). Therefore, accurate project estimates are essential to delivering the project within the budget (Malkanthi et al., 2017). Hence, to execute the project within the expected budget, the accuracy of the estimation is required so that the decision to go ahead with the project can be taken based on a reliable project budget (Malkanthi et al., 2017).

Although it is recognised that the project's successful completion depends on the achievement of time, cost, and quality factors, it is rare to complete a project within the budget without facing cost overruns (Endut et al., 2009). Considering its extended construction duration, it is a common phenomenon in infrastructure projects. It is found that infrastructure projects with significant cost overruns are much more common than projects finished within or under the budget (Odeck, 2019).

It was emphasised that nine out of ten infrastructure projects faced cost overruns, and the probability of the actual cost being higher than the estimated budget was 86% (Flyvbjerg et al., 2002). Furthermore, in their major studies, Flyvbjerg et al. (2002) and Flyvbjerg et al. (2003) concluded that the Mean Percentage Cost Overrun (MPCO) faced by infrastructure projects, in general, was 28%. For roads and highway projects, it was around 21%. Moreover, rail and fixed link projects such as bridges and tunnels were recorded with significantly different MPCOs, 45% and 34%, respectively. The study used two hundred and fifty-eight infrastructure projects from twenty nations covering all five continents. On the other hand, Odeck (2004) researched the cost overruns of six hundred and twenty road projects in Norway and concluded that the cost overruns could be as high as 183%, which makes it a critical matter to be addressed, especially in public-funded projects. In contrast to Flyvberg et al. (2002) work, Endut et al. (2009) studied a different

perspective on public and private projects separately without breaking them down as roads, bridges, and rails. They found that 47% of the public sector and 37% of the private sector projects were completed within the budget from research in the Malaysian construction industry. Further, out of the remaining projects, 84.3% of the private projects faced less than 10% cost overruns. Nevertheless, that was reduced to 76% in public projects, which means public projects face more cost overruns than private projects. A different study carried out by (Nicolaisen et al., 2012) discovered that the MPCO for fixed links, bypass projects, motorways (new), and motorways (upgrade) would be 28%, 17%, 9%, and 2%, respectively, after studying hundred and forty-six transportation infrastructure projects in the UK.

Odeck (2019) conducted the most recent research on the MPCO in infrastructure projects through literature covering all the continents using 20,833 cases from road construction and 579 cases from rail and light rail. It concluded that road projects would have 26.9% of MPCO while rail projects would have 36.3% of MPCO. The researcher further emphasised that the percentage is getting lower over time, but it still needs to be solved. Conversely, (Flyvbjerg et al., 2002) studied and drafted the data of 111 projects against the time and concluded that there were no significant improvements over time. However, it was expected to be decreased with the proper technology, models, and experience. Although the MPCOs are varied from country to country, these studies emphasised that cost overrun is a global phenomenon that needs to be addressed (Creedy et al., 2010).

4.5 Research methodology

This research was conducted through a case study. Therefore, the first step of this study was to collect cost data from road projects in NZ. Therefore, a criterion was established to ensure the collected data sample is large enough to allow data analysis and identify the trends. The road projects' scope includes new construction repair and maintenance work, alterations to existing

roads, and trim work such as pavements, bus stops, and road markings and painting. However, the collected data were only for projects with the scope to construct new roads or large-scale alterations to existing roads.

All the project data were collected through the government agencies that manage road projects in NZ. The reason for approaching government bodies is that the government has to maintain transparency in funding construction projects. On the other hand, getting approvals and permissions to use the project data in the private sector is difficult. Furthermore, some projects will not release any data to maintain confidentiality and avoid getting into the hands of competitors. Therefore, collecting all the required data for the study from the private sector was impossible. Moreover, the study was conducted during the COVID-19 background. Hence, meeting or approaching the people to collect the data was complex.

However, data were collected from one hundred and six NZ road projects despite the above restrictions. As mentioned above, the projects were limited to roads, highways, and bridges associated with roads and highways. Other transportation infrastructure projects, such as rails and tunnels, were not considered as there were fewer projects in NZ; thus, it is hard to use those data to study trends or behaviours. The collected project data were categorised into three main categories based on the project size. The size of the project was decided based on the contract price.

· Small-scale projects – contract price less than 10 million (NZD) – 41 projects

· Medium-scale projects – contract price between 10 to 100 million NZD – 55 projects

· Large-scale projects – contract price exceeding 100 million NZD – 10 projects

There were a limited number of projects with contract prices exceeding 100 million NZD. Moreover, out of the available few, only ten projects were approachable to collect data. Therefore, the data analysis was conducted considering these limitations.

Once the data was collected, the normality of the data set was tested as the analysis test selection depended on the distribution type. There are two main tests for normality. The most commonly used test is the Kolmogorov-Smirnov test. However, this test could perform well with larger data sets (Grech and Calleja, 2018). Villasenor Alva and Estrada (2009) emphasised that the Shapiro-Wilk test could be used for smaller data sets. The researchers added that if the *p*-value is higher than 0.05, the data set is considered a normal distribution. If the data set is normally distributed, then the data analysis can be done using parametric tests such as the analysis of the variance (ANOVA) test (Grech and Calleja, 2018).

In a different view, Grech and Calleja (2018) emphasised that the normal Q-Q plot is a better normality test than the Shapiro-Wilk test. Q-Q plot compares the quantiles between the data distribution and the standardised theoretical distribution from a specified family of distributions (Augustin et al., 2012). It was further explained that the data points would fall on the 45-degree reference line if the data is normally distributed. Conversely, the points will deviate from the reference line if the data is not normally distributed. Based on the two contrasting views above, this study used the Shapiro-Wilk test and the Q-Q plot test to check the normality of the distribution.

In this analysis, the normal distribution was not completely satisfied as their distribution is skewed with an enormous upper tail. Therefore, the investigation should be carried out using non-parametric tests. The Kruskal-Wallis test is a non-parametric test similar to the ANOVA test (Grech and Calleja, 2018; Vargha and Delaney, 1998). Vargha and Delaney (1998) further explained that the Kruskal-Wallis test compares several populations. In this study, there were three

groups of data with different means. Therefore, the Kruskal-Wallis test aimed to test the null hypothesis that there is no effect on the magnitude of cost overrun by the project size or the time.

Although it was required to consider the start year as the time factor, the required information was unavailable in most projects. The reason is that the collected data were from projects within the past twenty years. So, most of the historical records were difficult to find. So, instead, the year of completion was used as the time factor to analyse the data. Then, histograms were created for all three groups separately and combined to show and analyse the distribution of the percentage cost overrun of each project. The following section summarises the key findings of the data analysis of this study.

4.6 Case study findings

Figure 4.1 shows the histograms of the cost overruns of one hundred and six NZ road projects. Histograms are categorised according to the project size, considering the project value. The small-scale projects' cost overruns were seemingly distributed symmetrically. However, the number of projects on the right side of the distribution is higher than on the left. Therefore, there is a slight skewness in the data. However, that will be tested through descriptive statistics before conducting other statistical tests. According to the histogram, the highest number of projects in small-scale and large-scale categories were within 10% to 20% of cost overruns. In medium-scale projects, the majority of the projects faced cost overruns within 0 to 10%.

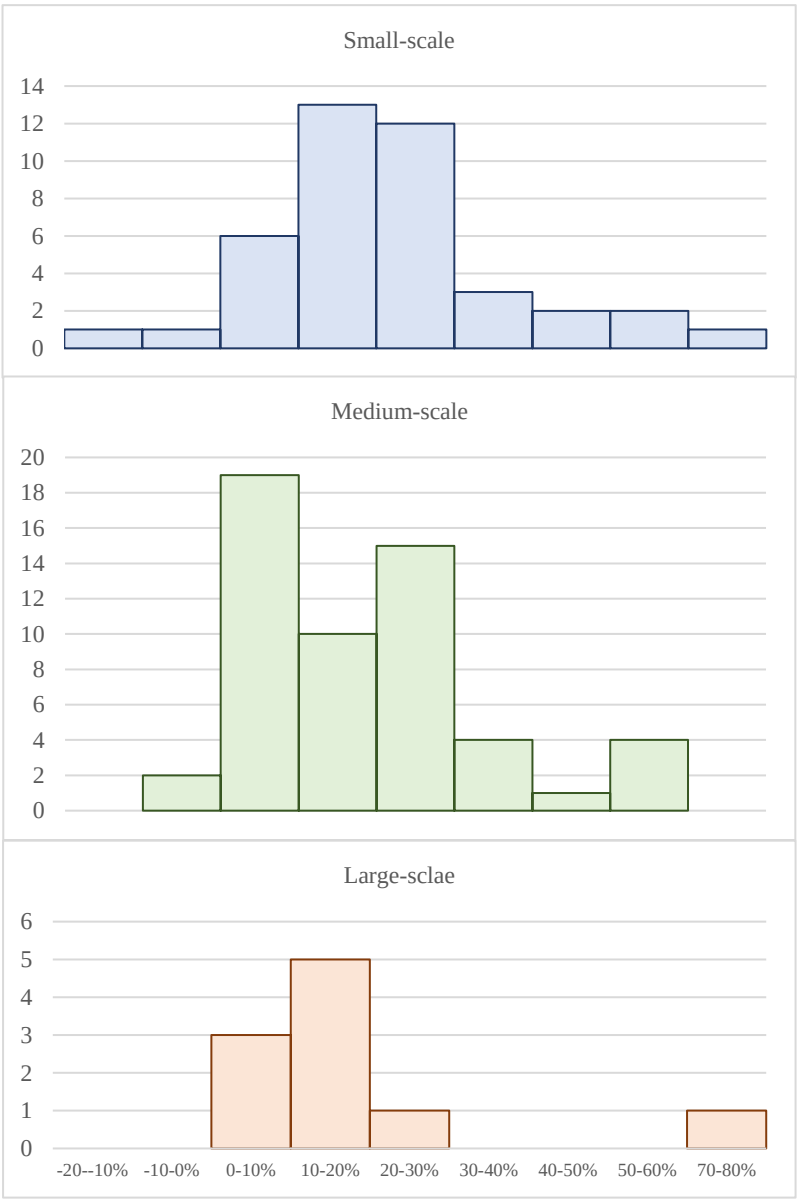

***Figure 4.1**. Cost overruns of 106 projects - histogram based on the project scale*

The cost overrun was calculated by taking the difference between the final account and the forecasted cost as a percentage of the forecasted cost. If the cost overrun percentage is zero, the project is completed within the forecasted budget. On the other hand, the histogram will be narrowly concentrated around zero if the difference between the final cost and the forecast is slight. In addition, when the cost overruns and cost underrun frequencies are the same, the histogram will be symmetrically distributed with zero skewness (Flyvbjerg et al., 2003). Conversely, according to Figure 4.1, the histogram is not symmetrical around zero and tends to grow more towards cost

overruns than cost underruns. According to this sample data, projects had more potential for cost overruns than underruns.

Further, the skewness indicates that a distribution leans towards the positive or negative side of the distribution. Through this value, it is possible to identify if the distribution contains an asymmetrical tail (Cain et al., 2017). Therefore, according to Table 3.1, small-scale projects show a negative skewness value, while the other two show positive skewness values. According to Cain et al. (2017), the negative skewness value indicates a tail on the left side, while positive skewness values indicate that the data set has a tail on the right side. Additionally, if the data set is normal, the kurtosis should be close to zero (Cain et al., 2017). However, according to Table 4.1, the Kurtosis values of all three groups are not close to zero. Nevertheless, the medium-scale projects show a minimum skewness close to zero. Therefore, a normality test is required before deciding the normality of the distribution.

First, the Shapiro-Wilk test was conducted to check the normality of the data set. As shown in Table 4.1, the p-values of all three groups separately and together were less than 0.05. Therefore, the data distribution was not normal. Subsequently, the above observation was further validated using the Q-Q plot, as shown in Figure 4.2. Although some data points were very close to the trend line, the slope of the trend line was not at 45 degrees. The data set significantly deviated from the trend line. Therefore, statistics confirmed that the data set is not a normal distribution. Accordingly, the data should be analysed using non-parametric tests.

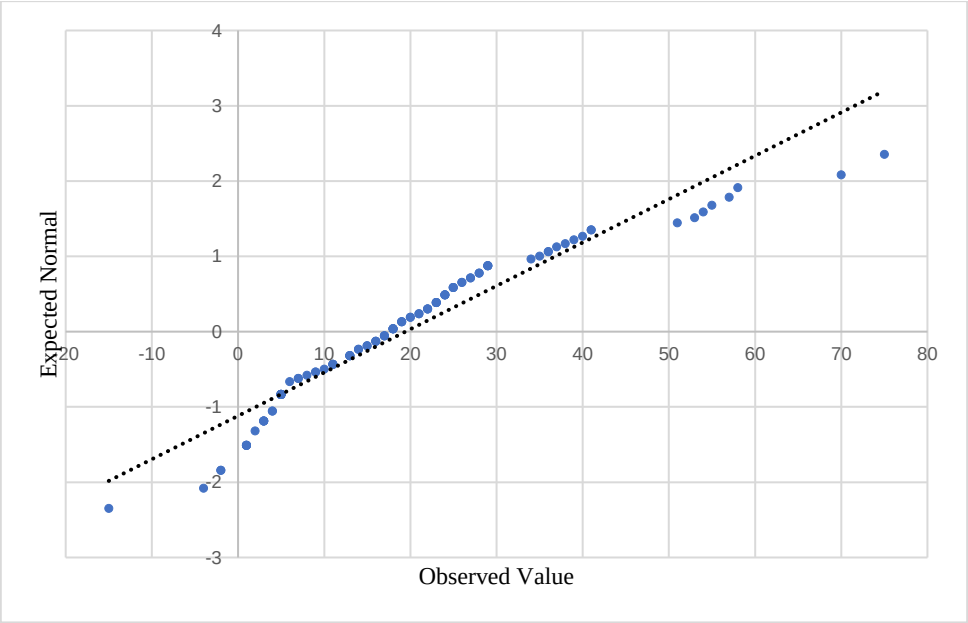

Figure 4.2. *Normal Q-Q Plot*

Table 4.1. *Descriptive statistics of the 106 road projects*

	Small (≤10 million NZD)	Medium (>10≤100 million NZD)	Large (<100 million NZD)	Total
Number of projects	41	55	10	106
Number of projects as a percentage	39%	52%	9%	100%
Shapiro-Wilk test	0.04	<0.001	<0.001	<0.001
Skewness	-0.127	0.09	2.419	
Kurtosis	0.588	-1.826	6.704	
MPCO (%)	21.6%	17.8%	19.6%	19.5%
Median	19%	15%	16%	18%
The sum of cost overruns (million NZD)	523.3	254.2	38.2	815.7
The sum of total project cost (million NZD)	240	2,104	2,733	5,077
The total cost overruns of all the projects as a % of the total cost of all projects	19%	14%	24%	20%

Legend: *NZD – New Zealand Dollars; MPCO – Mean Percentage Cost Overrun*

According to Table 4.1, small-scale projects faced an MPCO of 22%, the highest among all three groups. Medium-scale projects faced 18% of MPCO, while large-scale projects faced 20% of MPCO. In addition, once the total cost overruns of all one hundred and six projects were added together, it was approximately 815.7 million NZD, a significant amount compared to the NZ

economy. A major fraction of that came from small-scale projects totalling approximately 523.3 million NZD. Ten large-scale projects faced 38.2 million NZD worth of total cost overruns. Although the amount appears minor compared to the small and medium-scale projects, it must be addressed as the figure was calculated using only ten projects. It can be further validated by converting the total cost overruns into percentages. The total cost overruns of all ten large-scale projects as a percentage of the total cost of all ten projects were approximately 24%, which was the highest among all three groups. The same was true in small-scale projects, at 19%, while in medium-scale projects, it was 14%.

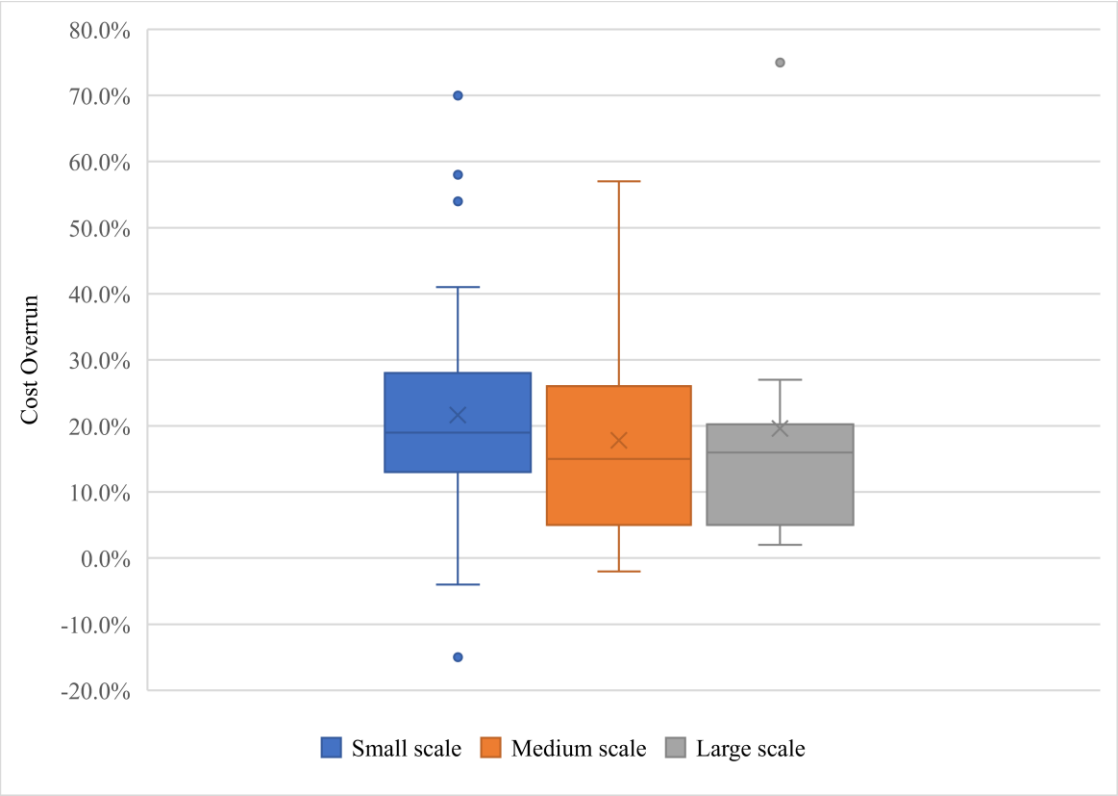

Figure 4.3. *Boxplot comparison of three project groups*

Based on the descriptive statistics, a boxplot was created to identify the error percentage ranges of all three groups. According to Figure 4.3, the means and medians of all three groups were nearly in the same range with slight deviations. However, the Kruskal-Wallis test will be conducted to validate the above observation. Hence, the Kruskal-Wallis test was conducted based on the null

hypothesis that there is no significant difference between the means of the three groups. If the null hypothesis is accepted, there is no effect or relationship with project size on MPCO.

The Kruskal-Wallis test revealed an impact of the project size on the magnitude of the cost overruns, $X2$ (2, N = 106) = 2.416, p = 0.029. Further, cost overruns were higher among small-scale projects (*Mean* = 22%) in comparison to medium-scale (*Mean* = 18%) and large-scale (*Mean* = 20%) projects. Therefore, the null hypothesis is rejected and establishes that the project size significantly affects the magnitude of the cost overrun.

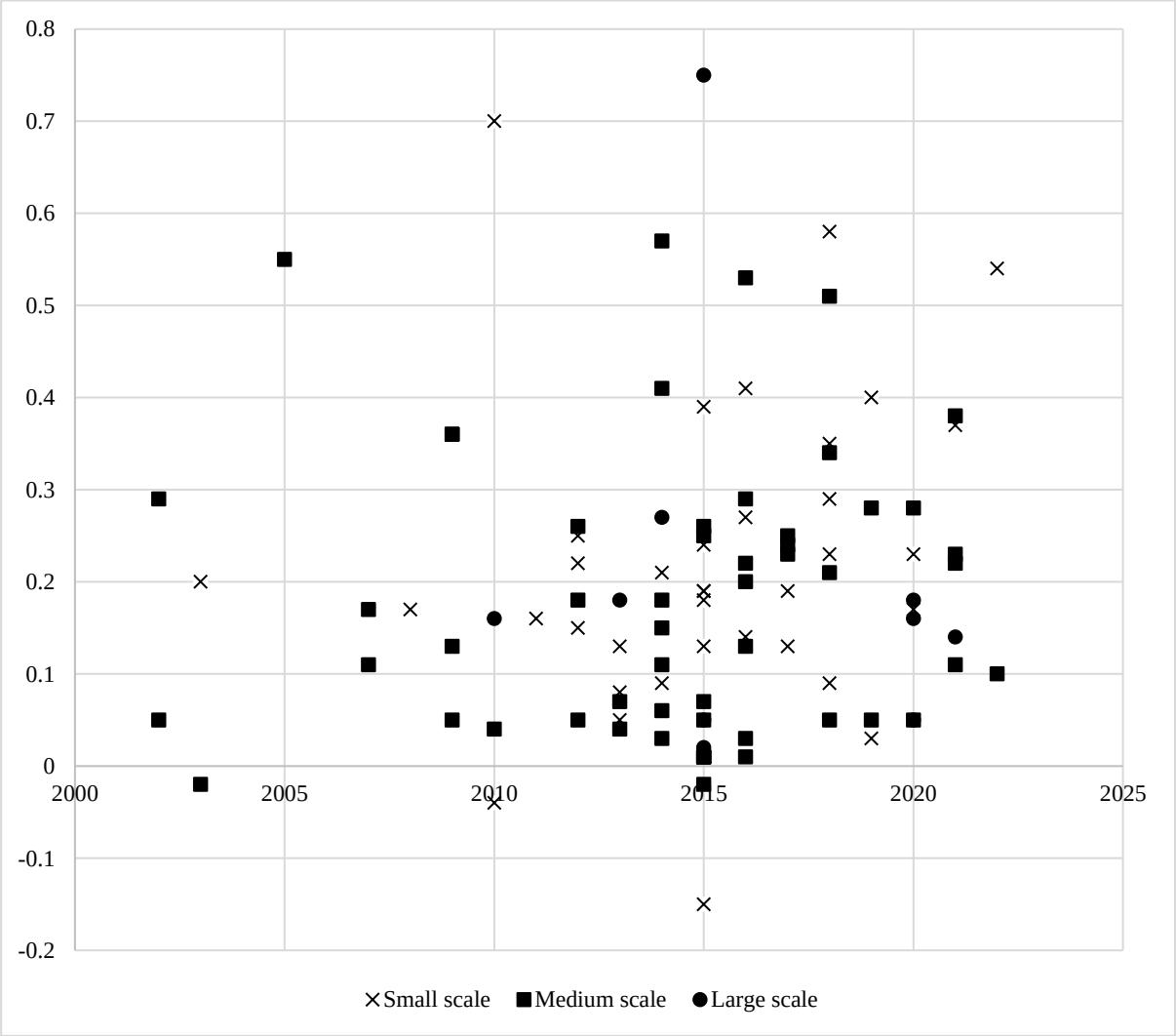

Figure 4.4. *Cost overrun of road projects over the past 20 years in NZ*

The second test was to check the relationship between the cost overruns and the time. Figure 4.4 is a scatter plot showing the spread of cost overrun percentages against the year of completion for

one hundred and six project data. According to the diagram, no trends or patterns are visible in all three types of projects. Even if all the projects are considered one group without separating them according to the project value, no particular pattern is visible. Consequently, the draft denotes no significant impact on the project cost overrun by the year of completion. However, the observation has to be validated through the Kruskal-Wallis test before conclusions can be established. Hence, the Kruskal-Wallis test was conducted based on the null hypothesis that there is no effect on the cost overrun by the time represented by the year of completion. However, the Kruskal-Wallis test showed no effect on the magnitude of cost overruns by the time or the year of completion, $X2$ (18, $N = 106$) = 17.93, $p = 0.46$. Therefore, the null hypothesis is accepted and concludes that the cost overrun percentage of NZ road projects does not depend on the time when the project is completed.

Since the study confirmed no connection between the cost overruns and the project execution time, this study carried out another test to evaluate the connection between the time and the cost overruns. The measure used for this test was the project duration. All the project durations of the one hundred and six cases were tested using the Kruskal-Wallis test. According to the Kruskal-Wallis test, the project's duration significantly contributes to the magnitude of the cost overrun, $X2$ (105, $N = 106$) = 12.82, $p = 0.04$. Therefore, the null hypothesis is rejected and established that the cost overrun magnitude of the NZ road projects is significantly impacted by the project duration.

4.7 Discussion

The study focussed on identifying the significance of cost overruns in NZ road projects. The paper presented the results of an investigation using one hundred and six NZ road projects under three subcategories based on the project size. The project size was differentiated based on the contract price. The total value of the selected road projects was approximately 1.5 billion NZD. Road projects would tend to perform differently than planned because none of the projects considered

in the data analysis faced zero cost overruns. In other words, finishing a road project precisely on the budget is rare and extremely difficult. Therefore, the project cost is highly uncertain based on cost, time, and quality constraints.

The following observations can be made based on the above research findings (Table 2 and Figure 4). Only 3.8% (≈4%) of the road projects in NZ finished on or below the forecasted budget, while the rest of the 96.2% (≈96%) projects faced cost overruns. That means 96 projects out of 100 have a higher chance of spending more than the planned budget for the project. According to the analysis, the general cost overrun percentage of road projects in NZ is 19.5%, with a standard deviation of 16.07%.

According to the data analysis, 96% of road projects faced positive cost overruns, while 4% faced negative cost overruns or savings. Flyvbjerg et al. (2003) conducted a similar study based on 258 TI projects across 20 nations. Their analysis considered all TI project types, including roads, rails, and fixed links such as bridges and tunnels. According to their study, 9 out of 10 projects (90%) generally faced cost overruns in any TI project. Although the results were similar, our research identified that NZ road projects face even more severe issues, as the percentage of projects that face cost overrun had increased from 90% to 96%.

On the other hand, Love et al. (2016) and Odeck (2004) established that a project's cost overrun could be as high as 70% and 183%, respectively, more than the estimated budget. In our study, the maximum cost overrun was noted as 75%. Therefore, this research study is in the same phase as the findings of Love *et al*. (2016).

Furthermore, our study revealed that small-scale projects with contract prices of less than 10 million NZD faced 22% of MPCO, while medium-scale projects with contract prices between 10 and 100 million NZD faced 18% of MPCO. In addition, large-scale projects with contract prices exceeding 100 million NZD experienced 20% of MPCO. Therefore, considering all the above

project categories, the overall MPCO was approximately 20%. Similarly, Flyvbjerg et al. (2002) found that road projects would face an MPCO of 20%, which matched the results of our study. On the other hand, Odeck's (2019) investigation showed an MPCO of 26% for road projects, which was slightly higher than the results of our study. Similarly, Creedy (2010) and Lee (2008) also confirmed relatively closer results of 28.8% and 29.5%, respectively. In a different view, Odeck (2019) also carried out a literature survey based on 48 publications to examine the MPCO results and found that the MPCO could be as high as 214% (from the research conducted by Al-Hazim and Salem (2015)). However, the average of all the MPCOs from 48 publication results was 34.12%. That was approximately 14% higher than the results of our study.

Even though the projects were categorised into three sub-groups, the mean cost overruns of all three groups were almost similar. Small, medium, and large-scale projects would face 22%, 18%, and 20% cost overruns. It was found that the project size impacted the cost overrun percentage. Accordingly, the more significant projects tend to experience lower cost overruns, while the smaller projects may experience considerably higher cost overruns. Similar findings were observed in the study carried out by Huo et al. (2018).

In contrast, a similar investigation was conducted by Flyvbjerg et al. (2003) and observed that the project size did not impact the cost overrun magnitude across all the TI projects. However, if the project is large-scale, the risk and uncertainties are comparatively higher than in small-scale projects. Nevertheless, large-scale projects have better planning, management, and monitoring systems than small projects. Therefore, the cost overruns in large-scale projects can be mitigated. Further, the project is more straightforward when it requires proper initial planning and management processes. Therefore, the findings of our research are more sensible.

Furthermore, the study investigated the impact of cost overrun magnitude on time represented by the year of completion of the project. According to the findings, it was clear that the time or the

year of completion does not substantially impact the cost overrun percentage. However, the study further investigated the connection between the cost overrun and the time by considering the project duration. Findings showed that the scope and definition are inadequate if the project duration is longer, and the uncertainty level is higher than small-duration projects. Therefore, planning lengthier road projects requires comprehensive attention compared to shorter projects.

Further, the mean cost overrun of each year varied due to external impacts during the time. For instance, according to the analysis, the MPCE in 2020 is considerably low. However, after that, with COVID-19, MPCO started to go up. Currently, the MPCO is showing an increasing trend. Therefore, the current estimation practice in the NZ road projects has not improved significantly over the past 20 years. The researchers supported the above observation based on studies over 70 years and concluded that there was no improvement in cost overrun percentage over time in all the TI project types (Flyvbjerg et al., 2003). Therefore, the cost overruns issue is significant in NZ road projects. The literature states that NZ has planned several major TI projects in the coming decade. The five projects mentioned in the literature were budgeted for 31.3 billion NZD. According to the investigation, MPCO faced by large-scale NZ road projects were around 20%. Therefore, the above projects may face approximately six billion NZD cost overruns. Considering the background of the NZ economy, that is a significant amount. The following section concludes the findings of this research.

4.8 Conclusions

Generally, TI projects incur enormous sums of money over a considerably extended period compared to most building projects. In addition, it has been common for TI projects to face significant cost overruns. Although much research has been done to identify the causes of cost overruns, proper action has yet to be taken to minimise the effect of cost overruns. Therefore, this research aimed to identify the significance of cost overruns in NZ road projects so that researchers

and industry practitioners can understand how severe the problem is and investigate further to find solutions. TI projects generally include roads, bridges, tunnels, and railways. However, due to the unavailability of enough data to carry out an in-depth analysis, only the road projects and bridges associated with roads were considered in this study.

It was concluded that cost overrun is an inevitable issue NZ road projects face. Moreover, over the past twenty years, there has been no improvement in cost performance. Hence, the current estimation practice must be upgraded to address this issue. 96% of the NZ road projects faced significant cost overruns, with an MPCO of 20% overall. Moreover, it was found that there is no substantial effect on the cost overruns from the year of completion. However, the NZ road projects demonstrated a significant impact on the project cost overrun magnitude by the project size and the project duration. The project size was considered based on the contract price of the project, while the time was studied based on the year of project completion.

The study was conducted based on the cost data collected from NZ road projects over the past twenty years. Therefore, the results can differ for the projects finished before the considered timeline or the projects in other countries and for other project types. However, the study concluded that NZ road projects faced significant cost overruns, and this statement is consistent with the literature findings. Therefore, the study's results can be used as background evidence to investigate further to find a solution to minimise the cost overruns.

Further, the future pipeline projects of the NZ road construction sector were planned to consume multi-billions, as identified in the introduction section. Therefore, with several upcoming road projects of large scale and longer duration, finding a solution to overcome the cost overruns is essential. Therefore, the client knows the entire budgetary plan from the conceptual stage.

Although the case study was carried out using NZ road projects, the findings on project size and duration can also be generalised to other contexts. The study recommends further analysing the

cases to identify potential solutions to overcome the cost overruns. Thus, the author explores the possibility of addressing the conceptual estimation errors by incorporating proper modelling techniques into the current estimation practice.

4.9 Epilogue

According to the findings, discussion, and conclusions of this chapter, it was resolved that the project size (determined by the project value) and the project time (determined by the project's start date) would not significantly impact the project cost overruns. The phenomenon of cost overruns in road projects is a complex and multifaceted problem that has garnered the attention of researchers, practitioners, and policymakers. This issue is not limited to a specific region or type of project. However, it has been observed worldwide across various project scales, from local road improvements to large-scale highway construction. The consequences of cost overruns are profound and extend beyond financial implications, affecting project stakeholders, the general public and the economy. Therefore, the next chapter investigates the modelling techniques to address the current conceptual estimation practice issues. If a modelling technique can produce a more reliable and realistic estimate with minimum cost overrun, that will benefit the project, project team and the general public.

5 Cost modelling techniques for conceptual cost estimation of infrastructure projects – a literature review

5.1 Prologue

Chapter 3 investigated the factors affecting the project's cost as the first step of finding a solution to the cost overrun matter. Further, in Chapter 4, it was identified that cost overrun is a significant and continuous ongoing issue in road projects in NZ. Based on the background set in these two previous chapters, this current chapter examines the literature on the possible statistical techniques that can be used to develop cost estimation models. Subsequently, the chapter investigates further the cost modelling techniques for infrastructure projects. Finally, recommendations are provided for reliable cost-modelling techniques for road projects.[3].

[3] This chapter is based on the following journal paper.

Atapattu, C.N., Domingo, N., and Sutrisna, M. (2023). A bibliometric review of the statistical modelling techniques for cost estimation of infrastructure projects. *Smart and Sustainable Built Environment, (*Ahead of Print). DOI: https://doi.org/10.1108/SASBE-01-2023-0005.

Atapattu, C. N., Domingo, N. D. and Sutrisna, M. (2022). Statistical cost modelling for preliminary stage cost estimation of infrastructure projects. IOP Conference Series: Earth and Environmental Science, Vol. 1101, pp. 052031. DOI: https://doi.org/10.1088/1755-1315/1101/5/052031

5.2 Abstract

Cost overrun in infrastructure projects is a constant concern with a need for a proper solution. The current estimation practice needs improvement to reduce cost overruns. This study aimed to find possible statistical modelling techniques that could be used to develop cost models to produce more reliable cost estimates. A bibliographic literature review was conducted using a two-stage selection method to compile the relevant publications from Scopus. Then, VOS-Viewer was used to develop the visualisation maps for co-occurrence keyword analysis and yearly trends in research topics. The study found seven primary techniques used as cost models in construction projects: regression analysis (RA), artificial neural network (ANN), case-based reasoning (CBR), fuzzy logic, Monte Carlo simulation, support vector machine, and reference class forecasting. RA, ANN and CBR were the most researched techniques. Furthermore, it was observed that the model's performance could be improved by combining two or more techniques into one model. This study mapped the research carried out on cost-modelling techniques and analysed the trends. It also reviewed the performance of the models developed for infrastructure projects. The findings could be used to research further to develop more reliable cost models using statistical modelling techniques with better performances. The research was limited to the findings from the bibliometric literature review. The findings provided an assessment of statistical techniques that the industry can adopt to improve the traditional estimation practice of infrastructure projects.

5.3 Introduction

Cost overrun is considered a global phenomenon in the construction industry, especially in long-run mega-scale projects such as infrastructure projects. For example, Huo *et al*. (2018) researched infrastructure projects in Hong Kong. They identified a mean percentage of cost overrun of 32.52% in road projects, 34.83% in rail projects, and 37.48% in bridge and tunnel projects. A proper cost

forecasting model is strongly needed because infrastructure projects usually experience extended construction periods.

On the other hand, the project management survey done by KPMG New Zealand observed that the number of projects delivered within budget decreased from 48% to 29% from 2010 to 2017 (Barlow *et al.,* 2017). Hence, the budget's reliability and accuracy are decreasing over time. The project's success will also be affected if the budget is inaccurate. Therefore, cost estimation at the pre-contract stage plays a significant role. However, Challal and Tkiouat (2012) explained that, compared to other industries, construction projects face difficulty anticipating the characteristics and the work to be done in the future, considering the long-term project delivery. Therefore, achieving a reliable and accurate estimate for the project is not straightforward. Furthermore, they emphasised the importance of cost estimation to project development.

Researchers also found two types of cost overruns: avoidable and unavoidable. The avoidable costs are the matters that could have been foreseen by the management beforehand. Conversely, unavoidable costs are caused by events that cannot be foreseen beforehand (Shanmugam *et al*., 2006). According to Shehu *et al*. (2014), construction projects must deal with various issues, mainly the waste of funds and investments, the loss of the expected end-product quality of the outcome, and several other issues arising from project delays. Budget estimation at the preliminary stage is commonly developed using historical data. Various traditionally used estimation methods are available at the preliminary stage, where information availability is minor and uncertainty is high (Challal and Tkiouat, 2012). However, the reliability of these traditional models is often questioned because proper judgment is critical in estimation. Hence, the estimation methods should be able to develop robust estimates that address the uncertain factors. However, the traditional preliminary estimation methods do not support that.

In a different view, Skitmore and Marston (2005) emphasised that historical trend-based data and judgment based on construction knowledge and experience are vital components of a proper cost forecasting technique. Thus, in the past few years, researchers have focused on developing new cost models that can incorporate both these components while minimising the errors and biases of the judgments resulting from the human mind's cognitive behaviour. Hence, researchers adapt various statistical modelling techniques into cost estimation models.

This study aimed to identify the significant statistical modelling techniques that can be used for the cost estimation of infrastructure projects. The following section discusses the research methodology of this research.

5.4 Research Methodology

5.4.1 Bibliometric literature review

The methodology was comprised of three major stages. First, the study focussed on identifying the primary statistical techniques in the cost estimation of construction projects. Second, the study further evaluated the models developed using the identified techniques but within the selected literature in the first step. Then, the models developed for infrastructure projects were filtered out for further review. Finally, it was decided to analyse the keywords related to statistical modelling used for cost estimation in construction projects. Therefore, the bibliometric literature review method was implemented because many researchers stressed that bibliometric analysis is much more suitable for identifying emerging trends in research areas, collaboration patterns, and performances of articles and journals (Donthu et al., 2021; Hallinger and Kovačević, 2019; Merigo et al., 2017; Zhao et al., 2019).

Moreover, Donthu et al. (2021) suggested that the bibliometric review should be used when the scope is broad, while the systematic literature review is suitable when the scope is specific. Therefore, the study chose the bibliometric review over the systematic literature review.

Merigo et al. (2017) stated that bibliometric review studies the bibliographic materials quantitatively by analysing and identifying the trends related to a particular research theme. According to Hallinger and Kovačević (2019), bibliometric analysis provides more comprehensive and deeper analyses with the advancement in text mining and citation analysis tools. Therefore, the researchers adopting this method for literature review has become a growing trend.

In addition, Falagas et al. (2007) stated that Scopus covers more recent publications than other digital sources, such as the Web of Science. On the other hand, Zhao et al. (2019) emphasised that there are no significant differences between the bibliometric analysis results of Scopus and Web of Science. Therefore, the study used the Scopus database to collect 147 articles for the analysis. Figure 5.1 illustrates the publication selection process and the criterion used for the bibliometric literature review. As illustrated in Figure 5.1, the keywords were entered into the Scopus Database without any restrictions on time.

Consequently, the search contains records that expand from 1997 to 2021. The purpose was to study the development of the research over the years. The search identified 219 records, out of which nine records were removed as they were not in English. In the next stage, another 41 records were removed and 169 were selected for construction. Then, only the journal and conference papers were retained, and 12 records were removed. The remaining 157 articles went through a screening of titles, abstracts, and keywords. Finally, 147 articles were retained for the bibliometric review.

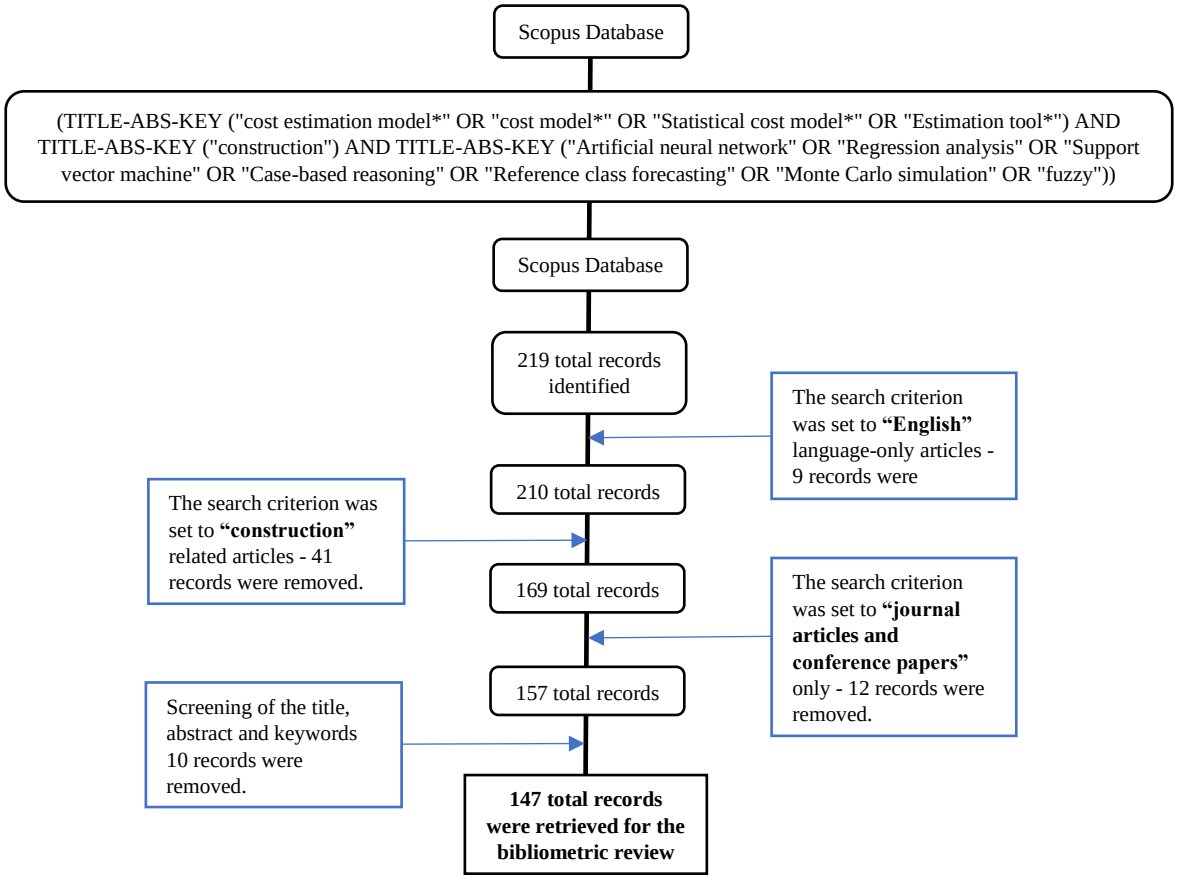

Figure 5.1. *Criteria used to retrieve the documents for bibliometric review*

In addition, Table 5.1 shows the publication source analysis for the selected articles. The table contains the publication sources from highest to lowest based on the number of citations received for the selected articles. The sources with an insignificant number of citations were not considered here.

Table 5.1. *Most cited publication sources*

No.	Source	Citations	Documents
1	Construction Management and Economics.	457	9
2	Building and Environment	360	3
3	Construction Engineering and Management	355	12
4	Automation in construction	105	3
5	Computer-Aided Civil and Infrastructure Engineering	95	2
6	Engineering, Construction and Architectural Management	82	6
7	Canadian Journal of Civil Engineering	74	4
8	Financial Management of Property and Construction	73	2
9	Management in Engineering	52	3
10	Construction Management	14	2

5.4.2 Visualisation of data analysis

Usually, the network visualisation software provides both options of bibliometric analyses and the graphic visualisation function (Donthu et al., 2021). Therefore, VOS-Viewer was used to conduct the analyses, map, and display the findings. Van Eck and Waltman (2014) emphasised that the VOS-Viewer is a comprehensive tool based on the Visualisation of Similarities (VOS) technology. They added that it could cluster the fragmented knowledge from different domains according to their similarity and relatedness. Once the clusters are identified, VOS-Viewer will map the similarities to visualise the links and connections. A node will denote the keywords, and the node size of the map signifies the counting of the appraised item, such as citation and occurrence. Similarly, a link represents the co-citation and co-occurrence (Van Eck and Waltman, 2014).

Therefore, VOS-Viewer can be used for several analyses, such as keyword co-occurrence, country-specific, co-authorship, and yearly trends in research topic analyses (Randhawa et al., 2016). However, this study was focused on keyword analysis and yearly trend analysis as the aim was to identify the statistical modelling techniques and their growth potential in research.

5.5 Research findings

5.5.1 Co-occurrence keyword analysis

According to Randhawa et al. (2016), keywords provide an understanding of the research concept of the article by being points of reference. The traditional cost estimation practice must be aligned with the development of design, technology, and construction methodologies. Therefore, more research has focused on new cost modelling techniques in recent years. As a result, there were seven basic cost modelling techniques identified through literature: regression analysis (RA), artificial neural network (ANN), Monte Carlo simulation (MCS), support vector machine (SVM), case-based reasoning (CBR), reference class forecasting (RCF), and fuzzy logic (Shabniya and

Dilruba, 2017). While some researchers developed models based on these basic techniques, it was observed that some studies were performed by combining these techniques or adopting the basics while developing custom-made models. Therefore, it was not certain what techniques were the most considered in cost modelling. In that case, keyword occurrence analysis would provide a fundamental analysis of the importance of the seven techniques mentioned above in cost modelling for construction projects.

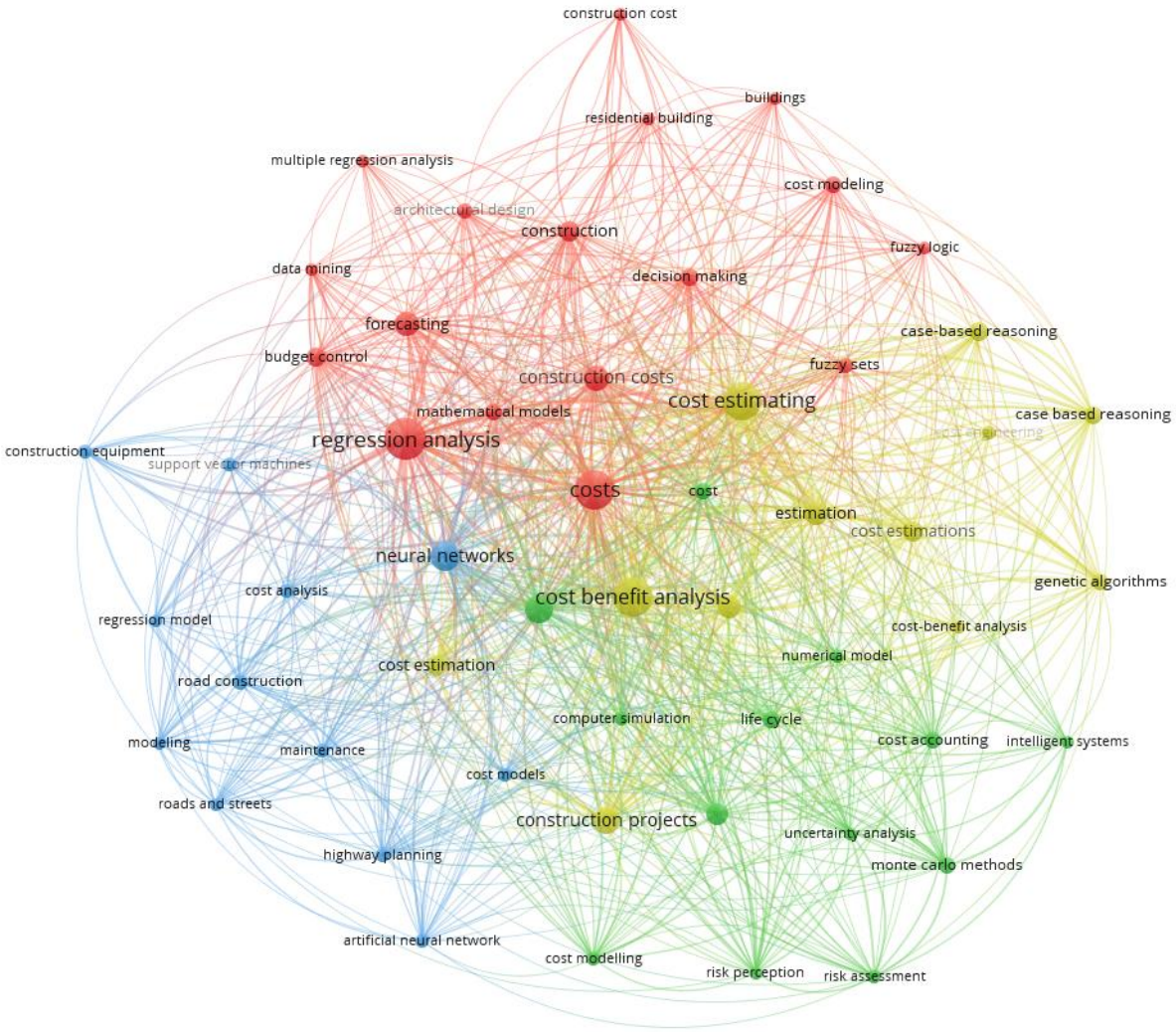

Figure 5.2. *Cluster visualisation map for co-occurring keywords (Source: VOC-Viewer)*

Once the publication record was uploaded to VOS-Viewer, keyword analysis was carried out through VOS-Viewer using the options of "all keywords" and "full counting". The minimum occurrence of a keyword was set at 4. Out of 1,289 keywords, 54 met the criteria. The proximity of keywords determines the degree of co-occurrence (Van Eck and Waltman, 2014). On that basis,

the identified keywords showed 891 links with 2560 total link strength. Four clusters were identified in the map as cluster 1 (17 keywords), cluster 2 (13 keywords), cluster 3 (12 keywords), and cluster 4 (12 keywords). Figure 2 illustrates the cluster visualisation for research keywords and their co-relationship based on the cost estimation modelling in construction projects. According to Figure 5.2, the primary cost modelling technique identified was RA.

RA and fuzzy sets/ logics were the techniques identified in Cluster 1. That means there could be co-occurrence and relationships between these two techniques. Otherwise, researchers might tend to compare models with the two techniques. Cluster 2 identified the MCS modelling. The rest of the keywords in this cluster were mainly related to risk. In addition, cluster 3 dominated all clusters with three effective techniques: RA, ANN, and SVM. It was noted in the literature search that researchers often compare the three techniques in their models. Finally, the third cluster recognised CBR as a cost modelling technique.

Furthermore, Figure 5.2 also highlights the primary modelling techniques identified with their respective number of occurrences and the total links. RA achieved the highest occurrences, while ANN and CBR were ranked second and third, respectively. According to the analysis, reference class forecasting was not identified in the keyword occurrence because the number of occurrences of the word was below the set criteria.

5.5.2 Yearly trends on research topic analysis

The same network visualisation map was converted into a sensor map that analyses the yearly trend from 1997 to 2021 in researching the areas related to the keywords. The publications in 2022 were not considered since the study was conducted in mid-2022, and the graph would show a sudden downfall in growth. Therefore, it could suggest wrong conclusions. Figure 5.3 illustrates the visualisation map generated by VOS-Viewer, analysing the yearly trends in research. It was observed that recent research was more focused on RA, ANN and SVM. In addition, Figure 5.4

shows the growth of research related to cost modelling techniques in the construction sector. There is an increasing demand for research in this area.

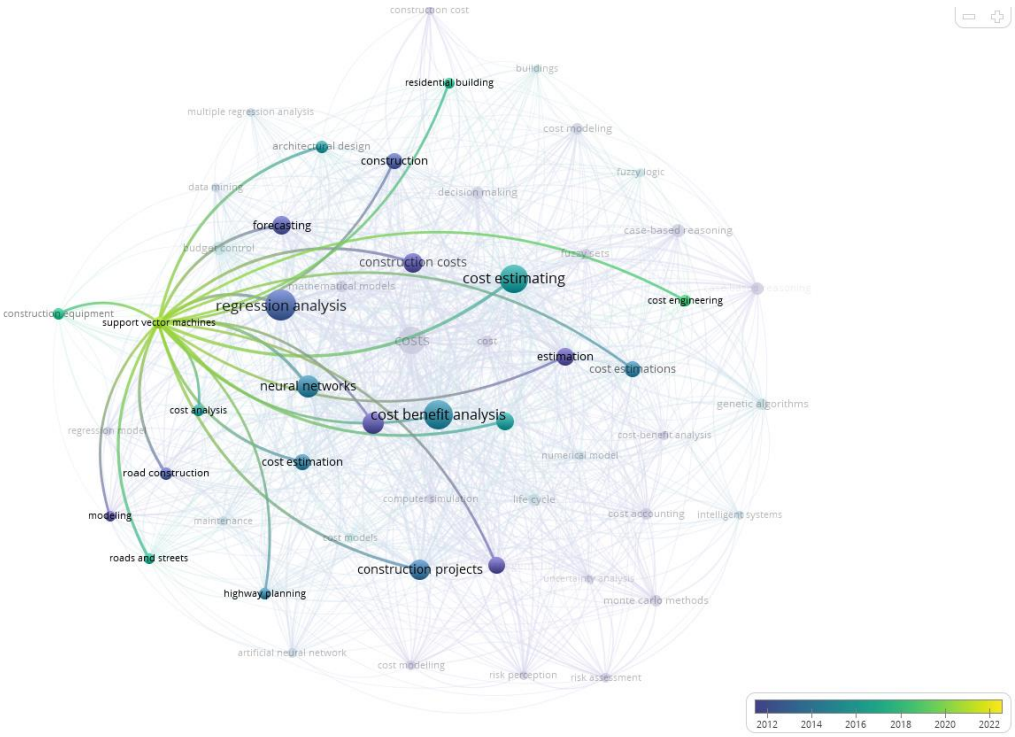

Figure 5.3. *Yearly trends on research topics (Source: VOC-Viewer)*

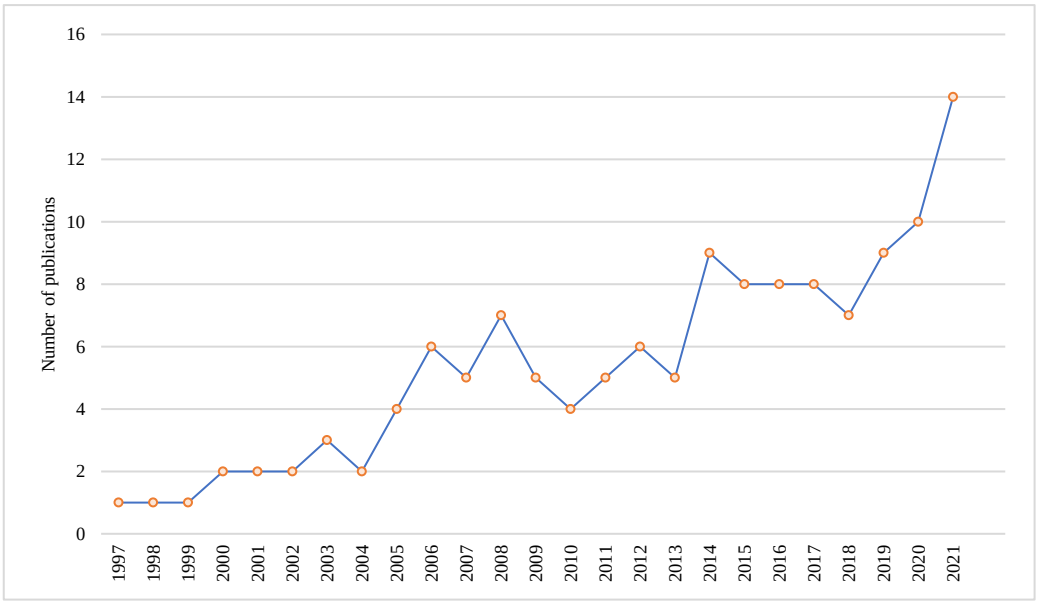

Figure 5.4. *Annual publication distribution*

5.5.3 Country-wise publication distribution

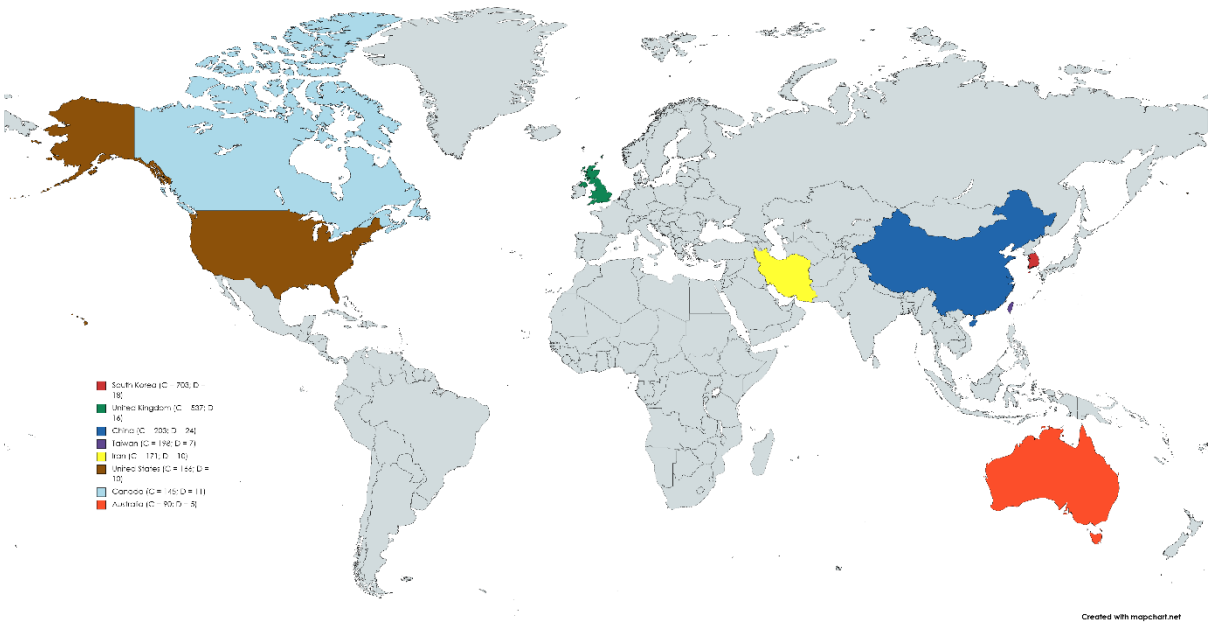

Figure 5.5. *Country-wise publication distribution (created with www.mapchart.net)*

Legend for Figure 5.6: C – Number of Citations, D – Number of Documents

Figure 5.5 shows the publication distribution based on the country. The corresponding author's affiliation was considered for the analysis. First, citations and the number of documents were obtained through the VOS-Viewer analytical function set for a minimum of five publications per country. Then, the map was generated through www.mapchart.net. Only the countries with citations on or above 90 were considered for the analysis.

According to Figure 5.5, South Korea published only 18 articles but received the highest number of citations. Meanwhile, China had the highest number of publications, with 24 articles. Nevertheless, based on the number of citations, China was in third place. In addition, the United Kingdom, Taiwan, Iran, the United States, Canada, and Australia have significantly contributed to the construction industry's research in cost modelling techniques. The following section will discuss the findings of the literature search.

5.6 Discussion

According to the findings, RA and ANN were the most significant statistical modelling techniques considered in cost estimation in construction projects. Furthermore, there was a strong correlation between these two techniques. It was observed that most of the research concentrated on both techniques to compare the models' performances. CBR ranked third, while the fourth, fifth and sixth were fuzzy logic, MCS, and SVM, respectively. In cluster 2, all the keywords were related to construction risks, and the technique included in this cluster was the MCS. Cluster 3 comprised three effective techniques mentioned in the findings: RA, ANN and SVM. Nevertheless, the rest of the keywords in this cluster were mainly related to road construction. Therefore, the cost estimation models developed for road projects were mainly based on the above three techniques.

According to Figure 5.4, the research trend on the statistical modelling techniques in the construction industry showed significant growth. The trend analysis emphasised that with the current development in design, technology, and construction methods, researchers tend to search for better and more reliable techniques to improve traditional estimation practices. The following sections will further discuss the findings of each technique based on the literature review.

5.6.1 Regression analysis (RA)

The bibliometric review emphasised that the RA was the most cited technique. It is a technique that estimates the relationship between the parameters. RA considers the relationship between each dependent variable against one or more independent variables. In other words, the model includes identifying the impact on a dependent parameter when an independent parameter is varied and all other independent parameters are fixed (Shabniya and Dilruba, 2017). In contrast, a large and growing body of literature has investigated the drawbacks of RA and emphasised that the technique needs a more straightforward approach to comparing the new project with historical data

to identify the most suitable model for the estimation (Kim et al., 2004). Hence, certain assumptions related to the equations would have to be made to make them suitable for the regression equations.

Additionally, surveys such as those conducted by Vahdani *et al*. (2012) emphasised that the regression technique is based on a linear nature while the cost data of construction projects are generally non-linear. However, the researchers highlighted that the technique comprises analytical and predictive capabilities. In contrast, Ahn *et al*. (2017) argued that the relationship between project parameters could be either linear or non-linear.

Using the regression model concept, Herbsman (1986) developed a model to forecast the costs of highway projects using cost indices and historical cost data. However, it was stressed that a cost forecasting model should be based on price escalation and other influencing factors. El-Maaty *et al*. (2017) developed a linear regression-based cost model to predict the percentage of cost overruns using 46 highway projects. The model showed an average percentage error of 30.42%. Their research also included a statistical fuzzy-based model using the same data set. The conclusion was that the regression-based model outperformed the fuzzy-based model. In another significant study, Hammad *et al*. (2008) highlighted the need for a custom model based on project type rather than a generic formula. Furthermore, the researchers developed a regression model with a 95% probability of accuracy. The study used 140 various building projects in Jordan.

Another investigation was carried out by drawing on the concept of RA; Lowe *et al*. (2006) developed six cost models that differ based on the number of variables used in each model. The models were based on cost data from 286 building projects in the United Kingdom. Out of the six, the best model had an error of 19.3%. The researchers compared the model against ANN using the same data and variables to validate the model. The best outcome of the ANN utilised 41 variables and showed a 16.6% error. Therefore, it can be emphasised that ANN is more effective than RA.

Günaydın and Doğan (2004) emphasised that RA performs poorly with early-stage design parameters.

Kim *et al*. (2013) and Wang *et al*. (2012) compared ANN, regression, and SVM. They emphasised that the ANN models produced better output results than the regression models. However, Kim *et al*. (2013) achieved a significantly lower MAER of 5.68% for their regression model. Unlike the other models, Ganiyu and Zubairu (2010) focused on variables significant to cost, with numerical relationships that are challenging to identify. For example, the complexity level, the project's importance, on-time completion, and construction complexity level were numerically comprised in the model. However, the model displayed an error of 23.8%.

According to the above findings, RA has mixed performances. There were both good and poor models, and the differences could be due to the cost database or the variable selection.

5.6.2 Artificial Neural Network (ANN)

Based on the findings, ANN was recognised as the second technique cited in construction cost estimation research. Similar to RA, there has been a growing trend in recent years on this topic. However, according to Ahiaga-Dagbui *et al*. (2013) and El-Kholy (2019), ANN is one of the most sophisticated models capable of non-linear approximation and modelling complex functions. Nevertheless, much research has stressed that most modelling techniques, such as regression, are incapable of non-linear approximation (Adel *et al*., 2016; Cheng and Hoang, 2014). Therefore, RA is not a good modelling technique for construction project estimation compared to ANN.

It was also identified that ANN could use the lessons learned from previously completed cases and generalise the knowledge to future projects (Adel *et al*., 2016; Cheng and Hoang, 2014; El-Kholy, 2019). On the other hand, Cheng and Hoang (2014) stated that the training process of ANN is overly complex as it is attained through a gradient descent algorithm on the error space. However,

they also found that the genetic algorithm and Particle Swarm Optimisation (PSO) can be used to overcome this issue.

Tarek *et al*. (1998) developed a cost model which adopted the ANN concept using 18 highway projects in Canada. That model predicted the cost with a weighted error of 0.98. Another study was carried out by Adel *et al*. (2016), who developed a model based on ANN using the data from 75 highway projects in Egypt. The study carried out several training, evaluation, and validation tests to measure the model's performance, and the results were found to be 4.51%, 5.8%, and 16.0%, respectively.

Meanwhile, El-Kholy (2019) identified fifteen variables to develop two models for predicting the time and cost overrun based on the ANN technique through the "NeuroSolutions (6.12) (2012)" software. The model was based on data from 56 highway projects in Egypt. The average error was concluded to be 39.8%. Therefore, El-Kholy's (2019) model was comparatively more accurate than the model developed by El-Maaty *et al*. (2017) using the same data set based on linear regression and a fuzzy statistical model.

Kim *et al*. (2004) took a different approach to neural networks. They combined genetic algorithms and neural networks to optimise the model by selecting the number of neurons in the hidden layer. In this study, several models adjusted the number of layers and neurons. The best model consisted of 2 hidden layers. The number of neurons was 12 in the input layer, 18 in the first hidden layer, 12 in the second hidden layer, and 1 in the output layer. This model achieved an MAER of 2.62%. The study concluded that the genetic algorithm is better for selecting parameters than the trial-and-error method.

As discussed previously, the primary study by Kim *et al*. (2013), comparing three modelling techniques, concluded that the ANN was better for cost forecasting than RA and SVM because it can deal with multifaceted problems while producing user-friendly models. Based on ANN, the

model developed in the study by Kim et al. (2013) they achieved an MAER of 5.27%. Similarly, Sodikov (2005) compared ANN with RA by developing two models with the same data set from Poland and Thailand. As expected, the ANN-based model exhibited a lower estimation error rate than the regression-based model. However, the model developed by Günaydın and Doğan (2004) accomplished the lowest error of 3.8%. This model was developed for the early-stage cost estimation of apartment buildings; however, this model was focused only on design and site-related variables. Therefore, the contract-related parameters, such as tendering strategy, contract type, procurement methods, and the like, were not considered.

Therefore, the findings suggested that ANN is an effective modelling technique for cost estimation in construction projects.

5.6.3 Case-based reasoning (CBR)

CBR is the process of solving a new case or a project based on previously resolved cases using the knowledge base of the model. Therefore, this model compares the new project with the old data set and finds a similar case (Agnieszka and Krzysztof, 2018; Tah *et al*., 1998; Zima, 2015). Several studies investigating this technique described four steps of the CBR model: retrieve, reuse, revise, and retain (Ahn *et al*., 2017; Kim and Shim, 2014; Shabniya and Dilruba, 2017). Furthermore, researchers emphasised that the model is based on the hypothesis that similar cases will have similar problems with similar solutions. However, Ahn *et al*. (2017) argued that this technique encounters barriers such as distance measurement, selection of characteristics, assigning the weightage, and thresholds of cases in reuse. In contrast, Won-Gil *et al*. (2019), Kim and Shim (2014), and Lee *et al*. (2013) suggested that CBR can override most of the shortcomings that regression and neural networks come across by utilising the lessons learned from the historical data to solve issues in future projects.

This technique was demonstrated by Agnieszka and Krzysztof (2018) when the cost model was developed. The study developed a model with an MAER of 14% using the cost data from 143 sports field construction projects in Poland. One of the main focuses of this study was the sustainable development factors accounted for by the model, which most other models needed to be considered.

On the other hand, a broader perspective was adopted by Ahn et al. (2017). Firstly, several distance measurement tools were adopted, such as the Mahalanobis concept, Euclidean concept, arithmetic summation-based similarity measure, and fractional function-based similarity measures. Then, a cost model was developed using a database of 99 complex multi-family housing projects in Korea. In their study, 3 test rounds were carried out on the several models developed using the measurement concepts. All the models except the Mahalanobis-based model performed well, showing an MAER between 0.085 and 0.096. However, the model based on the Mahalanobis concept showed an MAER of 0.31 to 0.47.

In an investigation into optimising the weights of the cost factors of models, Kim and Shim (2014) and Lee *et al*. (2013) adapted a genetic algorithm for the CBR model. The former researchers developed a model for building projects in Korea with a minimal error rate of 5.6%. The latter developed a cost model for river facility construction projects with an average error rate of 15.49% using only eight parameters. These findings were also supported by Won-Gil *et al*. (2019), who demonstrated the high forecasting capability of their model with a low MAER of 47.9% for building projects based on the data from South Korea. In another study, Zima (2015) developed a cost model based on CBR for sports field construction works while accounting for an MAER of 5.7%. This study further stressed that the model encounters more significant errors when quantifying the numerical impact of the location and inflation factors. The evidence suggests that

having a more extensive knowledge base of cases is essential to achieve the desired results using CBR.

5.6.4 Fuzzy logic

Fuzzy logic is recognised as a model used for human communication, reasoning, and decision-making abilities in a formalised way where the information is conflicting, incomplete, and uncertain (Ahiaga-Dagbui *et al*., 2013; Fayek and Rodriguez Flores, 2010). Nevertheless, Fayek and Rodriguez Flores (2010) added that the fuzzy models are used in risk assessment, range estimating, forecasting construction performance, contractor selection, working condition assessment, and determining cost-estimating relationships. Therefore, this technique could be used more effectively if combined with other techniques.

Researchers have recently started combining several modelling techniques to compare their performance with the mono-technique models. For instance, Ahiaga-Dagbui *et al*. (2013) developed a model by combining the ANN and fuzzy set theory and considering the data of 98 water infrastructure projects completed between 2007 and 2011 in Scotland. They adopted ANN's learning and generalisation capabilities and combined them with the positive advantages of fuzzy logic. Further, the model included the reasoning and decision-making abilities of the human brain with incomplete information. As a result, the model forecasted the final cost with a significantly low marginal error between 0.6% and 0.8%. Therefore, combining several techniques while being aware of how the techniques would co-relate with each other can provide better results.

5.6.5 Monte-Carlo simulation (MCS)

Through drawing on the MCS concept, Touran and Lopez (2006) emphasised that considering factors of uncertainty in a project will lead to quantifying the project cost overruns. The previous studies agreed that the simulation could predict complicated systems where the uncertainty of

available information is higher (Shabniya and Dilruba, 2017). Shanmugam *et al*. (2006) adapted the MCS to develop a cost estimation model for building projects that forecast the cost overrun percentage. Simulation models are time-consuming in optimisation, but they allow for quicker and easier complex approximations through stage-wise refinements. Their model predicted that 65 of the 500 projects would expect a 27.49% cost overrun relative to the contract sum. Rather than providing a single outcome, they proposed various cost overrun possibilities against the respective occurrence probability.

5.6.6 Support vector machine (SVM)

This technique is based on structural risk minimisation and statistical learning theory. Additionally, models can be developed based on SVM to identify the factors, data sample collection, and training process (Cheng and Hoang, 2014). Researchers identified several benefits of using SVM as a cost model. For example, convex optimisation issues can be solved quicker than other models and with a reliable and accurate solution (El-Sawalhi, 2015). El-Sawalhi (2015) further emphasised that many research studies that developed SVM-based cost models performed better than traditional ones. Also, SVM's generalisation capabilities and sparse representation are better than other models, such as neural networks.

Additionally, SVM comprises six significant properties identified by El-Sawalhi (2015), who identified SVM as the best modelling technique in conceptual cost estimates. Finally, the research concluded with a conceptual stage cost estimation model for road projects in the Gaza Strip. The study used seventy projects with a 95% accuracy performance. Contrarily, in an investigation comparing SVM with other models, Kim et al. (2013) identified several disadvantages of SVM. The algorithmic complexity required for the model is significantly higher than that of other models, and thus, a widespread memory is essential. Also, the model needs a trial-and-error period to determine a suitable mathematical function, defined as kernel, and the variables of the selected

kernel function. Researchers highlighted that the SVM models performed equally or notably better when comparing their results to ANN and fuzzy system-based models. However, Kim *et al.* (2013) developed the SVM-based model for school building projects in Korea, achieving an MAER of 7.48%. In contrast, the ANN model, developed using the same data, achieved better results.

5.6.7 Reference class forecasting (RCF)

The literature identified RCF as a technique for cost modelling. However, the number of occurrences and the total links of the technique were below the analysis criteria. Therefore, the technique was excluded from the keyword co-occurrence analysis.

The concept of reference class forecasting is similar to the CBR model. Hence, this technique also analyses future cases and finds solutions using historical data from similar cases (Shabniya and Dilruba, 2017). However, Awojobi and Jenkins (2016) highlighted that the two techniques differ in certain degrees as CBR does not consider human behaviour's irrationality as its outcome depends on expert judgment. Furthermore, Bayram and Al-Jibouri (2016) recommend RCF at the early stage of estimation, where the uncertainty is at its highest. These findings were supported by Bayram and Al-Jibouri (2016) in their model for estimating the cost overruns of public construction projects in Turkey and concluded that the average cost overrun is 11.33%, which could vary between 22.94% to 133.48%. Another study by Fridgeirsson (2016) identified that 60% of Icelandic transportation projects faced cost overruns of 95% of their initial cost plan. Although the RCF model developed in the study provided a comprehensive insight into this matter, the researchers emphasised that the current method used in the Icelandic Road Administration (ICERA) in cost forecasting addressed the matter with a 6% cost overrun for the five years of the study.

In contrast, Themsen (2019) studied the RCF model application in a Danish mega railing project. His study concluded that the model did not support the estimation accuracy while the project faced

cost and time overruns. The research further emphasised that the model could not prevent strategic misrepresentation and optimism bias.

5.6.8 Cost modelling for infrastructure projects

Table 5.2 displays the finding of the previous studies. It summarises the models used for infrastructure projects identified in the study, the variables considered in the models, and their performances. Although CBR was identified as the third technique for cost modelling, no cost models were developed for infrastructure projects within the selected publications. The reason could be that the CBR requires an extensive cost database of previously completed projects to perform well.

Table 5.2. *Cost models developed for infrastructure projects*

Model number	Modelling technique	Reference	Project type	Variables	Performance (MSE*)
01	RA	Shr and Chen (2006)	Roads and highways	The final cost, awarded bid, days used, actual contract duration, and initial duration.	±5%
02	RA	El-Maaty *et al*. (2017)	Roads and highways	Inadequate project planning and execution, insufficient cost planning and monitoring, lack of communication between construction parties, price fluctuations, lack of proper technical study before the tender by the contractor, errors in project quantities measurements, slow decision-making process, equipment failures, lack of adequate field visits before tendering by the contractor, inappropriate use of project site, improper use of materials, inaccurate drawings and contract documents, material monopoly by suppliers, inflation, and rework.	+30.42%
03	RA	Sodikov (2005)	Roads and highways	Predominant work activity, project duration, pavement width, shoulder width, ground rise fall, average site clear, earthwork volume, surface class, and base material.	+30% to +36%
04	RA	Sonmez and Ontepeli (2009)	Urban railways	*Primary parameters* Percentage of tunnel section over the total length of the rail, percentage of the total length of elevated stations over the total rail length, percentage of the total length of at-grade stations over the rail length, percentage of the total length of cut-and-fill method over the main line length, supply and installation of the rails, and the number of underground stations. *Secondary parameters* Contract type, number of at-grade stations, number of elevated stations, the main line length, and percentage of the total length of depressed-open sections (ramps) to the total rail length.	+35.2%
05	ANN	Adel *et al*. (2016)	Roads and highways	Project scope, duration, year of construction, project region, mainline length, mainline classification.	Training - +4.51%; Evaluation-

					+5.8%; Validation - +16.0%
06	ANN	El-Kholy (2019)	Roads and highways	Inadequate project planning and execution, insufficient cost planning and monitoring, lack of communication between construction parties, price fluctuations, lack of proper technical study before the tender by the contractor, errors in project quantities measurements, slow decision-making process, equipment failures, lack of adequate field visits before tendering by the contractor, inappropriate use of project site, improper use of materials, inaccurate drawings and contract documents, material monopoly by suppliers, inflation, and rework.	+39.8%
07	ANN	Sodikov (2005)	Roads and highways	Predominant work activity, project duration, pavement width, shoulder width, ground rise fall, average site clear, earthwork volume, surface class, and the base material.	+24% to +26%
08	ANN	Sonmez and Ontepeli (2009)	Urban railways	*Primary parameters* Percentage of tunnel section over the total length of the rail, percentage of the total length of elevated stations over the total rail length, percentage of the total length of at-grade stations over the rail length, percentage of the total length of cut-and-fill method over the main line length, supply and installation of the rails, and the number of underground stations. *Secondary parameters* Contract type, number of at-grade stations, number of elevated stations, the main line length, and percentage of the total length of depressed-open sections (ramps) to the total rail length.	Model 1 - +49.8%; Model 2 - +33.3%
09	SVM	El-Sawalhi (2015)	Roads and highways	Road area, road surface type, base course type, base course thickness, interlock thickness, asphalt thickness, pipe diameter, manhole depth, cut and fill volume, curb length	-5% average error
10	Fuzzy logic	El-Maaty *et al*. (2017)	Roads and highways	Inadequate project planning and execution, insufficient cost planning and monitoring, lack of communication between construction parties, price fluctuations, lack of proper technical study before the tender by the contractor, errors in project quantities measurements, slow decision-making process, equipment failures, lack of adequate field visits before tendering by the contractor, inappropriate use of project site, improper use of materials, inaccurate drawings and contract documents, material monopoly by suppliers, inflation, and rework.	+40.37%
11	ANN and Fuzzy hybrid	Ahiaga-Dagbui *et al*. (2013)	Water infrastructure	Tendering strategy, site access, type of location, project type, Contractor's need, soil type, initial cost, and the initial duration.	+0.6% to +0.8%
12	RA, SVM and data mining hybrid	Ahiaga-Dagbui and Smith (2014)	Water infrastructure	Tendering strategy, procurement strategy, ground condition, soil type, delivery partner, scope, the purpose of the project, and location	-3.83% to +2.33%

**MSE – Mean Squared Error*

The findings of Table 5.2 were further analysed using a boxplot in Figure 5.6. In Table 5.2, it is not clear how the distribution of error percentage varies. However, the research of Shr and Chen (2006) and Sonmez and Ontepeli (2009) did not include the details of the error percentages of the

model for each validation data except the mean error. Therefore, the boxplot did not consider models 1, 4, and 8.

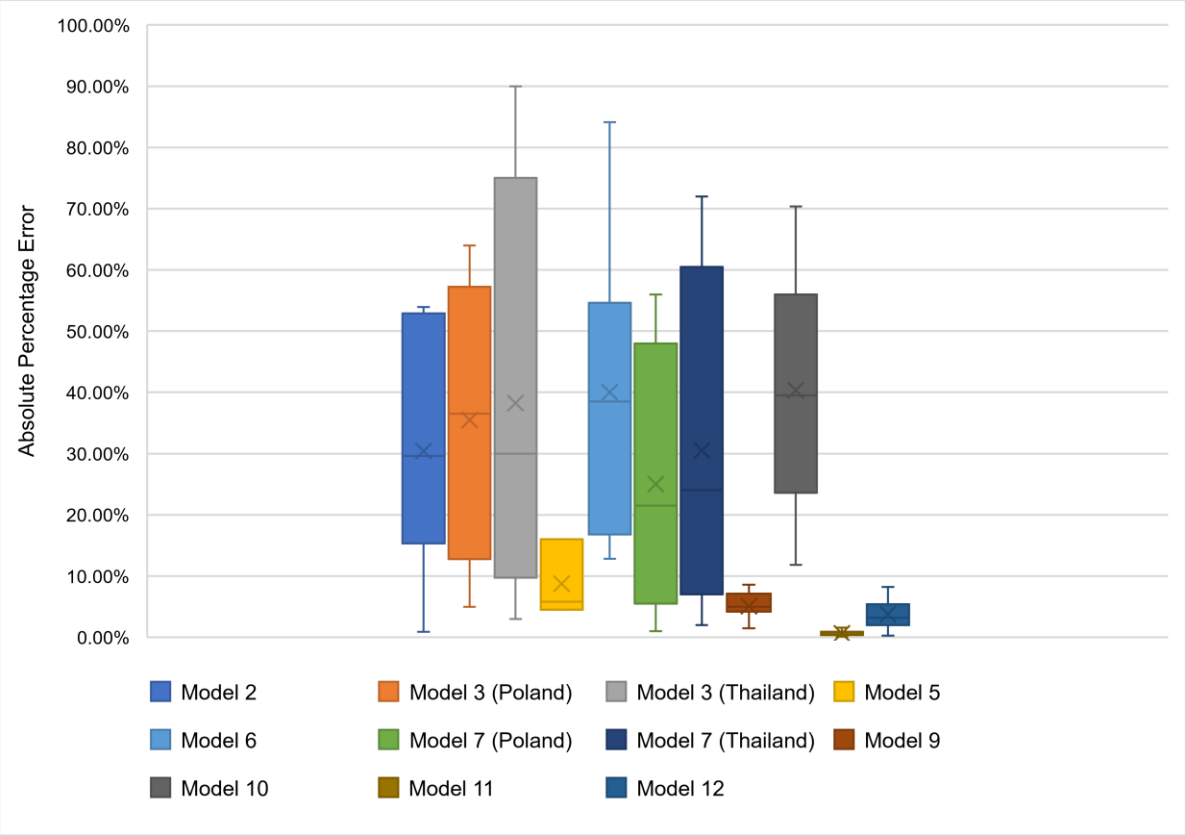

Figure 5.6. *Model performance comparison based on the Absolute Percentage Error*

According to the results, ANN showed a lower performance when the model considered more variables. For example, the two ANN models that Sonmez and Ontepeli (2009) developed comprised six primary and five secondary variables. As a result, the models showed a higher rate of errors ranging from 33% – 50%. However, Adel *et al*. (2016) created a model that achieved significantly better performance using only six technical variables. The model achieved a 4.51% error rate in training, 5.8% error in evaluation and +16% error in validation. Similar behaviour was noticed with RA as well. The RA model with fewer variables achieved ±5% marginal error (Shr and Chen, 2006). Conversely, Sodikov's (2005) model showed a 30% - 36% error, considering only seven numerical variables.

El-Sawalhi (2015) developed a model based on SVM with an error of -5%. Therefore, SVM can be a sound modelling technique for estimating infrastructure projects. Meanwhile, both of the models El-Kholy (2019) developed performed poorly. One model was based on ANN, while the other was based on fuzzy. The error rate of both models was approximately 40%. The reason could be that the models considered several qualitative risk factors. Therefore, selecting variables for the model is crucial for the model's performance.

On the other hand, it was observed that hybrid models could achieve better results than mono-technique-based models. For example, Ahiaga-Dagbui *et al*. (2013) and Ahiaga-Dagbui and Smith (2014) developed two hybrid models for the same data set. One model combined ANN and fuzzy logic, while the other combined RA, SVM, and data mining techniques to develop the hybrid model. Both models achieved significantly better MAER. Moreover, the ANN-fuzzy hybrid achieved the lowest performance of the two models (Ahiaga-Dagbui *et al*., 2013; Ahiaga-Dagbui and Smith, 2014). Therefore, it can be observed that combining several modelling techniques helps mitigate the disadvantages of one technique with the advantages of the other.

Consequently, Figure 6 indicates that models 5, 9, 11, and 12 show lower error rates. In addition, these models exhibit a minimum spread of the error percentage. In contrast, other models express a widespread error range. Therefore, models 5, 9, 11, and 12 produce more reliable results than the other models. For instance, although less information was available regarding the error percentages, model 01 indicates an error range between ±5%. Analysing the crucial characteristics of the models in Table 2, it can be emphasised that the models with fewer variables achieved better results than those with an extensive list of variables. Significantly, those identified variables were also simplified and more technical. For instance, model 5 variables were project scope, duration, year of construction, project region, mainline length, and mainline classification. Therefore,

modelling techniques such as RA, ANN, and SVM are not suitable for predicting the risk-related variables because all the other models contain variables related to the qualitative risk factors.

Based on the above findings, the most effective modelling techniques for estimating infrastructure projects are RA, ANN, SVM, and hybrid models.

5.7 Conclusions

A construction project's cost estimation must be accurate and reliable, especially when the uncertainty is high and little information is available. Therefore, it is crucial to develop a realistic budget as it affects most design, construction, and technology decisions. Cost overrun in construction projects is a global phenomenon due to a need for more accuracy in traditional estimation techniques. Therefore, researchers have focused on developing cost models using statistical techniques in the past few years.

The literature review identified seven formidable statistical modelling techniques: RA, ANN, CBR, fuzzy logic, MCS, SVM, and RCF. These techniques were used in cost estimation as single technique-based models or multi-techniques combined as hybrids. Further, it was observed that the cost estimation models based on these techniques could provide more reliable estimates. Therefore, the above could be used to improve traditional practice. Moreover, the most reliable modelling techniques with better performance were RA, ANN, SVM, and hybrids. Although CBR and RCF are also modelling techniques with better performances, they must have a more extensive cost database to provide more reliable outputs.

Consequently, the study identified twelve models developed for infrastructure projects covering roads and highways, railway, and water infrastructure. Based on the percentage error, the models were evaluated and observed that RA, ANN and SVM techniques were used in models with lower error rates. In addition, models with several techniques combined indicated even lower error rates.

Subsequently, the study emphasised that the techniques mentioned earlier perform better with technical variables than qualitative risk-related variables.

Considering the study analysis and findings, it is recommended to incorporate these techniques to mitigate the current issues and enhance the estimation practice in infrastructure projects. Based on the findings of this study, several future studies emerged. The hybrid models identified indicated the lowest error rates. However, those models were developed for water projects. Therefore, is it possible to develop such a hybrid model with a similar low error rate for other infrastructure project types? Subsequently, although much research has been done, how much of this research has impacted the industry? Are these models used in the industry to reduce cost overruns, and if not, how can these research outcomes impact the industry?

The study was limited to the findings from a bibliometric literature review conducted using the Scopus database. Therefore, other techniques and models could not be identified in this review.

5.8 Epilogue

This Chapter reviewed the literature systematically to identify the possible and most used statistical techniques for cost estimation. It also analyses the techniques with low error rates and high performance. Consequently, the Chapter investigates the models developed for infrastructure projects and compares them against their performance and the variables considered. Based on the findings, recommendations were also provided for the future development of cost models for infrastructure projects, specifically for road projects. The next Chapter develops a cost model for road projects in NZ using the regression analysis technique.

6 Conceptual cost estimation model for pre-design stage of road projects in New Zealand using regression analysis

6.1 Prologue

Based on the findings of Chapter 05, regression analysis is an effective modelling technique often used in the construction industry for cost estimation. Therefore, in this chapter, a regression analysis model will be developed, tested, and validated using actual cost data from the NZ road projects. Subsequently, conclusions and recommendations are presented comparing the model's performance and reliability of regression analysis as a technique for cost estimation[4].

[4] This chapter is based on the following journal paper (under review)

Atapattu, C.N., Domingo, N.D., and Sutrisna, M. (2023). A conceptual cost estimation model for the pre-design stage of road projects using regression analysis. *Journal of Financial Management of Property and Construction*. Manuscript ID: JFMPC-08-2023-0052R1

6.2 Abstract

The current estimation practice in construction projects greatly needs upgrading as there has been no improvement in the cost overrun issue over the past seventy years. This research developed a new multiple regression analysis (MRA) based model to forecast the final cost of road projects at the pre-design stage using data from forty-three projects in New Zealand (NZ). The research used the case study of forty-three completed road projects in NZ. Document analysis was conducted to collect data, and statistical tests were used for model development and analysis. Eight models were developed, and all models achieved the required F statistics and met the regression assumptions. The models' mean absolute percentage error (MAPE) was between 21.25% and 22.77%. The model with the lowest MAPE comprised the road length and width, number of bridges, pavement area, cut and fill area, preliminary cost and cost indices change. No research was conducted to adopt cost modelling techniques to the conceptual estimation practice in the NZ construction industry. The model is based on road projects in NZ. However, it was designed to be able to adapt to other contexts. The findings suggest that the model can be used to improve traditional conceptual estimating methods. The project team often stores past project data but is rarely used for analysing and forecasting purposes. This research emphasises that past data can be utilised effectively to predict the project cost at the pre-design stage with limited information.

6.3 Introduction

Cost estimation is a crucial task at the early design stage of a project because it is essential in decision-making. Notably, it is scarce to complete a project within the budget (Flyvbjerg et al., 2002). It is even more challenging in road projects, considering the scope, site conditions, and duration of projects (Ammar et al., 2022). Therefore, road construction is exposed to higher risks than building projects. Hence, researchers stress the importance of accurate and reliable estimates because underestimation causes overruns in cost and time (Flyvbjerg et al., 2002; Odeck, 2004).

At the same time, overestimation may detain funds unnecessarily, which could be used for essential purposes (Cantarelli et al., 2012). However, underestimation is the most common and more severe issue, as Flyvbjerg et al. (2002) found that 9 out of 10 projects faced cost overruns.

Barlow et al. (2017) reported that according to the project management survey carried out by KPMG New Zealand (NZ), from 2010 to 2017, the number of projects delivered within the budget had dropped from 48% to 29%. Furthermore, Flyvbjerg et al. (2003) investigated 258 infrastructure projects worldwide and concluded that road projects generally face 21% of the mean cost overruns. That means 21% of multi-million dollar road projects can negatively affect the country's economy. However, if the initial budget can predict the 21% cost overrun at the beginning of the project, then the investors could make factual decisions before facing unnecessary risks during the construction stage. Moreover, the construction industry is evolving with new technology, such as building information modelling (BIM), virtual reality, and blockchain technology. Thus, the current estimation practice needs improvement to address these changes. However, Flyvbjerg et al. (2003) observed that the current estimation practice had not improved over the past 70 years. For this reason, the cost overrun issue has been continuously challenging.

Over the years, much research has been done to develop cost models using various modelling techniques. Chapter 5 identified seven techniques that can be used to develop cost-forecasting models. The authors further discovered that regression analysis (RA), artificial neural network (ANN) and support vector machine (SVM) showed better performance for infrastructure project cost estimation.

Even though NZ road projects have faced significant cost overruns continuously over the past 20 years (NZTA, 2022), no research has been done to minimise the cost overruns and improve the conceptual cost estimation methods at the pre-design stage. According to Atapattu et al. (2023), most of the models developed using modelling techniques cannot be used for the conceptual

estimate because they require information for variables that are not available during the pre-design stage of the project (Ahiaga-Dagbui et al., 2013; El-Maaty et al., 2017; El-Sawalhi, 2015; Shr and Chen, 2006; Sodikov, 2005; Sonmez and Ontepeli, 2009). Therefore, this study proposes a model using RA to determine the final cost of road projects in NZ that can be calculated during the pre-design stage of the project. The study's outcome includes a regression analysis-based cost estimation model aimed at improving the performance of traditional estimation practices.

The following section gives an overview of the multiple regression analysis as a technique for cost estimation modelling.

6.4 Multiple linear regression analysis

The primary aim of the multiple regression analysis is to estimate the relationship between a dependent variable and a set of independent variables that affect the dependent variables (Jablonowski and MacEachern, 2009; Young, 2017). Multiple regression analysis can be generally represented in the following equation (Kim et al., 2004; Young, 2017).

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots\ldots + \beta_n * X_n + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots (Eq.\ 1)$$

Where 'Y' is the response or the dependent variable, 'Xn' is the independent variable, and 'βn' is the Coefficients or parameters to be estimated for the independent variables. Further, Young, 2017) explained that 'ε' is an error term derived from the difference between the dependent variable's actual and predictor values. The regression analysis aims to estimate each variable's coefficients, which will become the β in accordance with the above equation (Jablonowski and MacEachern, 2009; Kim et al., 2004; Young, 2017).

There must be no multicollinearity among the independent variables. Otherwise, even if the regression model is passed model fitness, the effect of the individual variables cannot be estimated

independently (Gordon, 2010; Young, 2017). That is because when multicollinearity is present in the model, the coefficients of the variables are correlated (Fordon, 2010; Young, 2017).

6.5 Regression assumption

Before developing the regression model, several assumptions were made at the beginning that will be tested. Young (2017) explained that four major theoretical assumptions are required for regression model development: linearity, normality, independence, and homoscedasticity.

In the linearity assumption, it is assumed that there is a linear relationship among the variables (Freund et al., 2006; Young, 2017). The second assumption is normality, where it is assumed that the residuals of the regression are normally distributed with a mean of zero (Freund et al., 2006; Gordon, 2010; Young, 2017). The next assumption is that the variables are independent without a significant correlation (Young, 2017). The final assumption is that the variance of the data points is consistent for all the data points, known as homoscedasticity (Panik, 2009; Young, 2017). Once the model is developed, the above four assumptions will be tested.

The following section describes the research methodology adopted in this research.

6.6 Research methodology

The study aimed to develop a model to estimate the cost of road projects in NZ at the early stage of the design. The detailed unit estimation method cannot be applied because the design is not finalised. Cost data were collected from 50 road projects constructed between 2002 and 2022. A limited number of road projects were carried out with a significant scope of work in NZ. Further, NZ has more alterations to existing roads and few new construction projects.

Moreover, the study was conducted during the COVID-19 period. That means there were difficulties in approaching companies and people to obtain the data. as well as the fact that private

companies have certain restrictions on giving out the sensitive data of their projects. Hence, government agencies were approached to collect the data. Road projects can include new construction, major and minor alterations, renovation and maintenance, and miscellaneous works. However, this study collected only the projects of either new construction or significant alterations.

Having limited the scope, the next focus was to decide how the procurement, tendering, and contract methods affect the cost. Furthermore, the model should forecast the effect of these three on the project cost and whether the cost database should include a sufficient number of projects from all types of procurement, tendering, and contract methods. Quantifying the monetary effect based on the type of procurement, tendering, or contract method is in itself challenging. However, considering the difficulties mentioned above, it was determined that the regression-based model would not consider any qualitative variables.

The Statistical Package for Social Science (SPSS) software was used to run the regression model. The general model statistics are tested with parametric tests such as analysis of variance (ANOVA) because *F* statistics provide the significance of the model in predicting the dependent variables. However, the data distribution must be normal to use parametric tests in regression models. Therefore, before the model development, the data distribution is assumed to be normal. Once the model is developed, a normality test must be conducted for the model output. Young (2017) provided several tests for normality, such as the Anderson-Darling test, Kolmogorov-Smirnov test, Shapiro-Wilk test, Ryan-Joiner test, and Chi-Square goodness-of-fit test. The most common normality tests are the Kolmogorov-Smirnov and the Shapiro-Wilk tests, supported by SPSS. Therefore, this study conducted these two tests for normality.

Then, the regression function in SPSS was used in model development, analysis, testing, and validation of the model. There are several approaches to finding the best variable combination,

such as backward elimination and forward selection. According to Ryan (2009), backward elimination is preferred over forward selection by most data analysts. Therefore, this study used backward elimination for the selection of variables.

Once the model was developed, required tests were carried out to check the regression assumptions' validity. Then, the model was validated using additional project data collected from sixteen projects of a similar nature. Finally, the model results were compared with and discussed similar literature findings on road project cost estimation. Figure 6.1 illustrates the research process carried out from the data collection to the model validation.

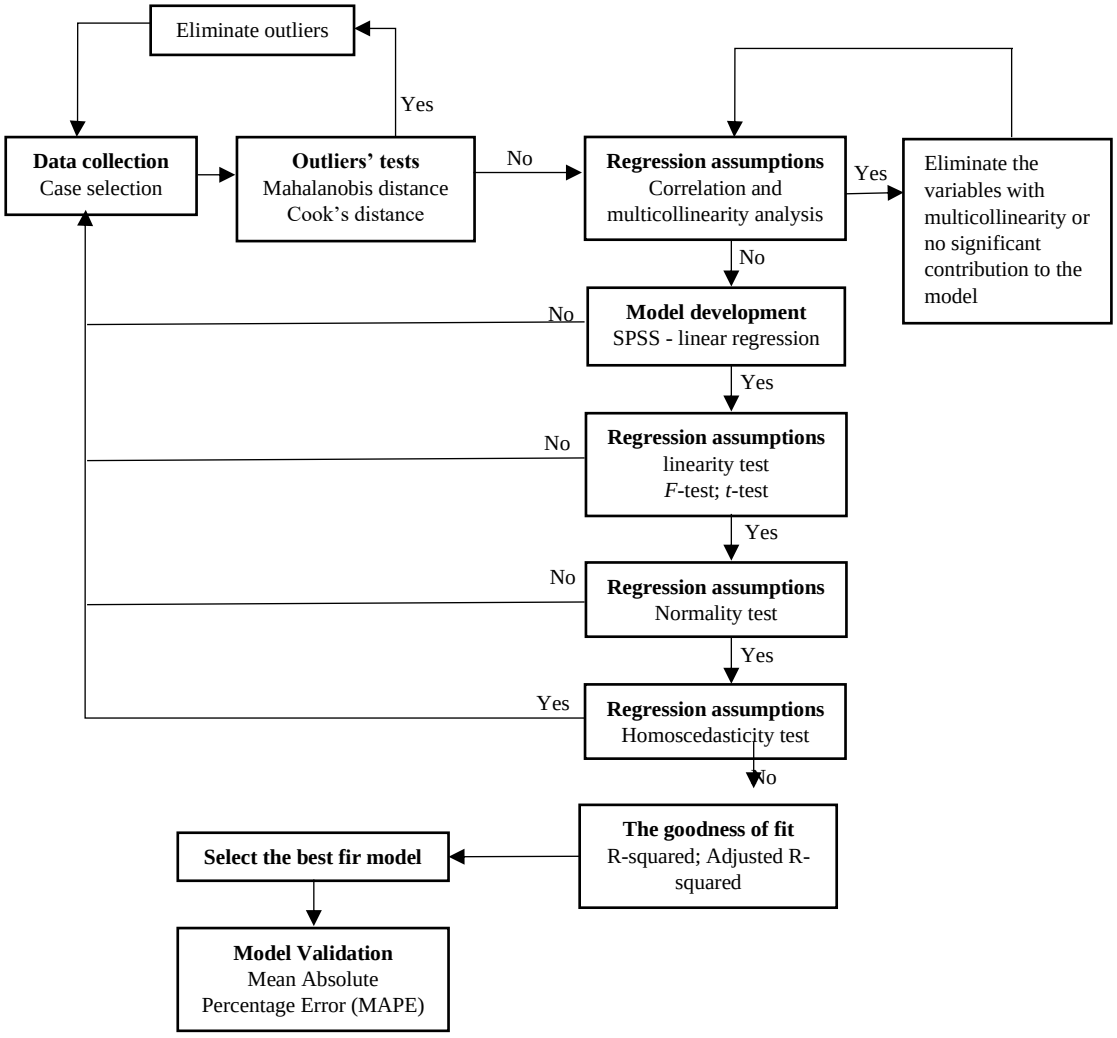

Figure 6.1. *Detailed model development process*

The following section discusses the findings and compares them with the models developed in the relevant literature.

6.7 Regression model development

6.7.1 Case selection

Before conducting the regression analysis, it is crucial to test if all the cases are suitable for developing the model. If there are any cases with unusual data, then the model-fitness test will fail, and the accuracy and reliability of the model will be decreased (Freund et al., 2006; Panik, 2009).

Mahalanobis distance shows the distance of the case from the centroid of all the cases for the independent variables. A considerable distance indicates that the case is an outlier for the variables. To determine the significance of the distance, Mahalanobis distance must compare its *p*-value with the margin of the alpha level. Tabachnick and Fidell (2001) suggested that if the *p*-value of the Mahalanobis distance is less than 0.001 alpha level, then that case is considered an outlier. Out of the 50 road projects, 7 showed Mahalanobis distances with *p*-values less than 0.001. Even though it was retested, removing each independent variable, there was no significant improvement. Therefore, 7 cases were deleted from the database as they could reduce the reliability of the data. According to Jablonowski and MacEachern (2009), a regression model with a sample size larger than 30 can provide more reliable results. Therefore, the model was developed based on 43 projects which were not outliers.

The outliers were further confirmed with Cook's distance measure. It measures the influence of a particular data point on the predicted values of all observations (Freund et al., 2006; Young, 2017). According to Gordon (2010), Cook's distance should be less than the "4/ sample size" value; if not, datapoint is considered an outlier. However, all 43 cases selected above showed Cook's

distance less than 0.093 (derived from 4/43). Therefore, it was further confirmed that the outliers were eliminated.

6.7.2 Variables selection

This study is focused on the technical variables of the project cost. First, variables were listed based on the selection from other similar RA-based models. Once all the variables were listed, the next step was identifying the variables that could be used for pre-design stage estimation. This step was further verified by the experts involved in the data collection. The third step was determining which variables' data can be extracted from the selected cases. According to the above three-step criteria, fifteen variables were selected. Table 6.1 shows the independent variables selected for the regression models. Twelve independent variables were selected, excluding three variables, and eight different models were created based on the selection of the variables.

Table 6.1. *Variables considered in the regression model*

Independent Variables	Unit	*p*-values of the variables							
		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
Road length	m	0.002	0.002	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Road width	m	0.710	0.707	0.703	0.703	0.687	0.741	-	-
Distance from the nearest major city	m	0.899	-	-	-	-	-	-	-
The number of bridges	m^2	0.021	0.017	0.01	0.007	0.005	0.004	0.003	0.003
The approximate length of retaining walls	m^2	0.791	0.778	0.751	0.727	-	-	-	-
Ground improvements area	m^2	0.746	0.747	0.755	-	-	-	-	-
Pavement area	m^2	0.053	0.046	0.040	0.027	0.017	0.014	0.013	0.025
Cut and fill area	m^3	0.307	0.296	0.285	0.294	0.226	0.227	0.192	-
Project duration	Days	0.881	0.894	-	-	-	-	-	-
Expected year of completion	Year	0.774	0.723	0.709	0.709	0.776	-	-	-
Expected preliminary cost (% of the total cost)	%	0.049	0.043	0.034	0.022	0.019	0.002	0.002	0.001
Expected % change in Construction Cost Index*	%	0.0875	0.067	0.059	0.054	0.049	0.023	0.012	0.016
Excluded variables									
Kerbs length	m								
Road markings	m								
Base course volume	m3								

m - linear meters; m^2 – square meters; m^3 – cubic meters; Days – calendar days; % - Percentage
** 2022 was considered the base year*

Backward elimination was used to find the best combination of the variables. Ryan (2009) explained that if there are a "*k*" number of variables, there are (2*k*-1) possible combinations. Therefore, in this study, there are 4,095 combinations with twelve variables. It is impossible to check all the possible combinations to find the best combination. Then, if the forward selection method is chosen, it is difficult to decide which variable combination to choose. Consequently,

backward elimination was chosen to eliminate the variable with the highest *p*-value in each step. In contrast to forward elimination, only a maximum of twelve variable combinations are to be examined in this method.

The variables were selected depending on the *p*-values. However, SPSS stops variable elimination once all the variables achieve *p*-values less than 0.05. Therefore, in this study, only eight models were examined instead of twelve. According to Table 6.1, Model 1 is comprised of all twelve variables. Then, from models 1 to 8, one variable was eliminated at each step according to the backward elimination technique. According to Table 6.1, model 8 is the only model with only the significant variables. However, the best-fit model cannot be decided based on the individual *p*-values of the independent variables (Hoffmann, 2022). Therefore, the model statistics must be compared with the variable statistics. Nonetheless, multicollinearity and correlation issues must be eliminated before the model development.

6.7.3 Multicollinearity

In addition to the selected twelve variables, three other variables were considered: kerb length, road markings, and base course volume. However, the regression showed a multicollinearity issue in these variables because all three factors showed Variance Inflation Factors (VIF) higher than 14. According to Young (2017), if the VIF>5, the factor shows high multicollinearity. Moreover, the road length variable was eliminated from the second model when the above three variables were present. However, road length is one of the significant variables for road project cost. The reason could be that the above three variables significantly correlate with road length. Therefore, those variables were excluded from further analysis. Once the three variables were removed from the model, X1 became a significant variable and was included in all eight models.

Nevertheless, the VIF of all twelve variables in Table 6.1 were below 5 in all eight models. According to Snee (1983), if the VIF<10, observing the eigenvalues for multicollinearity is not

required. Therefore, multicollinearity was eliminated in these models. The VIF values of the variables are shown in Table 6.2. However, even if the multicollinearity is not detected, it is essential to check for any strong correlations among the variables.

Table 6.2. *VIF of the variables in each model*

Independent Variables	Unit	VIF of the variables							
		Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**	**Model 7**	**Model 8**
Road length	m	1.982	1.943	1.806	1.771	1.738	1.731	1.727	1.692
Road width	m	1.463	1.459	1.455	1.454	1.453	1.270	-	-
Distance from the nearest major city	m	1.362	-	-	-	-	-	-	-
The number of bridges	m^2	1.490	1.482	1.386	1.351	1.168	1.104	1.098	1.096
The approximate length of retaining walls	m^2	1.588	1.582	1.538	1.530	-	-	-	-
Ground improvements area	m^2	1.775	1.757	1.698	-	-	-	-	-
Pavement area	m^2	4.746	4.663	4.661	3.648	3.463	3.244	3.212	1.710
Cut and fill area	m^3	3.004	2.974	2.903	2.672	2.475	2.285	2.221	-
Project duration	Days	1.248	1.212	-	-	-	-	-	-
Expected year of completion	Year	3.333	3.010	3.000	2.997	2.707	-	-	-
Expected preliminaries cost (% of the total cost)	%	2.088	2.071	2.026	1.927	1.924	1.187	1.175	1.154
Expected % change in Construction Cost Index*	%	2.021	2.019	2.018	2.015	1.798	1.154	1.045	1.039

m - linear meters; m^2 – square meters; m^3 – cubic meters; Days – calendar days; % - Percentage
** 2022 was considered the base year*

6.7.4 Correlation analysis

It is vital to have a set of variables with no correlation. If there is a strong correlation between two variables, the estimate of one regression coefficient is affected by the presence of the other predictor variables (Young, 2017). The author further explained that the coefficient of correlation gives the proportional change in a variable when the other variables are held constant. Generally, this value lies between -1 and +1. If the coefficient is +1, it is a strong positive correlation, while -1 means a strong negative correlation (Asuero et al., 2006). However, the authors suggest that with actual data, it is challenging to achieve either extremes or zero correlation (Asuero et al., 2006; Young, 2017).

Asuero et al. (2006) further argued that if the correlation coefficient is zero, it does not necessarily mean the two variables are statistically independent. However, according to the correlation matrix of this regression model, the correlation coefficients of all the above twelve variables were less than 0.5. Therefore, none of the models developed in this study contain any variables with strong

correlations. Hence, the model statistics can be examined for further analysis. Table 6.3 shows the correlation matrix of the variables.

Table 6.3. *Correlation matrix*

Variables	Road length	Road width	Distance from the nearest major city	The number of bridges	The approximate length of retaining walls	Ground improvements area	Pavement area	Cut and fill area	Project duration	Expected year of completion	Expected preliminaries cost (% of the total cost)
Road width	0.160	-	-	-	-	-	-	-	-	-	-
Distance from the nearest major city	0.140	-0.206	-	-	-	-	-	-	-	-	-
The number of bridges	-0.038	0.015	-0.045	-	-	-	-	-	-	-	-
The approximate length of retaining walls	-0.130	-0.136	0.059	0.283	-	-	-	-	-	-	-
Ground improvements area	0.407	0.125	0.044	0.207	0.062	-	-	-	-	-	-
Pavement area	0.591	0.328	0.102	0.214	0.012	0.056	-	-	-	-	-
Cut and fill area	0.325	0.325	-0.025	0.204	-0.186	0.260	0.023	-	-	-	-
Project duration	0.077	-0.135	-0.084	0.092	-0.058	-0.130	-0.155	-0.172	-	-	-
Expected year of completion	-0.201	-0.462	0.385	-0.167	0.226	-0.062	-0.242	-0.253	0.110	-	-
Expected preliminaries cost (% of the total cost)	-0.326	-0.097	-0.024	-0.032	0.183	-0.011	-0.253	-0.078	0.115	0.459	-
Expected % change in Construction Cost Index*	-0.053	-0.305	0.282	-0.006	-0.168	-0.068	-0.113	-0.043	0.050	0.487	-0.121

6.8 Regression model

Hypothesis testing is required to test the regression model. Accordingly, the null and alternate hypotheses can be written as follows.

H_0: $\beta_1 = \beta_2 = \ldots\ldots = \beta_n = 0$

H_a: Any of the coefficients or at least one coefficient is not equal to zero.

If the null hypothesis is accepted, no regression model exists in this dataset. The above hypotheses can be tested using the statistical *F*-test. Gordon (2010) explained that the purpose of the *F*-test is to test whether at least one of the model's predictor variables is significant and can predict the

dependent variable. Once the model is developed, regression assumptions must be tested before investigating the model's fitness. Out of the four regression assumptions, only normality and homoscedasticity are the remaining assumptions to be tested. The following sections discuss the assumption testing.

6.8.1 Regression assumption testing – Normality test

Once the regression model is created, the normality assumptions must be confirmed. Otherwise, the model cannot be validated based on the experiment's parametric tests. For the normality test, the hypotheses can be written as follows.

H_0: The errors follow a normal distribution

H_a: The errors do not follow a normal distribution

The Kolmogorov-Smirnov test and Shapiro-Wilk tests were conducted to test the normality. According to Table 6.4, the *p*-values of both normality tests are greater than 0.05 alpha level. Therefore, the null hypothesis cannot be rejected. Hence, it is confirmed that the data distribution is normal. Therefore, the ANOVA test results of the model are valid.

Table 6.4. *Normality test results*

Normality test	Test Statistics	*p*-value
Kolmogorov-Smirnov test	0.116	0.200
Shapiro-Wilk test	0.933	0.068

6.8.2 Regression assumption testing – Homoscedasticity test

The last regression assumption to be tested is homoscedasticity. Under this assumption, the regression model is expected to show a constant variance in its variance or the error term (ε) (Young, 2017). If the variance is not constant, then the behaviour is called heteroscedasticity. That can be tested through the scatter plot between the standardised predicted value and the standardised residuals. If the scatter plot shows a pattern, then the model faces the issue of heteroscedasticity

(Freund et al., 2006). The authors further stated that, alternately, the scatter plot without a pattern shows homoscedasticity.

Therefore, Figure 6.2 illustrates the scatter plot between the standardised predicted value and the standardised residuals of this study. However, the plot shows no pattern, and the data points are randomly scattered. Consequently, the regression models of this study have met the assumption of homoscedasticity.

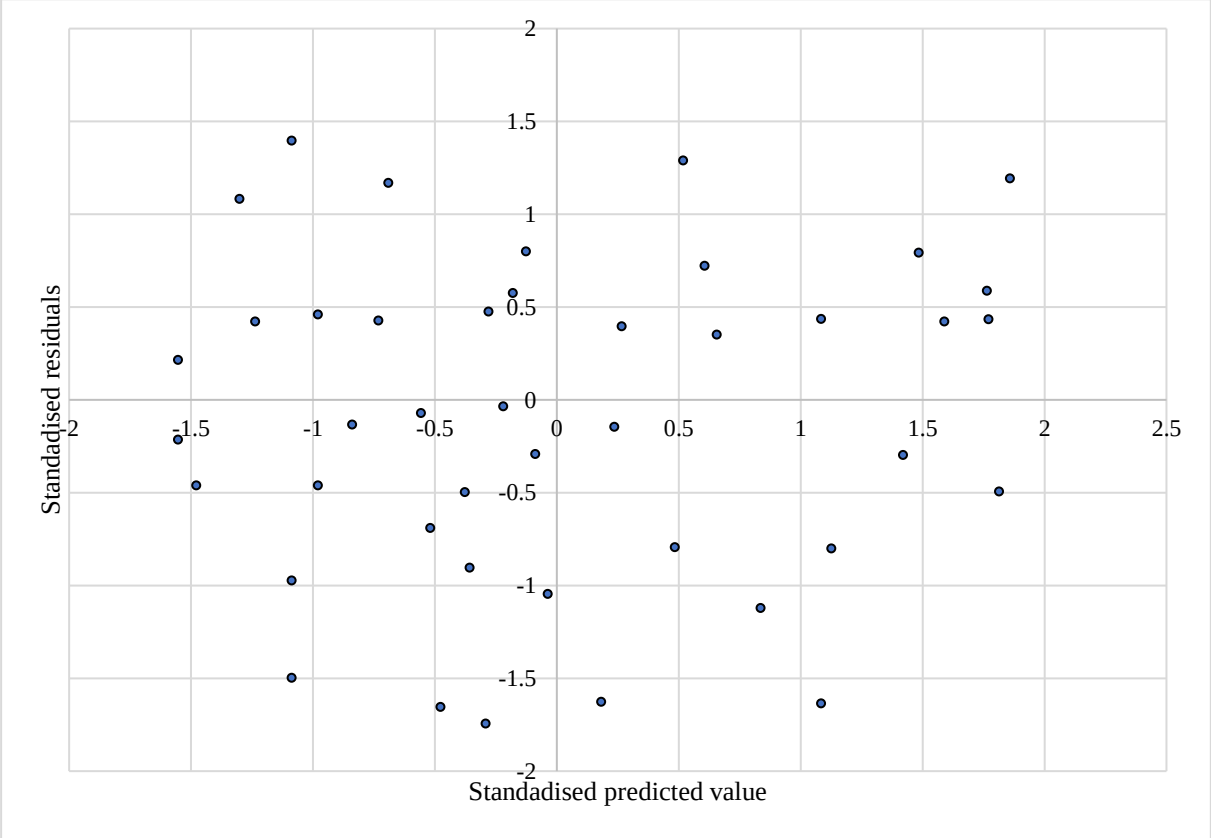

Figure 6.2*. Scatter plot of standardised predicted value against the standardised residuals*

6.8.3 Model fitness

Table 6.5 summarises the model statistics of the regressions. Table 6.5 shows that all the *F* values are greater than the *F*(critical) value. Therefore, the Null hypothesis is rejected as the dataset has a regression relationship (Gordon, 2010). Further, to accept the alternate hypothesis, the overall *p*-value of the model should be less than 0.05 alpha level. Conferring to Table 6.5, the *p*-values of

all eight models are less than the alpha level of 0.05. Therefore, the null hypothesis is rejected, and it is established that the independent variables can predict the dependent variable. Further to the above, the backward elimination has stopped at model 8 as all the *p*-values have reached below 0.05. Therefore, the other four models were not considered.

Table 6.5. *Statistics of the regression models*

	Model							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
R	0.829	0.829	0.829	0.828	0.826	0.825	0.826	0.808
R-Squared	0.687	0.687	0.687	0.685	0.683	0.680	0.682	0.654
Adjusted R-Squared	0.453	0.485	0.513	0.536	0.556	0.593	0.576	0.578
Standard error of the estimate (millions)	4.94	4.80	4.67	4.56	4.45	4.27	4.36	4.34
F-value	2.933	3.394	3.947	4.592	5.386	7.790	6.424	8.677
F(critical) value	2.092	2.114	2.142	2.179	2.225	2.364	2.285	2.470
df (degree of freedom)	12, 30	11, 31	10, 32	9, 33	8, 34	6, 36	7, 35	5, 37
p-value	0.023	0.012	0.006	0.003	0.001	<0.001	<0.001	<0.001

Furthermore, the model fit can be further tested using R values. R is the correlation between the observed and predicted values of the dependent variable, which is the final cost of the road projects. R-squared is the proportion of variance in the final project cost, which can be predicted from the variables considered in the respective models (Gordon, 2010).

According to the R-squared values shown in Table 6.5, all eight models perform well within a similar range. In models 1, 2, and 3, 68.7% of the project's final cost variance can be predicted using the considered variables. However, models 4, 5, 6, 7, and 8 can predict 68.5%, 68.3%, 68%, 68.2%, and 65.4% of the final cost variance, respectively, which is approximately close. Therefore, based on these values, the ability to predict the final cost is reduced when the number of variables considered is reduced. Hence, out of these models, models 1, 2, and 3 are better than the others based on the R square values.

However, researchers emphasised that the model's fitness cannot be decided based on the high R-squared values. They further explain that adding more variables will make the model more complicated to achieve a high R-squared value (Gordon, 2010; Freund et al., 2006; Ryan, 2009). Therefore, adjusted R-squared is used to overcome the issues with R-squared because Freund et

al. (2006) and Ryan (2009) explained that adjusted R-squared would not be increased by adding more variables to the model. Similarly, according to Table 6.5, model 6 has the highest adjusted R-squared compared to all the others, even though the R-squared of model 6 is less than that of the six other models. Therefore, based on the adjusted R-squared, model 6 has the highest capability of predicting the final cost compared to the other three models.

Furthermore, model fitness must be tested using adjusted R-squared and standard error (SE). Lower SE values indicate better model fitness (Freund et al., 2006; Gordon, 2010; Ryan, 2009). Consequently, model 6 considers all the most essential variables while indicating an appropriate adjusted R-squared value and lowest standard error.

6.8.4 Predicted values and residuals

Figure 6.3 shows the regression model's normal p-p plots of standardised residuals. It compares the cumulative probability of the observed value against the cumulative probability of the expected value. According to Figure 6.3, the probabilities of the cases used for the models lie very close to the reference line without significant deviations. Therefore, the data distribution is normal (Young, 2017). However, normality must be further confirmed with normality tests.

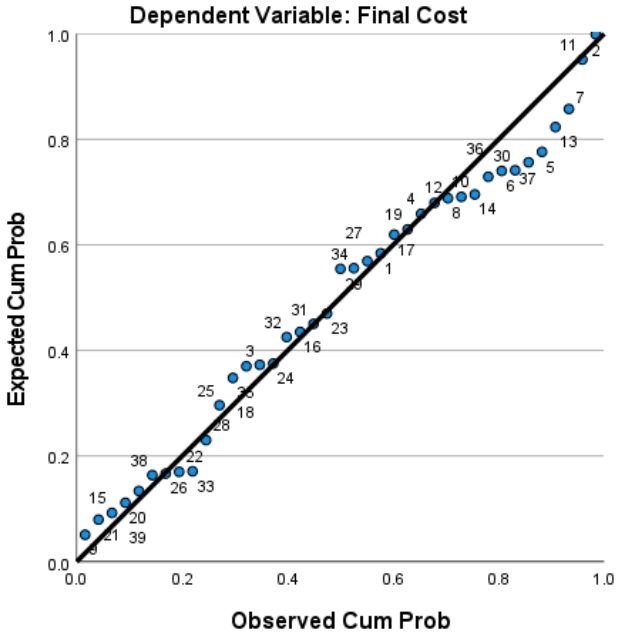

Figure 6.3. *Normal P-P plot of regression standardised residual*

6.8.5 Model validation

According to Ryan (2009), there are two ways to validate the model. One way is to split the data set in half and use one set for the model development and the other for validation. However, the author further pointed out that splitting the dataset may reduce the prediction efficiency. Therefore, Roecker (1991) suggested that the best approach to validating the model is to use a new dataset, thus ensuring the new data is in the same region as the old data. Therefore, this research used a new dataset to validate the model.

This study collected the data of sixteen projects for model validation. Therefore, forty-three cases were used for model development and sixteen for model validation. The ratio of cases was approximately 75: 25.

After that, the performances of the eight models were measured by the Mean Absolute Percentage Error (MAPE), which was calculated using the following equation (Eq. 2) (Kim et al., 2004).

$$\text{MAPE} = \frac{\sum \left| \frac{C_e - C_a}{C_a} \text{ x } 100 \right|}{n} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

C_e indicates the estimated cost through the regression model, C_a indicates the actual cost of the case, and n denotes the number of cases used for the validation. Table 6.6 shows the MAPE achieved for each model with respective adjusted R-squared values. MAPE was calculated separately for the negative and positive values of the C_e-C_a differences. Further, Table 6.6 also indicates the Mean Percentage Errors (MPE) of underestimating and overestimating errors. In a different view, the model's accuracy was also checked using the ratio between predicted cost (PC) and actual costs (AC). The mean ratios are also illustrated in Table 6.6.

Table 6.6. *Model validation results*

Model Number	MAPE	MPE of underestimating	MPE of overestimating	PC/ AC	Adjusted R-squared
1	22.36%	-18.57%	+26.15%	1.017	0.453
2	22.19%	-18.54%	+25.85%	1.003	0.485
3	22.11%	-18.86%	+25.35%	1.157	0.513
4	21.68%	-18.15%	+25.20%	1.001	0.536
5	21.29%	-18.17%	+24.41%	1.004	0.556
6	21.25%	-19.34%	+23.16%	0.999	0.593
7	21.35%	-17.83%	+24.87%	1.002	0.576
8	22.77%	-18.99%	+26.54%	1.001	0.578

According to Table 6.6, the lowest MAPE was from model 6, While the highest was from model 8. The lowest and highest MAPEs were reported as 21.25% and 22.77%, respectively. Therefore, the difference between the lowest and the highest was not significant. The lowest overestimating MPE was reported from model 6, whereas the highest was derived from model 8. On the other hand, the lowest underestimating MPE was reported by model 7.

Consequently, model 6 has the lowest ratio of predicted cost against actual cost. However, if the ratio exceeds 1, the model overestimates the cost. Conversely, a ratio below 1 indicates that the model underestimates the cost. Nonetheless, except for model 3, the rest of the models expressed a ratio relatively closer to 1. If the model achieved 1, then the model could predict precisely the same as the actual cost, which in reality is not possible.

The following section discusses the models developed in this study with findings similar to those in the literature.

6.9 Discussion

Based on the model fit test of this study (refer to Table 6.5), the models' predictability is around 68%. Therefore, approximately 32% is not captured by the model. Consequently, the independent variables selected for this model are not enough to explain much of the variation of the dependent variable. Other crucial variables could affect the project cost, which are not considered here. According to Atapattu et al. (2023), it is evident that there are significant risk factors affecting the project cost of road projects in NZ. However, due to the uncertainty of these factors, a regression model cannot fully capture such variables in the model. Therefore, the regression model can be improved by combining a modelling technique to capture the construction projects' uncertainty.

Over the years, much research has been carried out on regression analysis-based models developed for cost estimation of construction projects. However, this discussion only compares our results with the regression models developed for road projects.

One of the earliest research projects on regression models for road project cost was done by Ou and Swarthout (1986). The study was done using twenty-six projects from the Western United States. Researchers developed several unit price-based models and regression models for comparison. The models were developed using thirteen independent variables. The researchers observed that regression analysis could provide better estimates than the traditional unit-price method. Therefore, the regression technique can improve the current estimation practice for road projects.

Kim et al. (2008) developed a regression analysis-based model for cost estimation of highway projects with a MAPE of 13.28%. The model was based on twenty-seven highway projects completed between 1991 and 2001. Therefore, the model studies cover ten years, while our model is based on projects over twenty years.

However, Kim et al.'s (2008) models reported similar adjusted R-squared values compared to our models.

Mahamid (2011) developed ten regression models for road projects based on a hundred and thirty-one projects in Palestine. The model reported that the MAPE ranged from 13% to 31%. However, the researcher reported that the significant variables, such as road length and width, were excluded from the ten models with the highest accuracy. However, in our study, model 6, which reported the lowest MAPE, road length and width were considered. Nevertheless, Mahamid's (2011) models indicated higher adjusted R-squared values than our study.

The subsequent study was conducted by Cirilovic et al. (2014), which considered two hundred road projects covering fourteen countries in Europe and Asia. The researchers developed several models using multiple regression analysis and artificial neural network techniques to compare and find the better option. However, the model predictions are based on the variables related to the country's economy rather than the technical attributes of the project. Project duration and road length were the only variables that defined the project characteristics. However, in our models, all the independent variables can be used to define the project. Considering the background of New Zealand's economy, the variables considered in Cirilovic et al. (2014) models may not significantly impact road projects concerning the technical aspects of the project.

El-Maaty et al. (2017) conducted a similar study using highway projects. Their model showed an average percentage error of 30.42%. However, the model was developed to predict the cost overrun percentage instead of the project's final cost. Hence, the model cannot be used for conceptual cost estimation.

Another research was conducted in the same year by Zhang et al. (2017). Their research used absolute shrinkage and selection operator (LASSO) regression to compare with regular regression techniques. According to their findings, the ordinary least square regression model showed a

MAPE of 7.6%, while the LASSO regression model achieved a 7.1% MAPE. Therefore, both models had better performance compared to our study. However, the variables considered in Zhang et al. (2017) 's models cannot be used to anticipate the conceptual cost estimate. For instance, there were variables such as contract price, construction spending, prime loan rate, and weather days. Such factors are not easy to anticipate at the beginning of the project. Nonetheless, the models developed in our research considered the variables that can be anticipated at the project's early design stages.

Going forward, Dimitriou et al. (2018) compared the performance between regression analysis and artificial neural networks (ANN) using sixty-eight concrete bridges and roads along with the bridges. However, according to his findings, ANN reported better performance than regression analysis. ANN achieved an 11.48% MAPE, while the regression model achieved 12.29%. Nonetheless, the regression-based model performed better than the model developed in our analysis.

Similarly, Lin and Techapeeraparnich (2019) conducted similar research based on forty-four road projects in Thailand. Unlike most of the regression models discussed above, Lin and Techapeeraparnich (2019) took a similar approach to our study in finding the variables because the variables comprised in the model were road length and width, number of lanes, pavement type, earthworks, and miscellaneous work. The paper did not discuss the performance of the models in terms of error rates. However, in contrast to Dimitriou et al.'s (2018) findings, Lin and Techapeeraparnich (2019) observed that the regression model achieved better performance than the ANN-based model regarding R statistics. On the other hand, compared to the R statistics of our models, Lin and Techapeeraparnich (2019) accomplished better R statistics in their model.

The most recent study by Ahmed (2021) developed two regression-based models for roads and railways associated with tunnels. The model's accuracy was tested using the ratio of the predicted

cost to the actual cost. The mechanised tunnelling roads/ railways-based model showed 0.782, while the conventional road and railways showed 0.768 as the accuracy ratio. However, all eight models developed in this research (refer to Table 5) achieved ratios closer to 1 than Ahmed's (2021) study results.

Therefore, the other researchers developed better regression analysis-based models compared to the performance of the eight models of our study. However, those models were comprised of independent variables which could not be predictors at the beginning of the project. In contrast, the models created with similar variables did not perform better than those introduced in our research, except for Lin and Techapeeraparnich's (2019) model. Subsequently, our research developed a more reliable estimation model for New Zealand Road projects.

The following section concludes the research with recommendations and a discussion for further research.

6.10 Conclusions

This paper presents how multiple linear regression can be used as an analytical tool to develop a cost estimation model at the early design stage of road projects. Regression analysis is a standard statistical technique used for various analytical functions. Therefore, it is easy to train the professionals to utilise it for budget management during the project development.

Eight models were developed using the data from forty-three road projects in NZ. There were twelve independent variables, and the best sub-set was derived based on the backward elimination technique. The R-squared of the eight models ranged from 0.654 to 0.687, all of which were valid with *p*-values below the 0.05 alpha level.

Sixteen additional cases within the same range as the originals were used to validate the model. MAPE of the models were between 21.25% and 22.77%. Therefore, the differences among the

eight models were not significant. However, model 6 achieved the lowest MAPE, meaning the highest accuracy. Further, model 6 has the highest R-squared value and the lowest SE values. Therefore, it was the best-fit model. This model contained seven variables out of the twelve. They were road length, road width, the surface area of the bridges, pavement surface area, earthwork volume – cut, fill, and toppings, expected preliminary cost as a percentage of the total cost, possible cost indices change as a percentage to the base year (2022 was considered as the base year).

The models developed in this research still showed an error of approximately 21%-22%. The reason could be that the models considered only the technical independent variables, similar to the traditional estimation practice. However, the regression-based models can provide better estimates than the traditional estimation methods.

This paper contributes to the body of knowledge by studying regression analysis as a suitable modelling technique for conceptual cost estimation at the pre-design stage. Even though the estimation error is around 20%, the technique is easy to understand, and hence, industry experts can easily adapt to this new technique. Therefore, using this model, improving the current performance of the conceptual estimation methods in road projects is still possible. In the meantime, further investigation can be done to find other possible variables available at the pre-design stage, which could improve the model fitness and reduce the estimation error. Further, chapter 5 identified that ANN and SVM are good modelling techniques. Therefore, the same research can be done using those techniques to compare the performance and select the best model for the industry.

The model was limited to NZ road projects and did not consider qualitative construction risks. Nonetheless, the qualitative construction risks can still significantly impact the project cost. Thus, they should be incorporated into the estimation to further improve the regression model performance. Subsequently, combining the regression-based model with another analytical tool

that can address the impact of the qualitative risk factors better than regression analysis is recommended.

6.11 Epilogue

This chapter used the multiple regression analysis concept to develop a cost estimation model for the early design stage of road projects in NZ. Several models were developed using different combinations of the variables. Backward elimination was conducted to identify the best variable combination. Based on the outcome and the model's performance, recommendations were made for further research and further improvement of the model's performance and reliability.

7 Conceptual cost estimation model for the pre-design stage of road projects in New Zealand using artificial neural networks

7.1 Prologue

Based on the findings of Chapter 05, ANN was identified as an effective cost modelling technique often used in the construction industry for cost estimation. The regression model developed in Chapter 06 did not achieve the desired level of performance. Therefore, in this chapter, a model will be developed, tested, and validated using ANN using actual cost data from the NZ road projects (the same data set used in the regression analysis model). Subsequently, conclusions and recommendations are presented comparing the performance of the model and the reliability of ANN as a technique for cost estimation.[5]

[5] This chapter is based on the following journal paper (under review)

Atapattu, C.N., Domingo, N.D., and Sutrisna, M. (2023). A conceptual cost estimation model for the pre-design stage of road projects using artificial neural network. *Journal of Construction, Engineering and Management*. Manuscript ID: COENG-14398R1

7.2 Abstract

Providing robust cost estimates at the early design stage of road projects is often challenging due to the lack of available information. However, the conceptual estimate is crucial for the project's decision-making. Over the past twenty years, Road construction has faced approximately 20% cost overruns. This study used artificial neural networks (ANN) to develop a conceptual cost estimation model for road projects. Forty-three road projects from NZ were used to create the model, and sixteen cases were used for model validation. Twelve technical independent variables were selected based on previous literature, availability from the collected database, and availability of information during the pre-design stage. Six models were developed using different combinations of variables, the number of neurons in hidden layers, and activation functions. All the models achieved a Mean Absolute Percentage Error (MAPE) of less than 35.58%. The best model indicated a MAPE of 11.82% and an R-squared value of 0.877 for the validation phase. Subsequently, ANN can be used effectively to produce reliable conceptual cost estimates at the pre-design stage of road projects where less information is available.

7.3 Introduction

The conceptual stage estimate of the road project is a significant component in the planning and feasibility studies of the projects. Hence, the estimate’s accuracy is a fundamental requirement (Tijanic et al., 2020). The accuracy of the budgeting and estimation crucially depends on the reliability of the available information (Kim et al., 2004). However, less information is available on the project scope at the early design stage of the construction projects (Sodikov, 2005). Dimitriou et al. (2018) explained that the cost estimation accuracy increases when the project moves from conception to design, tendering, and construction. According to the Association for the Advancement of Cost Engineering (AACE), the conceptual cost estimate of a construction project would be expected to vary from the actual estimate. At the conceptual estimate stage, this

variation could be -20 % to -50 % of an underestimation range to +30% to +100% of an overestimation range (AACE, 2020). The unrealistic conceptual estimates can have a significant impact, mainly on the investment decisions of the project (Odeck, 2004). In addition, Dimitriou et al. (2018) further noted that the estimate could negatively influence design alternatives and the selection of the most efficient technical solutions.

Therefore, researchers suggested that the parametric cost model would be more ideal than traditional estimation methods at this stage (Hegazy and Ayed, 1998; Kim et al., 2004; Sodikov, 2005). Over the years, some researchers have identified several modelling techniques that can be used to predict the cost of construction projects, for example, regression analysis, neural networks, support vector machine, case-based reasoning, reference class forecasting, Monte Carlo simulation, and fuzzy logic (Shabniya and Dilruba, 2017).

A significantly greater number of cost estimation models were developed using the techniques mentioned above. Nevertheless, Mahalakshmi and Rajasekaran (2018) stressed that regression analysis is suitable for predicting a linear relationship. However, they further emphasised that the construction cost follows a non-linear relationship with its independent variables. Therefore, ANN is more suitable than regression analysis because of its predictability of non-linear relations.

Furthermore, many researchers confirmed the importance of the neural network in cost estimation over traditional approaches and other modelling techniques. For example, Adeli et al. (1998) emphasised that road and highway construction cost estimation crucially depends on human judgement, random market fluctuations and weather conditions. However, the researchers added that the neural networks could provide a solid analytical foundation for the estimate, resulting in reliable cost estimates. In another study, Kim et al. (2004) stated that neural networks could produce the best prediction with the highest accuracy compared to regression analysis and case-based reasoning. Nonetheless, the researcher further observed that case-based reasoning could

perform better in the long run. However, case-based reasoning performance crucially depends on the number of cases in the database because the accuracy will also increase when the number of projects in the database increases (Kim et al., 2004). A similar investigation was conducted by El-Kholy (2019), comparing neural networks against regression analysis and fuzzy logic. According to his results, the neural network-based model performed better than the other two techniques. At the same time, Roxas et al. (2019) also observed that ANN models show better generalisable capability and accuracy. Recently, Tijanic et al. (2020) emphasised that the neural network has the flexibility and tendency to predict and classify all sorts of data better than other techniques. In addition, the authors further observed that neural networks are more effective than traditional cost estimation methods in predicting the future, with a lack of information available.

A wide range of studies were conducted to study ANN as a tool for cost estimation (Adel et al., 2016; Cirilovic et al., 2014; Dimitriou et al., 2018; El-Kholy, 2019; Hegazy and Ayed, 1998; Jaafari et al., 2021; Mahalakshmi and Rajasekaran, 2018; Roxas et al., 2019; Sodikov, 2005; Tijanic et al., 2020; Xue et al., 2020). However, the contribution of such models in minimising the issues in conceptual estimation is minimal. Because the data or information required for the variables used in these models are not available until the design is done (Sodikov, 2005). Therefore, Flyvbjerg et al. (2002) noticed that there had been no improvement in the cost estimation of transportation infrastructure projects for the past seventy years. The statement was further tested and confirmed by a recent study based on one hundred and six New Zealand (NZ) road projects (Atapattu et al., 2023). According to the study, there has been no improvement in the conceptual cost estimation of NZ road projects for the past twenty years, and those projects are consistently facing approximately 20% cost overruns. Predominantly, the NZ industry reports stated that several significant investments are planned for road development for the next decade (Ministry of Transport NZ, 2020). Based on the literature, ANN-based models can improve the conceptual estimation and reduce the 20% above cost overrun. If an ANN-based model can be

developed with higher accuracy and less error, the model could predict the cost better than the current traditional models. Consequently, the difference between the conceptual estimate and the final cost at the project completion will be reduced. Therefore, to reduce the cost overrun, the performance of the conceptual estimation must be improved.

Further, the current ANN-based models identified in the literature contain several issues. Firstly, all the models considered a mix of technical, risk-based, and classification variables. When several classifications or risk factors are used, the number of cases used in model development should contain a substantial number of cases for each type in the classification variables (Ashtari et al., 2022). Further, ANN is not a technique for risk probability-based estimation (El-Kholy, 2019). Therefore, risk-related variables would reduce the model performance. Secondly, all the ANN-based models mentioned above contain variables that have less or no information during the early design stages to consider for estimation (Cirilovic et al., 2014; Dimitriou et al., 2018; El-Kholy, 2019; Hegazy and Ayed, 1998; Jaafari et al., 2021; Mahalakshmi and Rajasekaran, 2018; Roxas et al., 2019; Sodikov, 2005; Tijanic et al., 2020; Xue et al., 2020). Subsequently, considering the strengths and weaknesses of the current ANN-based models, this research intends to develop a cost estimation model for road projects to minimise the consistent cost overrun faced over the past twenty years and improve the reliability of early-design stage cost estimation. Because the current ANN-based models either used variables that cannot be quantified or extracted data during the conceptual stage or contained risk/uncertainty-related variables. Therefore, such models showed low performance. However, the intended model of this paper considers only the variables suitable for the conceptual stage. The following section discusses the basics of artificial neural networks.

7.4 Artificial neural networks

The artificial neural network has simply adopted the architecture and functions of human brain cells and the nervous system that learn from previous experience and respond to complex problems

(Tijanic et al., 2020). Silva et al. (2017) state that artificial neural networks can be used in universal functional approximation, pattern recognition, process identification and control, predictions, and system optimisation.

ANN models comprise three primary layers: input, hidden and output. The input layer receives the information/ data into the model. The number of neurons in this layer depends on the number of independent variables selected for the model. Each independent variable represents one neuron in the input layer. Thenceforth, the centre part of the model contains the hidden layer. An ANN model can have one or more hidden layers depending on the model's architecture. This layer is used to extract patterns associated with the process or system that is being analysed (Silva et al., 2017). The number of neurons in the hidden layer can be decided based on the model's performance. Additionally, the model can be re-run, adjusting the number of neurons in the hidden layer. Then the model with the highest performance will comprise the ideal number of neurons. Ultimately, the output layer is composed of the neurons that produce the result or the outcome based on the analysis and calculations performed in the model. The number of neurons is based on the model's expected output. Usually, the output layer consists of one neuron because most estimation models are developed to forecast the project's final cost.

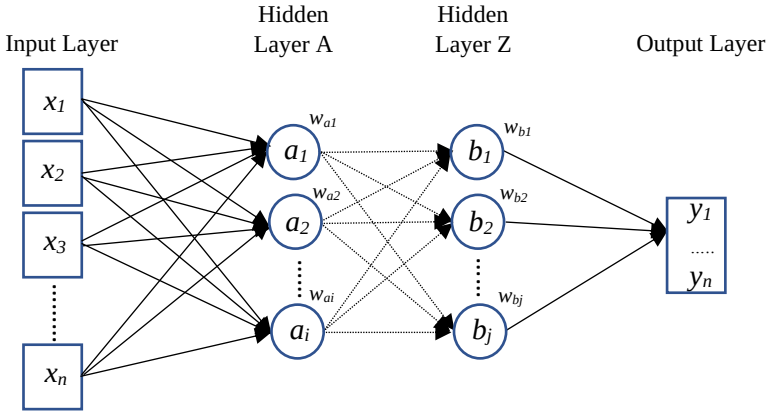

Figure 7.2. *ANN architecture*

The multiple-layer feedforward method is the most used architecture in ANN models because the single-layer feedforward models do not contain any hidden layers (Silva et al., 2017). Figure 1

shows a sample multiple-layer feedforward network with *n* number of independent variables (*x*), and two hidden layers with *i* and *j* numbers of neurons, respectively. "w_i" and "w_j" are the weights of the neurons. The output layer has only one neuron representing the respective output of the problem being analysed by the model. The Multiple Perceptron (MLP) and the Radial Basis Function (RBF) are the most widely recognised multiple-layer feedforward architectures (Ahiaga-Dagbui and Smith 2012; Silva et al. 2017; Tijanic et al. 2020). These two models utilise the generalised delta and competitive/delta rules as the learning algorithm in the training phase. However, Ahiaga-Dagbui and Smith (2012) emphasised that the MLP model is superior to the RBF model. This is because the MLP model focuses on identifying the relationships between the inputs and the output; on the other hand, the RBF model performs in two stages. The first stage serves a probability distribution of the inputs and identifies the relationships in the second stage. Therefore, this study only focuses on developing the ANN model adopting the MLP principles.

The neural networks' configuration is based on identifying network topology, the backpropagation learning rule, transfer function and learning control function (Adel et al., 2016). The backpropagation learning algorithm is also known as the generalised delta rule (Silva et al., 2017), performed in two stages, as shown in Figure 1. The first stage is known as forward propagation. At this stage, the model uses the training data to calculate the output using the neurons in each layer. In the next stage, the output of the first stage is compared with the desired output. Subsequently, the weights are modified based on the variance, and this method is called backward propagation. Therefore, successive application of the forward and backwards propagation allows adjusting the weights and thresholds of the neurons to adjust automatically in each iteration and gradually reduce the sum of the errors (Silva et al., 2017).

Jafarzadeh et al. (2014) emphasised that the backpropagation algorithm is more efficient than other learning algorithms in problem-solving. This idea was further confirmed by Emsley et al. (2002) by evaluating different algorithms. Their evaluation considered backpropagation, delta-bar-delta,

quasi-Newton training algorithm, conjugate gradient descent, and quick propagation. Although the researchers identified several network training functions, the most popular algorithm is the Levenberg-Marquardt (Adel et al., 2016; Marinelli et al., 2015) because it is one of the fastest backpropagation algorithms (Adel et al., 2016; Jafarzadeh et al., 2014). The training function is used to update the weights and bias nodes.

Model activation functions are used to process and interpret the data inside the neurons of the hidden and output layers. According to Adel et al. (2016), using one or more functions in one model is possible. The activation function aims to receive the output given by the summation function as an input and convert it into the final output of a node (Alaloul and Qureshi, 2019). Although there are several activation functions, non-linear functions are the most used as the linear activation function has limited power and cannot be used in complex data (Alaloul and Qureshi, 2019). Tanh and Sigmoid are the most common model activation functions (Adel et al., 2016). Therefore, the aforementioned activation functions will test in this study to develop the model.

Based on the above-discussed concepts of ANN, this study aims to develop a prediction model for road project cost estimation using the cost data from NZ. The following section reviews the ANN-based models produced by other researchers.

7.5 Literatuyre review

One of the earliest models developed using ANN was the study conducted by Hegazy and Ayed (1998). They used 18 highway projects in Canada as the data source and considered three different approaches in the network configuration. The first approach was the backpropagation training, which produced a weighted error of 10.4%. The second model was developed using simplex optimisation, producing a 1% weighted error result. Finally, they used genetic algorithms optimisation, but the weighted error was 21.8%, which is significantly higher than the other two

approaches. Although the model achieved good results, it was developed over twenty years ago. Therefore, it must be retested before using the model in today's industry.

Sodikov (2005) conducted similar research on road projects in Poland and Thailand and developed ANN-based cost models. However, the MAPEs of these two models were significantly high. Elbeltangi et al. (2014) developed an ANN model for highway cost estimation with a MAPE of 6.21% based on sixty-seven projects in Libya using eleven parameters. During the same time, Cirilovic et al. (2014) developed a cost model using a completely different approach. In their model, most variables were external factors affecting the road projects. Only a few variables were project-specific. However, the best model achieved an R-squared of 0.7469.

Further, Adel et al. (2016) developed a conceptual cost estimate model for highway projects using a database of seventy-five highway projects. Initially, they developed twelve models changing the number of neurons in the hidden layer. However, eleven models were rejected after the training, validation, and testing phases. Only one model was selected as a reliable model, which achieved MAPE of less than 20% in all three phases.

Mahalakshmi and Rajasekaran (2018) developed an ANN model using MATLAB software for highway projects in India. Their model was able to predict the outcome with a MAPE of 8.46%. However, Mahalakshmi and Rajasekaran's (2018) model considered topography, pavement type, soil condition, and drains. However, the information for these variables is most unlikely to be available during the conceptual estimation stage.

In another research, El-Kholy (2019) adopted modular neural networks, generalised regression networks, and time-lag recurrent networks to develop three different ANN models. However, all three models showed MAPE higher than 25%. It was highlighted in these models that out of the fifteen variables used, most of them are qualitative and risk-related factors. It further emphasised

that ANN cannot ascertain the project risks alone when risk factors are incorporated into the model as independent variables.

The research conducted by Maharie and Shaik (2020) emphasised that both ANN and support vector machines showed lower performance compared to regression-based models. Similarly, Tijanic et al. (2020) developed an ANN-based model using fifty-seven road project data. The research experimented with several types of ANN models and concluded that the general regression neural network-based model achieved the lowest error percentage.

The four models Xue et al. (2020) developed comprised two different components. Two sets of models considered the factors relating to bridges and tunnels and the general variables. In contrast, the other two models did not consider the variables of bridges and tunnels. However, the lowest error rates within these four models were the models with bridge and tunnel variables.

Jaafari et al. (2021) recently conducted a study on developing a cost model for road construction based on machine learning techniques. They adopted techniques such as linear regression, ANN, instance-based learning, K-star, and support vector machine in developing the models to compare the methods. According to their results, the ANN-based model also achieved good results, with a percentage of relative error index of 0.038% for the testing phase. Compared to most of the models developed for road projects, this model achieved an outstanding performance. However, this model considered variables such as the number of trees, size of trees, cut and fill slopes, drainage, and culvert details. These variables are most unlikely to be available during the project's conceptual stage.

Table 7.1 summarises the performance of the models identified in the literature compared to our model.

Table 7.1. *Summary of literature on similar ANN models developed for road projects*

Model	Performance (MAPE or R-squared or RMSE)	Database
Adel et al. 2016	Training – 4.51%; Evaluation – 5.8%; Validation – 16%	75 road projects
Cirilovic et al. (2014)	R-squared – 0.7469	200 road projects
Dimitriou et al. (2018)	Piers - 34%; Precast beams – 11.48%; Cast-in-situ – 13.94%; Cantilever – 16.12%	646 concrete bridges
El-Kholy (2019)	Modular neural network - 25.4%; Generalised regression network - 37.64%; Time-lag recurrent network - 29.43%	56 road projects
Hegazy and Ayed (1998)	Backpropagation – 10.4%; Simplex optimisation – 1%; Genetic algorithm – 21.8%	18 road projects
Jaafari et al. (2021)	0.038%	4811 road projects
Mahalakshmi and Rajasekaran (2018)	8.46%	52 road projects
Maharie and Shaik (2020)	Root Mean Squared Error (RMSE) – 1.18	74 road projects
Roxas et al. (2019)	Average 20% error	41 road projects
Sodikov (2005)	Poland – 24%; Thailand – 26%	38 in Poland 42 in Thailand
Tijanic et al. (2020)	Training – 10.22%; Validation – 13.06%	57 road projects
Xue et al. (2020)	With bridge and tunnel data CNN* model - 17%; ANN model - 22.84% Without bridge and tunnel data CNN* model – 23.78%; ANN model – 40.62%	415 road projects

*Legend - *CNN – Convolutional Neural Network*

The following section discusses the research methodology adopted in this study.

7.6 Research methodology

The cost data were collected from sixty-six projects in NZ completed between 2002 and 2022. National-level government agencies were approached to collect the data for the research. The scope considered for the project was national-level road projects with the scope of new construction and significant alteration of roads and bridges associated with the roads. Once the projects were gathered, the outliers were identified and eliminated using Mahalanobis distance measures because, according to Tabachnick and Fidell (2001), if the *p*-value of the Mahalanobis distance is less than 0.001 alpha level, then that case is an outlier. The outliers will show significant deviation from the other cases regarding scope, budget, and other independent variables. Having outliers within the database can reduce the expected performance of the model. Therefore, seven cases were eliminated from the database, and only fifty-nine cases were considered in model development. Thus, the scope, budgets and variables data were in a similar range for model development and validation cases. Of these fifty-nine cases, 43 cases were randomly selected by

the modelling software for the model development, while the other sixteen were retained for statistical validation of the model.

Since New Zealand does not have a considerable number of large-scale new road construction projects, it was challenging to gather data. In addition to that, the data collection was carried out during the COVID-19 pandemic and was severely impacted due to the restrictions on meeting personnel. Once the cases were gathered, all the possible variables were extracted separately. Initially, there were fifteen variables. After that, a multicollinearity test was conducted to examine the suitability of the variables because if multicollinearity is present, variables may not predict the output as expected. In addition, it can also lessen the impact of significant variables. Consequently, the selected variables were considered as the neurons of the input layer.

MATLAB software was adopted to create the neural network model. Using the same variables, several models were tested to identify the models with significant performance. The activation function, number of variables, and number of neurons in hidden layers were adjusted to see the models' capability. Initially, twenty models were developed, out of which only six models were considered for further analysis. The disregarded models showed significantly lower R-squared values that were less than 0.6. However, the selected models showed better R-squared values, as discussed in the model development section. Once the models were created, the validation was done using the retained sixteen cases. The following section presents the process of ANN model development using the adopted methodology described above, as well as the model's results.

7.7 Model development and discussion

7.7.1 Variable selection

Table 7.2 summarises the variables used in the current literature that developed similar models in other countries. Based on the findings in Table 1, fifteen independent variables were identified for

this research based on several criteria. The main criterion was to determine the variables that have information available to feed the model during the conceptual design stage because this research is focused on developing a model for the early design stage of the project. In the second step, the variables with data available from the selected cases used for the model development were further considered. The third criterion was to evaluate the variables that were considered by most authors in the current literature because if the same variable is used in several models, that variable could be significant to the project cost. However, once the model is developed, the significance of each variable about the project cost will be tested and verified. All the non-significant variables will be disregarded from the model. The fifteen selected variables fulfilled all the three requirements above.

Further, the factors selected from the literature were validated by the experts who met during the data collection to understand the availability and suitability during the conceptual estimate. According to industry experts, this study did not consider variables such as contract type, procurement type, tendering strategy, project scope, project type, and road classification. Because, if those variables were to be considered, then the data should be available for each type of classification, type, and scope. For instance, if the procurement type is a variable, then the model development cases should contain enough data to cover all the procurement strategies used in the road projects. Otherwise, the model outcome may not match the expected project cost. Further, the case number under each category above must be higher to develop a reliable model. Therefore, this study did not consider the classification variables.

Table 7.2 *Variables identified through literature*

Variables considered in this study	Availability during pre-design stage	Adel et al. (2016)	Cirilovic et al. (2014)	Dimitriou et al. (2018)	Hegazy and Ayed (1998)	Jaafari et al. (2021)	Mahalakshmi and Rajasekaran (2018)	Maharie and Shaik (2020)	Roxas et al. (2019)	Sodikov (2005)	Tijanic et al. (2020)	Xue et al. (2020)
Road length	√	√	√	√	√			√	√		√	√
Road width	√	√				√	√				√	
Distance from the nearest major city	√	√	√		√			√	√			√
The number of bridges	√		√	√				√				
The approximate length of retaining walls	√	√	√	√	√			√	√		√	√
Ground improvements area	√					√			√			
Pavement area	√						√		√	√		
Cut and fill area	√					√			√	√		
Project duration	√	√			√		√	√	√	√	√	√
Expected year of completion	√	√			√		√	√	√	√	√	√
Expected preliminary cost (% of the total cost)	√	√			√		√			√		√
Expected % change in the construction cost index	√		√					√	√			√
Kerbs length	X	√	√	√	√			√	√		√	√
Road markings	X	√	√	√	√			√	√		√	√
Base course volume	X	√	√	√	√			√	√		√	√

In addition, during the data collection, industry experts involved in those completed projects were inquired regarding the availability of the information for the selected variables during the pre-design stage. As mentioned in Table 1, except for the last three variables, the other twelve variables should be able to be quantified approximately by an experienced estimator during the conceptual estimation. Some of the experts noted that it is possible to quantify the last three variables based on the previous experience and cost analysis of previously completed projects. Therefore, none of the variables were eliminated at this stage and were kept for further analysis for the model's suitability.

Observing the significance of these variables is required to predict the projects' final cost. Therefore, the ANOVA test, as shown in Table 2, was conducted to examine the significance of the variables. If the *p*-value is more significant than 0.1, the variable would not significantly impact the project cost. Further, the identification and elimination of variables with multicollinearity are crucial. If the independent variables are highly correlated, then the dependent variable prediction may not be reliable. Therefore, the data was fed into SPSS to check the variance inflation factors (VIF). Young (2017) observed that the variable indicates multicollinearity if the variance inflation factors (VIF) are higher than five. Therefore, kerb length, road markings, and base course volume

were eliminated for two reasons. The first reason was that the VIFs of those variables were higher than five. That means there would be a cause of multicollinearity. The second reason is that those three variables did not significantly contribute to the project's final cost based on the *p*-values. Further, the "road length" variable did not show a considerable p-value when these three variables were present. However, "road length" was an essential variable to the road project cost. Once the three variables above were eliminated, the "road length" variable showed a very substantial *p*-value, as shown in Table 7.3. Accordingly, Table 7.3 shows the twelve selected variables for the ANN model.

Table 7.3. *Variables considered in the model*

Variable	Unit	VIF	*t*-test *p*-value	Model group 1	Model group 2
Road length	m	1.982	0.002	√	√
Road width	m	1.463	0.710	√	-
Distance from the nearest major city	m	1.362	0.899	√	-
The surface area of the bridges	m^2	1.490	0.021	√	√
The surface area of the retaining walls	m^2	1.588	0.791	√	-
Ground improvements area	m^2	1.775	0.746	√	-
Pavement area	m^2	4.746	0.053	√	√
Cut and fill area	m^3	3.004	0.192	√	-
Project duration	Days	1.248	0.881	√	-
Expected year of completion	Year	3.333	0.774	√	-
Expected preliminary cost (% of the total cost)	%	2.088	0.049	√	√
Expected % change in the construction cost index*	%	2.021	0.088	√	√
Excluded variables					
Kerbs length	m		-	-	-
Road markings	m		-	-	-
Base course volume	m3		-	-	-

7.7.2 Model development

According to Adel et al. (2016), the ANN training phase comprises an iterative process involving submitting training segment data to the network to conclude the relationship between the inputs and outputs. Once the models' variables were defined, the data were uploaded to MATLAB software to generate the model. Initially, all the variables were tested with the neural networks. Then, based on the *p*-values of the *F*-test using the analysis of variance (ANOVA) test, some variables were eliminated to identify the best variables group. In Table 7.3, Model Group 1 considered all the variables, while Model Group 2 considered only variables with significant *p*-

values for the ANOVA *F*-test. Based on the variables shown in Table 7.3, the ANN architecture developed for each model is shown in Table 7.4.

The architecture of the models can be written simply based on the number of neurons in each node. For instance, the ANN model architecture of this study consisted of an input layer, two hidden layers and an output layer. The input layer contained twelve neurons, one for each independent variable. Each hidden layer had '*i'* and '*j'* numbers of neurons in each layer, respectively. The output layer has only one neuron since the desired output was only the project's cost. Hence, the ANN architecture of this research can be identified as 12-*i*-*j*-1.

In MATLAB, Levenberg-Marquardt (trainlm) is the ideal backpropagation algorithm for ANN modelling. The model adopted two activation functions, which are Tanh and Sigmoid. Table 2 illustrates the activation functions used in each model for the hidden and output layers. Although several other combinations of the functions were tested, the table shows only the results of the successful models.

Developing an ANN model has three primary phases: training, validation, and testing. Generally, MATLAB randomly chooses the cases for each phase out of the total cases uploaded for the model. The first phase of neural network modelling is the training phase. By default, 70% of the cases are randomly selected for the model training. Some models use the same dataset for training, validation, and testing by dividing the cases into a 70%: 15%: 15% ratio. However, this study decided to validate the results using the cost data from sixteen instances. Therefore, the testing data set was 30% of the initial cases.

To check the reliability of the results, a sensitivity analysis was conducted, identifying the best model architecture that can predict the output with the slightest error. Therefore, the number of hidden layers, number of neurons in the hidden layers, and type of transfer function were varied

to determine the best ANN model to predict the cost estimate. Table 7.4 consists of only the models with the highest reliability compared to the other combinations.

R-squared and root mean squared error (RMSE) were chosen to measure the performance of the models. Table 7.4 illustrates the six highest-performing models among the models developed, including the performance measures at the training and testing phases.

Table 7.4. *Model training and testing results*

Model number	Number of variables	Number of nodes in the hidden layer		Activation function		Training Phase		Testing Phase	
		Layer 1	Layer 2	Hidden layer	Output layer	RMSE	R^2	RMSE	R^2
1	12	3	0	Tanh	Tanh	1.465	0.895	0.932	0.856
2	12	1	1	Sigmoid	Sigmoid	0.279	0.977	0.245	0.993
3	12	1	2	Sigmoid	Sigmoid	0.395	0.992	0.274	0.983
4	5	2	2	Sigmoid	Sigmoid	0.434	0.991	0.219	0.989
5	5	1	0	Sigmoid	Sigmoid	0.453	0.989	0.277	0.995
6	5	1	1	Tanh	Tanh	1.872	0.935	0.675	0.493

R-squared of all six models during the training phase showed a good model fit. However, models 1 (12-3-0-1) and 6 (5-1-1-1) indicated RMSE greater than 1 with the two lowest R-squared values. In the testing phase, the highest RMSE values were obtained from models 1 and 6. Moreover, those two models showed the two lowest R-squared values in a similar range to the testing phase. Further, the R-squared value of model 6 at the training phase is significantly lower than the other five models. Therefore, the Sigmoid function is performing better than the Tanh as the activation function for this dataset. Models 2 (12-1-1-1), 3 (12-1-2-1), 4 (5-2-2-1), and 5 (5-1-0-1) achieved performance very close to each other.

All the above models contain a maximum of three neurons in each layer. Models 1 and 5 have only one hidden layer, while the other four have two hidden layers in each model. The models' performances were checked up to twenty neurons in the hidden layer. It was noticed that the performance of the model decreased when the number of neurons increased. The model validation considered all six models above to determine the best model.

7.7.3 Model validation

Although model development consisted of model training and testing phases, evaluating validity using a new set of data not used in the development is ideal. Therefore, a separate analysis was carried out to test the data validity of the model using MATLAB. To evaluate the models' performance, Adel et al. (2016); El-Kholy (2019) and Kim et al. (2004) suggested that Mean Absolute Percentage Error (MAPE) is the most widely used formula. MAPE can be calculated using equation (1), as shown below.

$$\text{MAPE} = \frac{1}{n} \sum \left| \frac{C_e - C_a}{C_a} \text{ x } 100 \right| \text{ ……(EQ1)}$$

'C_e' indicates the estimated cost through the ANN model, 'C_a' signifies the actual cost of the case, and 'n' denotes the number of cases used for the validation. Table 7.5 illustrates the MAPE achieved by the ANN models developed in this research.

Table 7.5. *Model validation results*

Model Number	Number of independent variables	Number of neurons in the hidden layer		MAPE	R^2
		Layer 1	Layer 2		
1	12	3	0	35.58%	0.764
2	12	1	1	11.82%	0.877
3	12	1	2	16.08%	0.794
4	5	2	2	22.90%	0.785
5	5	1	0	17.15%	0.740
6	5	1	1	26.72%	0.736

According to Table 7.4, models 1 and 6 did not perform well in the training and testing phases due to high RMSE values. However, Table 7.5 shows that the above two models achieved comparatively reasonable MAPEs for the validation phase, 35.58% and 26.72%, respectively. In addition, the R-squared of models 1 and 6 were within the same range as the others. Further, Table 7.6 shows all the error results of all sixteen test cases for Model No. 02.

The output of Table 7.6 is further illustrated in Figure 7.3. Apart from seven cases, in all other nine cases, ANN predicted costs were in solid coherence with the actual cost of the projects. However, overall MAPE is 11.82%. In addition, Table 7.6 also shows the project's original estimate. The

initial estimates of all 16 projects showed 22.56% of overall MAPE. Consequently, compared to the original estimate, the ANN model developed in this study achieved much better results. Therefore, this ANN model can minimise the cost overruns by 10% to 11%. The following section discusses the model's limitations.

Table 7.6. *Validation results for Model No 02*

	Actual Cost (NZ$)	Traditional conceptual estimation		ANN Model No 02	
		Original Estimate (NZ$)	Absolute Percentage Error (APE) (%)	Predicted Cost (NZ$)	Absolute Percentage Error (APE) (%)
1	5,379,458.36	3,674,114.50	31.70%	6,899,887.60	28.26%
2	5,469,708.59	3,074,952.50	43.78%	5,843,324.38	6.83%
3	6,183,854.43	4,963,023.66	19.74%	6,589,217.26	6.56%
4	6,761,382.89	4,889,405.00	27.69%	5,083,693.28	24.81%
5	6,909,174.34	4,193,413.52	39.31%	5,477,487.87	20.72%
6	6,964,095.39	5,964,031.00	14.36%	6,641,822.06	4.63%
7	7,075,124.51	6,186,372.10	12.56%	5,082,925.09	28.16%
8	7,363,881.68	7,015,700.00	31.89%	5,499,259.58	25.32%
9	8,111,879.42	6,735,269.36	16.97%	7,556,502.07	6.85%
10	9,351,966.35	8,343,894.36	10.78%	8,994,509.72	3.82%
11	9,470,985.65	7,053,766.50	25.52%	8,044,963.22	15.06%
12	11,000,688.37	7,125,064.79	35.23%	11,331,302.27	3.01%
13	11,911,646.74	10,590,454.15	11.09%	11,251,509.20	5.54%
14	13,028,819.73	11,827,677.57	9.22%	12,815,319.43	1.64%
15	16,068,756.72	13,726,546.53	14.58%	16,946,004.03	5.46%
16	17,092,831.00	14,256,081.00	16.60%	16,678,525.04	2.42%
MAPE (%)			**22.56%**		**11.82%**

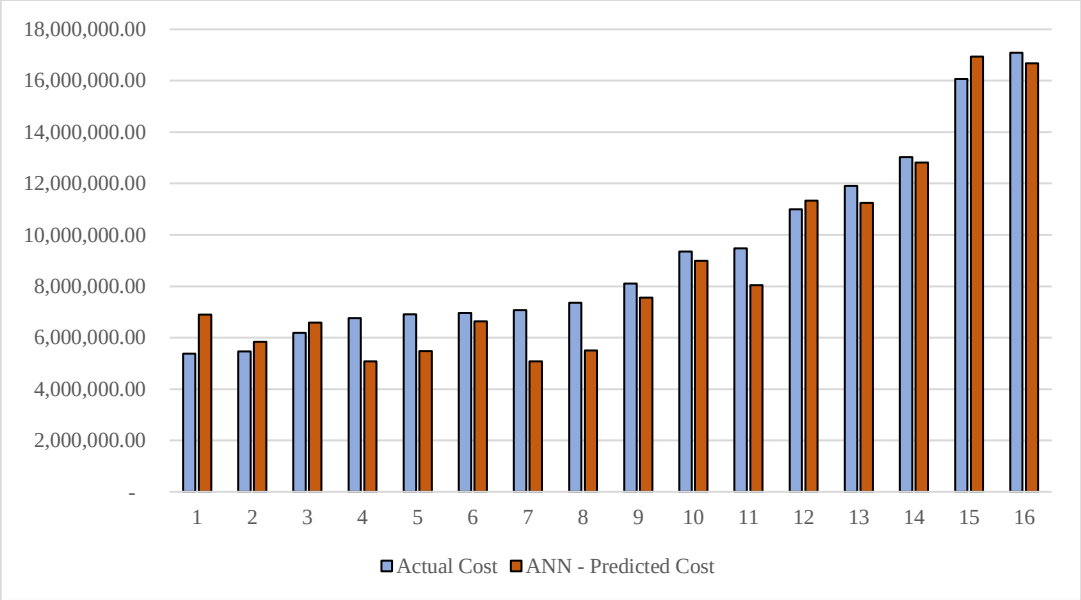

Figure 7.3. *Actual cost versus predicted cost (test data)*

The following section discusses the limitations of the ANN model developed in this study.

7.7.4 Model limitations

Based on the literature review, it is observed that the model performance can be improved when the number of projects in the cost database is higher. This is because when more cases are used to develop the model, the reliability is higher as there is more evidence to support the modelling equation. Our model is based on forty-three cases. Further, the model development did not consider variables such as procurement strategy, tendering, contract types, and road classifications. This is because if such classification variables are to be believed, there should be enough cases for each type of classification to run the model successfully. Otherwise, the model tends to generate unreliable outputs. However, in this study, within the forty-three cases, the majority of the cases belong to one category of procurement, tendering, contract and road classification. Therefore, those variables were not considered.

In contrast, Cirilovic et al. (2014) developed a model considering several variables that are not project-specific but are mainly external factors. However, these external factors are risk-related variables and often unpredictable based on the previous project data. The ANN model developed in our research did not incorporate any external factors or any other risk factors. A probability-based analysis should be done for a probability-based estimation and to produce a reliable estimate. Furthermore, our model showed an R-squared value of 0.877 and a MAPE of 11.82%. The major weakness in this model is the uncertainty of the project. Therefore, by incorporating the risk estimation method, the performance of this model may be improved.

Another reason for having a performance of 0.877 (R-squared) is that there may be variables with significant impact on the project cost that are not considered in the model. For example, as previously discussed, classification variables, such as road classification, procurement, tendering and contract types, and road layer materials, were not considered due to the smaller amount of project data available for each type. Subsequently, the model was developed using forty-three

samples, and an additional sixteen samples were used for model validation. However, using a minimum of a hundred samples is recommended to achieve better results. The model can perform better with a larger sample to apply to the proposed project forecasting.

The Table summarises the scope of all the variables as a recommendation for future road project estimation. The forthcoming project must be within the scope and limitations mentioned in Table 6 to apply the model.

Table 7.7. *Scope of the model and the variables*

Description	Scope and limitations
Independent variables	
Road length	Measured in meters Total length of the road
Road width	Measured in meters Total width of the road without the pavement
Distance from the nearest major city	Measured in meters Approximate distance to the site from the nearest major city where most materials, plant, labour, sub-contractors and suppliers are available.
The number of bridges	It is measured in numbers. Reinforced concrete bridges with The bridge span is the same as the span of the typical road.
The approximate length of retaining walls	It is measured in meters. Reinforced concrete wall
Ground improvements area	Measured in square meters Total project surface area (including roads, pavements, and other ground areas subjected to the changes by the project)
Pavement area	Measured in square meters The surface area of the side pavements, including walkways
Cut and fill area	Measured in square meters The surface area covers the total cutting and filling area. (The volume cannot be calculated at the conceptual stage without a proper ground investigation and a survey)
Project duration	Measured in Calendar Days As mentioned in the Contract Project construction start date to the practical completion date.
Expected year of completion	Based on the contractual programme, the intended year of practical completion
Expected preliminary cost (% of the total cost)	Expected preliminary cost as a percentage of the total project cost
Expected % change in the construction cost index	Expected change in the construction cost index as a percentage. 2002 was considered as the base year.
Road layers	
Main alignment	High-quality compacted earth-fill subgrade 1m height + 180 mm Sub-base + 190 mm base-course + Chip seal surface. Pavement layers should be quarried rock + 1.5% cement. To the NZTA specifications

The following section concludes the findings of this research and discusses the recommendations for further research.

7.8 Conclusions

A conceptual estimate of the construction project is crucial in making decisions. Therefore, it is vital to produce reliable and accurate estimates at the early design stage when less information is available compared to the other phases of the construction projects. Consequently, this study adopted the ANN technique to develop a conceptual cost estimation model for road projects. ANN provide accurate and reliable predictions compared to other modelling techniques. Six ANN-based models were developed with significant performance compared to the other models. Model no. 02 indicated the best results compared to the other five models. The MAPE of model no. 02 was 11.82% with 0.877 of R-squared.

A key takeaway from this study is that technical variables alone cannot produce accurate conceptual estimates due to the high level of uncertainty at the pre-design stage. However, without risk consideration, the ANN-based model developed in this study was able to calculate the conceptual estimate within approximately 12% error. Considering that the current NZ road projects face cost overruns of 20% (NZTA, 2022), this study can improve the estimation practice and reduce the cost overruns from 20% to 11.82%. All the literature identified in this paper that considered ANN-based models did not apply to the conceptual stage because those models contained variables that could not be calculated using the information available at the conceptual stage. Conversely, this study identified and recommended twelve independent variables that can be used for conceptual estimates. Even though this study focused on NZ road projects, the same model can be applied and tested in other contexts to improve the conceptual estimate because no NZ-specific variables were used in this model. The project location was the only variable that could affect the geographical area. However, in this study, geographical location was changed to

"distance from the nearest major city". Therefore, the model does not consider the country but how far the project is from the nearest major city where all the materials, labour, and plants are available. That affects the project cost more than the geographical location. Therefore, the model developed in this study can be generalised to other contexts.

The ANN was observed as a technique unsuitable for predicting the project's risk factors. Therefore, it is recommended to combine the outcome of this study with another method that can predict the risk factor to improve the current model performance. Another way to improve the performance is to increase the number of cases used to develop the model. When the database's reliability increases, the models' performance will also increase. Due to the level of availability and accessibility to the project information, the twelve variables considered were the maximum variables that could be obtained in comparison with literature findings, data availability, and expert opinions. However, there may be other variables that can make a significant impact on the project cost but are not incorporated here. By considering these, the model performance may further improve.

Subsequently, the models are developed using the cost data from NZ road projects comprising newly constructed and significant alterations. Therefore, research in other countries or different sectors, such as buildings or different scopes, such as maintenance of roads, can adapt the concept and the development process.

7.9 Epilogue

This chapter used the artificial neural network concept to develop a cost estimation model for the early design stage of road projects in NZ. Several models were developed using different combinations of the variables, number of hidden layers, and number of neurons in the hidden layers. Six models were developed with significant performance. Three stages of model development were carried out: model training, testing, and validation to ensure the model fits the

purpose. Based on the outcome and the model's performance, recommendations were made for further research and further improvement of the model's performance and reliability.

8 Final conceptual cost estimation model for the pre-design stage of road projects in New Zealand

8.1 Prologue

Road construction projects are crucial to infrastructure development, economic growth, and societal progress. Therefore, having a reliable budget at the beginning of the project enables the project funders and clients to make appropriate decisions. Over the years, various approaches have been proposed and implemented for estimating the cost of road construction projects. However, based on the findings from Chapter 4, road projects experience significant continuous cost overruns. As a solution to this, chapters 6 and 7 developed two cost models using regression analysis (RA) and artificial neural network (ANN) techniques. Although the ANN-based model minimised the cost overruns by approximately 8%, there should be ways to minimise the cost overruns further. Therefore, this chapter develops the final conceptual cost model by combining two techniques that can complement each other's disadvantages to improve the model performance.[6]

[6] This chapter is based on the following journal paper (manuscript is ready for submission)

Atapattu, C.N., Domingo, N.D., and Sutrisna, M. (2023). A conceptual cost estimation model for pre-design stage of road projects – A hybrid of artificial neural networks and Monte Carlo simulation, *Sustainable Cities and Society*

8.2 Abstract

The successful planning and execution of construction projects depend significantly on accurate cost estimation, as cost overruns can lead to budgetary constraints, delays, and compromised project quality. Therefore, the accuracy and reliability of conceptual estimates need to be improved. This study aims to develop a cost model usable at the concept stage of road projects. Forty-three cases from New Zealand (NZ) road construction were used to develop the model, and sixteen more were used for model validation. The study combines ANN with MCS to develop a risk-based estimation model. The final model achieved 3.53% of the mean absolute percentage error. However, several cost models have been developed, and only a few considered risk-based estimation. However, such models also cannot apply to the pre-design stage because the variables need in-depth information. Notably, the model of this study applies to other countries as it does not contain NZ-specific variables. The proposed model effectively determines a reliable pre-design stage cost estimate for road projects with little information.

8.3 Introduction

The world is evolving towards the digital age, called the fourth industrial revolution or Industry 4.0. Like every other sector, the construction and built environment constantly evolves and embraces new technologies (Ashtari et al., 2022). However, there are issues such as cost overruns without a proper solution or improvement over the past seventy years (Flyvbjerg et al., 2003). Providing a reliable cost estimate for road projects at the early design stages is often challenging, considering the less information available, project duration, and geographical spread compared to vertical construction (Xenidis and Stavrakas, 2013). According to ASCE (2020), the uncertainty at the project's conceptual stage is significantly high. Therefore, researchers have investigated different techniques to improve the current estimation practice. As noted in Chapter 5, the most reliable statistical techniques for developing the cost models were regression analysis (RA),

artificial neural network (ANN), and support vector machine (SVM). Further, the chapter explained that combining several techniques into a hybrid model can overcome the disadvantages of one technique with the advantages of another. Therefore, hybrid models have proven effective in cost estimation (Ahiaga-Dagbui et al., 2013; Ahiaga-Dagbui and Smith, 2014).

Further, the current research primarily focuses on estimation using Artificial intelligence, such as machine learning or developing automated models aligning with Industry 4.0. Therefore, this study explores a solution for conceptual estimation through machine learning, or a combination of several modelling techniques, to automate the process and minimise manual errors in road projects' conceptual estimation.

8.4 Research Methodology

This study aims to develop a sound and reliable cost estimation model for road projects in NZ. The same case study used to develop RA and ANN models in Chapters 6, and 7 were used in this study to enable a better cross-comparison among the models. Projects were used for this model. Fifty road projects in NZ were completed between 2002 and 2022, and the scope comprised new construction and significant alterations. Out of the fifty projects, seven were eliminated as outliers. Mahalanobis distance was used to identify the outliers. According to Tabachnick and Fidell (2001), the *p*-value of Mahalanobis distance must be higher than the alpha value of 0.001. Otherwise, the particular case is considered as an outlier. As discussed in previous chapters, fewer new road construction projects were available in NZ. Hence, the scope was expanded to include the significant alterations as well. Once the project cost database is established, Figure 8.1 highlights the entire process of developing the final statistical model. The process is described in detail under the model development section.

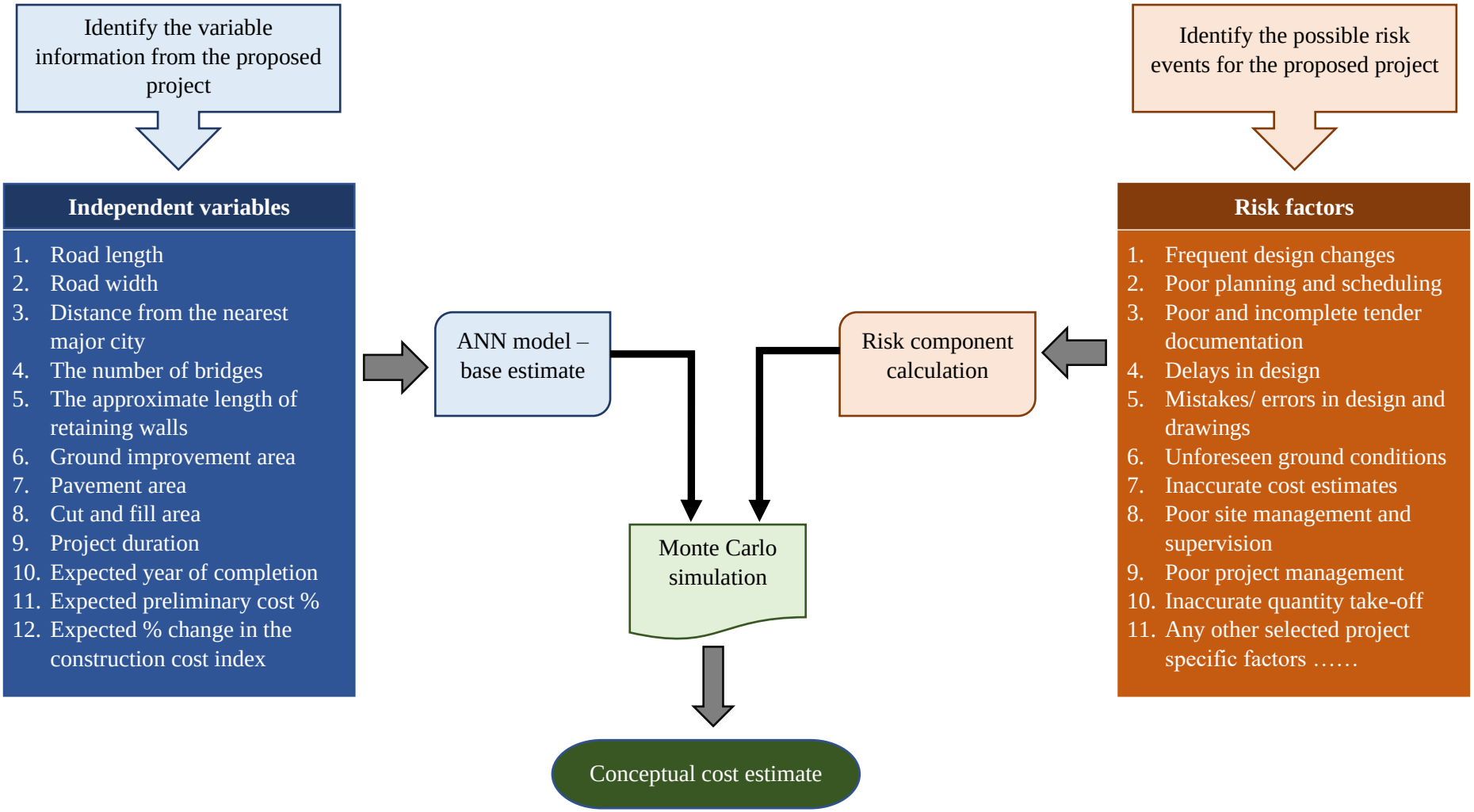

Figure 8.1. *Final model calculation process*

As the RA model was not considered for the hybrid model, the methodology used in developing the regression model is not discussed here. Nevertheless, the RA model development process is explained in-depth in Chapter 6. The final model consists of two parts. The first part is the ANN model. The ANN model developed in Chapter 7, which used technical variables, was considered to develop the base cost estimate. Therefore, the base estimate is a technical aspects-based estimate. The ANN model was developed using MATLAB, as explained in the chapter above. The second part of the model was to combine the ANN model output with the MCS model. For MCS, the @Risk Add-in provided by Palisade is used. Once the final model was created, the model was validated through the same data set of sixteen projects used for RA and ANN models.

The following section briefly discusses the RA-based model developed in this research.

8.5 Regression analysis-based cost model

In Chapter 6, a model was developed for cost estimation of NZ road projects using regression analysis (RA). The RA measures the relationship between the dependent variable and one or more independent variables (Jablonowski and MacEachern, 2009; Young, 2017). The following equation (Eq. 1) is the general representation of multiple regression analysis.

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ldots\ldots + \beta_n * X_n + \varepsilon \quad \ldots\ldots\ldots\ldots\ldots\ldots (Eq.\ 1)$$

Where "Y" is the overall budget of the project, which is the dependent variable, "Xn" are the factors that constitute the overall budget, identified as independent variables. "βn" is an estimated constant. Further, Young (2017) explained that "ε" is an error term derived from the difference between the dependent variable's actual and predictor values.

Based on the above equation, eight models were developed using twelve independent variables. The dependent variable was the project's final cost (refer to Chapter 6). Out of the eight models,

the most reliable model showed a Mean Absolute Percentage Error (MAPE) of 11.82%. However, due to the higher MAPE value, in Chapter 7, ANN was adopted to investigate if a more reliable cost model with a lesser error rate could be developed. The following section discusses, in summary, the ANN-based model developed in Chapter 7.

8.6 ANN-based cost model

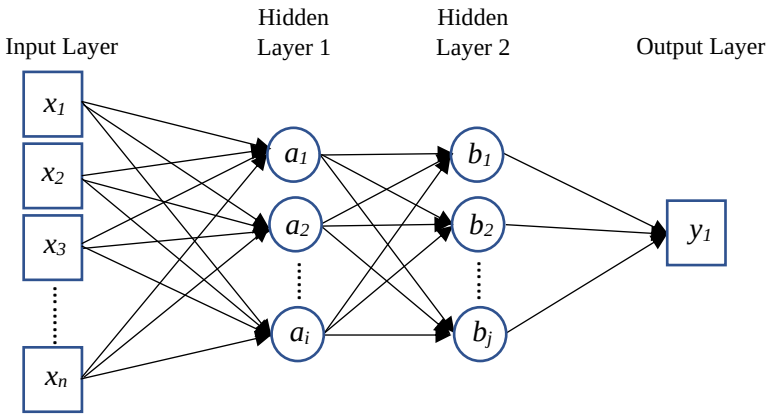

Figure 8.2. *ANN model architecture (Source – Chapter 7)*

According to Tijanic et al. (2020), ANN functions similarly to the human brain cells and nervous system. Therefore, ANN can be used for predictions and several other functions related to decision-making. Figure 1 shows the general architecture of the ANN model. There are three main phases of an ANN model. Firstly, the input layer is where all the input data is entered, and then the hidden layer. There can be one or more hidden layers in a model. Finally, the output layer is where the intended outcome is generated. Therefore, in most cases, the output layer has only one neuron.

Using this technique, Chapter 7 developed a set of cost estimation models by varying the number of independent variables, the number of neurons in hidden layers, and the activation function (Tanh and Sigmoid). The models were developed using data from forty-three completed projects. There were three phases in the model development: training, testing, and validation. Out of the six models developed, which comprised twelve independent variables, one neuron in each hidden layer with

the sigmoid activation function achieved the most reliable performance. The most accurate and reliable model achieved 0.977 and 0.993 as R-squared, respectively, in the training and testing phases, and the root mean squared error is also minimal compared to the other models. Further, during the validation phase, the best model showed the lowest mean absolute percentage error (MAPE), 11.82%.

Even though the results of the ANN model were significantly better than the RA-based model, the error is still considerably high. The main reason for this high error rate is that neither ANN nor RA techniques are suitable for predicting future uncertainties. For instance, El-Maaty et al. (2017) developed an RA-based model, while El-Kholy (2019) developed an ANN-based model for road projects. Further, both models used the same independent variables, mainly risk factors. However, both models showed more than 30% of error and failed to achieve a reliable outcome. Therefore, the ANN-based estimates did not consider the risk factors.

In Chapter 5, the critical risk factors that affect the project cost were identified. Out of the most significant, ten factors were frequent design changes, poor planning and management, poor and incomplete tender documentation, delay in design, mistakes/ errors in design and drawings, unforeseen ground conditions, inaccurate cost estimates, poor site management and supervision, poor project management, and inaccurate quantity take-off. Consequently, the effect of these risk factors cannot be quantified numerically using any straightforward methods. Therefore, qualitative risk analysis is required to identify the magnitude of the risk caused by each factor.

The following section discusses the risk factor and its estimation in-depth to incorporate the risk into the conceptual estimation.

8.7 Risk-based estimation

Both the RA and ANN models failed to consider the risk estimation part. Although ANN is an AI-based cost model that performs well in uncertain and complex situations, ANN needs comprehensive cost data from previous projects to deal with similar risks expected in the proposed projects (Ashtari et al., 2022). However, it is often difficult to find such information from previous projects to work the ANN model alone to accurately calculate both the base estimate and risk estimation. In Chapter 5, two hybrids were identified with significantly low error percentages. One model combined ANN and Fuzzy (Ahiaga-Dagbui et al., 2013), while the other combined RA, SVM and data mining techniques (Ahiaga-Dagbui and Smith, 2014). However, both models were developed for water infrastructure projects. Therefore, we cannot apply these hybrid models to road projects directly. Therefore, this chapter looks into combining the RA or ANN models with another technique that can reduce errors and increase performance. Ideally, the technique should address the construction risks of the project, as this is the component missing in the RA and ANN models.

According to the NZTA Cost Estimation Manual (NZTA, 2021), NZTA uses the Monte Carlo simulation (MCS) technique to analyse the risk. By using MCS already used in NZTA, it would be easy to communicate the research impact to the current NZ industry. Although the industry uses MCS for risk analysis, the base estimate is based on manual and expert judgment-based calculations. Therefore, the reliability of the overall estimate has not improved (NZTA, 2021). Therefore, this study examines the possibility of having a hybrid model by combining ANN with MCS. Since the MAPE of RA is higher than ANN, this study only focuses on combining ANN with MCS.

8.8 Monte Carlo Simulation

Construction projects face various types of risks throughout their construction phase. Considering the scale of investments or funds in construction projects, it is essential to forecast the risks at the estimation phase. However, construction projects generally face cost overruns (Peleskei et al., 2015). Flyvbjerg et al. (2003) also confirmed the above statement and stated that 9 out of 10 transportation infrastructure projects face cost overruns.

In Chapter 5, it was noticed that statistical techniques such as RA, ANN, and SVM do not perform well with risk-related variables. Therefore, these techniques cannot correctly predict the risk involved in construction projects. However, risk estimation is a significant component of the project cost. Without proper risk analysis and estimation, the project may experience cost overruns. Therefore, this chapter investigates the risk estimation of construction projects.

Construction risk cannot be estimated with exact figures because the risk is an uncertain future situation, which could either happen or not happen. Sometimes, the risk may occur but at a different magnitude than estimated. Therefore, the best way to estimate the risk is through its probability. However, Peleskei et al. (2015) stated that proper risk analysis is often challenging during the estimation stage, yet it is very significant for the reliability of the estimation. Over the years, researchers have developed several models and techniques to analyse the project risk. MCS is a well-known technique commonly used in the construction industry to estimate risk probability. Therefore, this study analyses the road project data used in chapters 6 and 7 to analyse the risk probability using MCS.

Risk can be defined from various perspectives. However, Hilson (2004) defined *risk* as any potential uncertainties that will affect one or more objectives if it occurs. Risks are threats or negative uncertainties and opportunities or positive uncertainties (Peleskei et al., 2015). Further,

Girmscheid and Busch (2014) identified six risk types involved in construction projects: legal, scheduling, technical, financial, management, and environmental risks. Peleskei et al. (2015) added that the magnitude of the impact and possibility of risk occurrence, as mentioned above, differ from project to project.

According to the estimation classification system introduced by AACE (2020), at the conceptual stage, usually, the estimate should be expected to have -20% to -50% of a lower range to +30% to +100% of a higher range of variation. Therefore, it is clear that the project risk level is higher at this stage due to the limited information availability. However, the conceptual estimate will lead to significant project decisions and budget establishment. Hence, it is crucial to have a reliable estimate. Consequently, risk estimation is required to ascertain the possible percentage of variation from the primary estimate. AACE (2020) defined contingency as an amount added to the project estimate to allow for items, conditions, or events for which the state, occurrence, or effect is uncertain and that occurrence will possibly demand additional cost. Generally, the contingency is estimated using statistical analysis of experts' judgement based on past experience and lessons learnt (Maronati and Petrovic, 2019). However, it was observed that some of the projects allow a general contingency sum as a percentage of the estimated value (Gbajobi et al., 2018; Peleskei et al., 2015). However, allowing the contingency without a proper risk analysis creates a riskier situation as it is uncertain what to expect during construction. Further, if the project faces unexpected risk with a higher financial impact, the allowed contingency may not be sufficient (Gbajobi et al., 2018).

According to researchers, there are several ways to handle the risk events. Nevertheless, the most effective way is to take mitigation actions. However, mitigating or reducing risks involves using the contingency allowance, which means requiring additional money. When a risk is encountered in a project, there could be two significant mitigation actions. The most likely action would be the

design change to reduce the cost, while the most unlikely action would be to allow additional funding. Road projects, mainly funded by the government, would face challenges in allocating additional funds for an ongoing project, mainly if funds were not allocated during the conceptual stage when the decision-making process happens.

Therefore, contingency allowance should be calculated systematically to cover all expected risks. Girmscheid and Busch (2007) and Peleskei et al. (2015) identified MCS as an effective technique for quantifying the risk, which can be assessed as the contingency to the project estimation. Peleskei et al. (2015) further emphasised that much research has been done to analyse the risks and their potential impacts. Conversely, the criterion for input parameter selection needs to be adequately investigated. Nonetheless, MCS is a widespread technique in risk probability estimation. Therefore, this study investigates MCS as a technique of risk estimation to combine with the ANN-based model developed in Chapter 7 to improve its performance and reliability.

The following section discusses the research methodology adopted in this study to develop the final cost model.

8.9 Risk component estimation

Even though the ANN-based model developed a reliable base estimate, the estimation of the expected risk component plays a crucial role in achieving the best outcome of the estimate. In Atapattu et al. (2023), fifty-three factors that affect the project cost were identified. In any construction project, these risk management steps, risk identification, analysis, response, and monitoring, must be conducted. The fifty-three factors identified by Atapattu et al. (2023) generally happen in the risk identification step. In most projects, a risk register is prepared based on the identified risks. Then, the Risk analysis is the most critical phase of all. Similar to the study of Atapattu et al. (2023), prioritising and identifying the most significant factors is the second step. The risk matrix is generally used in most projects to identify the severity of the risks. Table 8.1

shows a general risk matrix used in NZ projects. Similarly, for the opportunities, a similar matrix can be used.

Table 8.1. *Risk matrix (Source: NZ Government Procurement, 2019)*

LIKELIHOOD						
	Almost certain	Moderate risk	High risk	High risk	Extreme risk	Extreme risk
	Likely	Moderate risk	Moderate risk	High risk	Extreme risk	Extreme risk
	Possible	Low risk	Moderate risk	High risk	High risk	Extreme risk
	Unlikely	Low risk	Moderate risk	Moderate risk	High risk	High risk
	Rare	Low risk	Low risk	Low risk	Moderate risk	High risk
		Very low	Minor	Medium	Major	Substantial
				IMPACT		

Figure 8.3 shows the significant factors identified through a questionnaire survey distributed to NZ Quantity Surveyors. Once the risks are identified, each risk component has to be estimated. The risks in the red zone (high and extreme) within the Table 8.1 risk matrix should be estimated and considered as part of the project cost contingency (NZ Government Procurement, 2019). However, according to Laryea and Hughes (2006), there is no systematic process to estimate the cost contribution of significant risk activities. The most common way is to use expert judgment.

In contrast, Potts and Ankrah (2014) identified three methods to calculate the risk allowance of a construction project. The first method is the decision tree method, which analyses the different risk combination possibilities the project may face and identifies the risk component accordingly. The second method is to calculate the risk-based estimation using the MCS method. The third method is the central limit theory, which calculates the overall risk allowance for the project using a simple calculation formula. The fourth method is multiple estimating using root mean square.

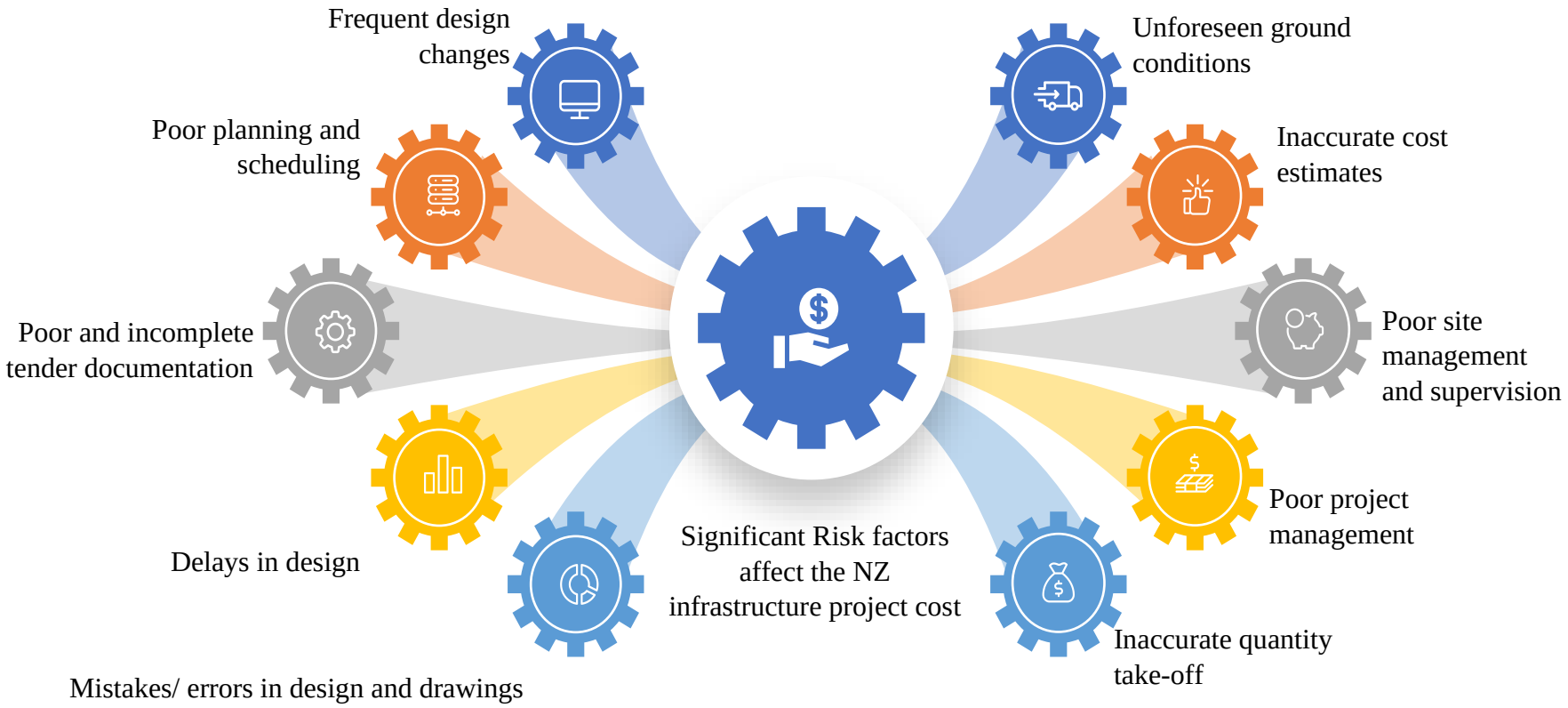

Figure 8.3. *Significant risk factors affect the NZ transportation infrastructure project cost (Source: Atapattu et al., 2023)*

In this study, the method adopted to calculate the risk component was the second method mentioned above. This method requires the risk component to be calculated in three different likelihoods: minimum cost, most likely cost, and maximum cost. The expert's opinions and experiences and the lessons learned from past completed projects can be used to establish the costs mentioned above. Once these three measures are established, MCS can calculate the total risk-based estimation, combining the risk-free base estimate.

8.10 Model development

Risk-based estimation of MCS needs two primary measures to run the model. The first measure is the base estimate of the project. Generally, the base estimate comprises the project's basic direct cost, such as materials, labour, and plant, and other indirect costs, such as profits and on-site and offsite overheads (Liu and Napier, 2010). Since there is less information available, parametric estimation methods are required at the conceptual stage of the project. However, in this study, the aim is to improve the reliability of the conceptual estimate of the NZ road projects. Therefore, the predicted cost of the ANN-based model developed in Chapter 7 is the base estimate for the final model development.

The second measure, the expected risk component, is the most critical ingredient of the final model. In any construction project, the risk management process is implemented throughout the project to identify, analyse, estimate, and mitigate the expected risks. It is necessary to have the perspective of different project stakeholders to ascertain the risk component more accurately (Liu and Napier, 2010). Consequently, the purpose of the risk management workshop is to use the professional judgement of different project stakeholders to determine and quantify the cost-effectiveness of the expected risk events. This risk component will be this model's second variable or measure. Table 8.2 shows the risk calculation table. The cost data related to this risk analysis

were selected from all the cases used for the model development. Subsequently, it was evident that most cases had high risks in the top ten risk factors listed in Table 8.2.

Table 8.2. *Risk component calculation table*

	Risk event	Impact	Likelihood	Risk matrix	Risk cost (NZ$)		
					Lower-level cost	Most likely cost	Higher level cost
1	Frequent design changes	Substantial	Almost certain	Extreme high	XX	XX	XX
2	Poor planning and scheduling				XX	XX	XX
3	Poor and incomplete tender documentation				XX	XX	XX
4	Delays in design				XX	XX	XX
5	Mistakes/ errors in design and drawings				XX	XX	XX
6	Unforeseen ground conditions				XX	XX	XX
7	Inaccurate cost estimates				XX	XX	XX
8	Poor site management and supervision				XX	XX	XX
9	Poor project management				XX	XX	XX
10	Inaccurate quantity take-off				XX	XX	XX
11	Any other uncertain events specific to the proposed project				XX	XX	XX
	Total risk component				**XXX**	**XXX**	**XXX**

The ten factors identified by Atapattu et al. (2023) are also listed in the above table. According to the study carried out among the NZ transportation infrastructure experts, these ten factors create extremely high risks to the projects (Atapattu et al., 2023). In addition, there would be risk events specific to the proposed project. These events must be identified and their risk probability evaluated. Project risk analysis involves identifying the risks and uncertainties that may occur during the project delivery and analysing the severity of their impact on the project. About the cost estimation, the main concern is the impact of the risk events on the project cost. According to Vose (2008), cost-related uncertainties are modelled using PERT or triangular distribution distributions. The two variables mentioned above, base estimate and risk estimate, are added to the MCS and the model is run to estimate the most probable final cost of the model.

MCS requires data on the probability density function (PDF) to analyse the construction risks (Wing Chau, 1995). In the MCS method, PDF is generated using the provided range for the abovementioned measures by generating many combinations of potential cost outcomes.

Therefore, MCS considers the probability of the risk occurrence based on the series of possible outcomes. Therefore, MCS is more robust in risk-based estimation than other statistical techniques.

Once the ANN-generated base estimate and the possible risk estimate are combined, the next stage is to generate the probability distribution using the PDF generated. In this distribution, there are countless possible outcomes for the project cost. The only change in this is the probability of the risk. Since this is developed at the project's conceptual stage, the project team cannot produce an exact risk estimate. Consequently, general norms are accepted in the industry for probability-based risk estimation. The cumulative probability distribution has two significant positions: P50 and P95 estimates. Based on the provided base estimate, the P50 estimate assumes the probability of risk occurrence is 50%. In P95, the probability of occurrence is assumed to be 95%. According to the cost estimation manual of Waka Kotahi/ New Zealand Transport Agency (NZTA, 2021), the difference between the base and P50 estimates is considered the contingency. However, if the base estimate shows a low level of accuracy, then the P50 estimate will also not be reliable. Therefore, the estimate provided by the ANN model developed in this study shows a higher level of accuracy.

Further, the difference between P50 and P95 is also a considerable risk allowance in road projects, considering its high-risk involvement. Unavoidable reasons may cause the cost to go beyond the P50 estimate. Hence, it is advisable to have this difference as a known factor for the funding party of the project. Figure 8.4 shows the model developed in this study and how the final estimate is derived.

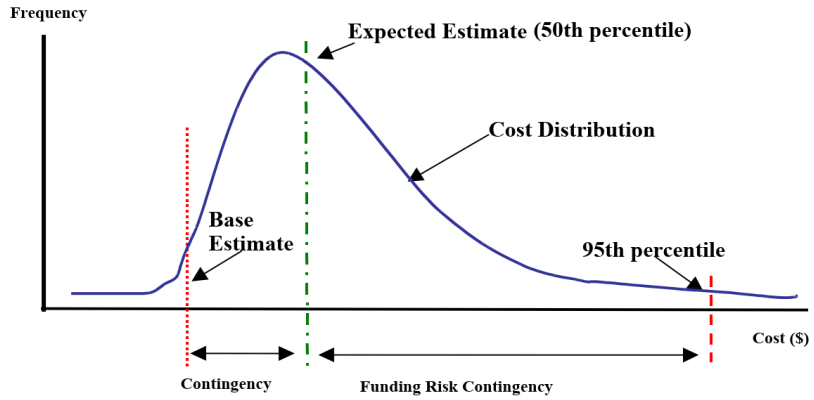

Figure 8.4. *Risk-based estimation process (Source: NZTA, 2021)*

According to Figure 8.4, the base estimate shows the ANN-generated estimate, which is more accurate than the traditional cost estimation models. Then, the P50 estimate shows the expected outcome of this study, which is the predicted final cost of the project. The predicted final cost is a composition of the base estimate and the risk contingency. The graph also shows the P95 estimate. If the project experiences further risks beyond the expected allowance, the additional cost must be between P50 and P95 estimates. This model was run for all sixteen projects used for model validation of the ANN model and derived the predicted final cost of the projects to compare with the actual cost.

The following section discusses model testing and validation.

8.11 Model testing and validation

The same data set used in chapters 6 and 7 will be used to validate the hybrid model. In this way, the performance of the three models can be compared efficiently and effectively. The equation (Eq. 2) introduced by Kim et al. (2004) for Mean Absolute Percentage Error (MAPE) is used to measure the error of the results in model validation.

$$\text{MAPE} = \frac{\sum \left| \frac{C_e - C_a}{C_a} \text{ x } 100 \right|}{n} \quad \text{.....(Eq. 2)}$$

C_e indicates the estimated cost through the new cost model, C_a indicates the actual cost of the case, and n denotes the number of cases used for the validation.

Using the base estimate generated from the ANN model, the MCS was run using the @Risk add-in. The simulation was based on 10,000 iterations. According to Bouayed (2016), it is suitable to have several iterations of more than 5,000 for project risk analysis. For each iteration, MCS select a different cost for the particular project based on the provided ANN-generated base estimate and the risk estimate. Once the simulation was completed, all the possible outcomes were plotted into a histogram, indicating the total project cost distribution depending on the expected risk contingency. Subsequently, the P50 estimate is selected as the final cost of each test case to compare with the project's actual cost. Figure 8.5 shows the 10,000 iterations generated by the MCS model for one of the test cases.

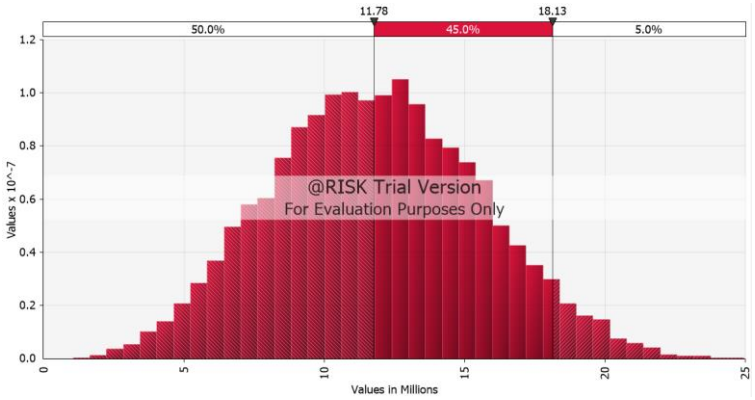

Figure 8.5. *Monte Carlo simulation model for test cases (Sources: @Risk add-in for Excel from Palisade)*
Legend – Base estimate = 11,331,302.27; P50 estimate (Base estimate + Risk component) = 11,737,125.78

Based on the MCS distribution, absolute percentage errors (APE) were calculated for each test case separately, and later, MAPE was calculated using Eq. 2, mentioned above. Table 8.3 shows the MAPEs of all three models developed for comparison. The models were the RA-based model developed in Chapter 6, the ANN-based model developed in Chapter 7, and the final ANN and MCS combined hybrid model.

Table 8.3. *Model validation results (Sources: RA model and ANN model data derived from chapters 6 and 7)*

	Actual Cost (NZ$)	RA Model		ANN Model		ANN+MCS Hybrid Model	
		Estimated Cost (NZ$)	APE (%)	Estimated Cost (NZ$)	APE (%)	Estimated Cost (NZ$)	APE (%)
1	5,379,458.36	3,972,192.05	26.16%	6,899,887.60	28.26%	5,593,479.92	3.98%
2	5,469,708.59	6,711,879.41	22.71%	5,843,324.38	6.83%	5,803,066.59	6.09%
3	6,183,854.43	5,083,746.73	17.79%	6,589,217.26	6.56%	6,207,842.51	0.39%
4	6,761,382.89	8,872,286.63	31.22%	5,083,693.28	24.81%	6,865,439.54	1.54%
5	6,909,174.34	4,227,723.78	38.81%	5,477,487.87	20.72%	6,864,505.18	0.65%
6	6,964,095.39	297,631.46	9.57%	6,641,822.06	4.63%	6,946,067.29	0.26%

7	7,075,124.51	6,051,353.99	14.47%	5,082,925.09	28.16%	7,095,423.07	0.29%
8	7,363,881.68	11,210,037.08	52.23%	5,499,259.58	25.32%	7,242,384.12	1.65%
9	8,111,879.42	10,784,743.69	32.95%	7,556,502.07	6.85%	7,770,549.26	4.21%
10	9,351,966.35	12,127,629.96	29.68%	8,994,509.72	3.82%	10,019,073.62	7.13%
11	9,470,985.65	10,055,345.46	6.17%	8,044,963.22	15.06%	8,931,138.80	5.70%
12	11,000,688.37	9,983,124.70	9.25%	11,331,302.27	3.01%	11,737,125.78	6.69%
13	11,911,646.74	10,828,878.05	9.09%	11,251,509.20	5.54%	12,655,189.15	6.24%
14	13,028,819.73	10,752,684.92	17.47%	12,815,319.43	1.64%	13,174,239.61	1.12%
15	16,068,756.72	17,984,152.52	11.92%	16,946,004.03	5.46%	15,338,080.06	4.55%
16	17,092,831.00	19,161,063.55	12.10%	16,678,525.04	2.42%	16,071,320.03	5.98%
MAPE (%)			**21.35%**		**11.82%**		**3.53%**

APE - Absolute Percentage Error; MAPE - Mean Absolute Percentage Error; RA – Regression Analysis; ANN – Artificial Neural Network; MCS – Monte Carlo Simulation

The data in the table can be further elaborated and clearly explained using a graph of the actual cost against the predicted cost of each model. Figure 8.6 compares the actual cost of the 16-test data set against the predicted cost of all three models developed in Chapters 6, 7, and 8: RA-based, ANN-based, and ANN and MCS combined hybrid models. Although Table 8.3 already clarified that the final model shows the lowest error rate, it is even more evident in Figure 8.6. All the test cases indicate a slight difference between actual and predicted values. In contrast, RA and ANN-based models show significant deviations in some cases, and their MAPE is higher than the hybrid model.

The following section will discuss the findings of this study in-depth, along with similarities and differences in the literature.

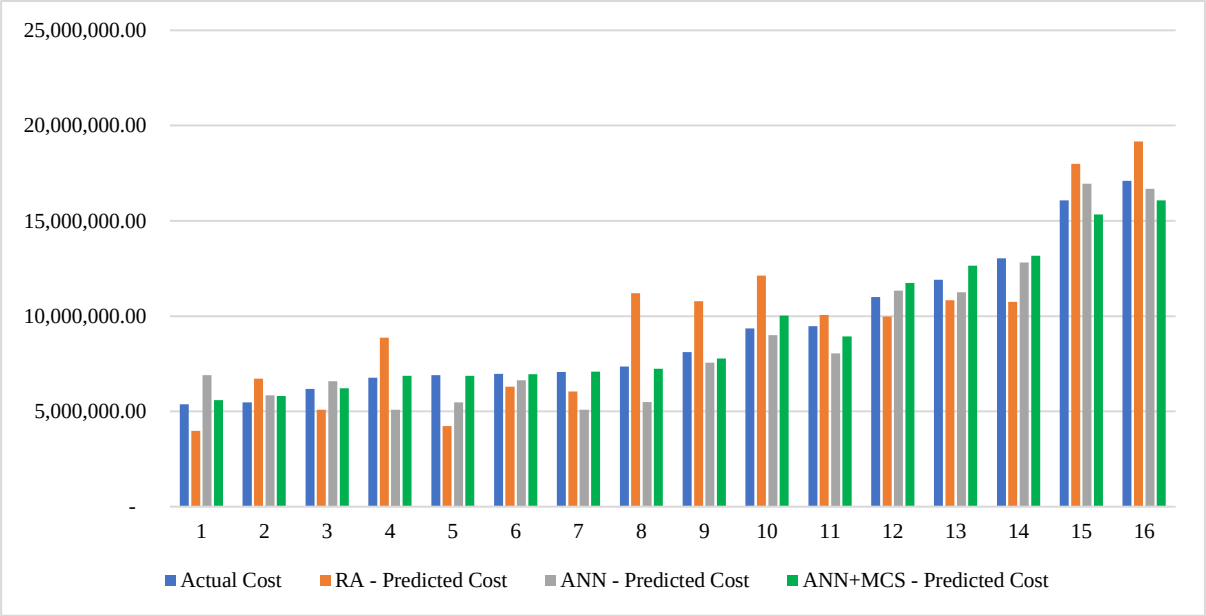

Figure 8.6. *Actual cost vs predicted cost of each model*

8.12 Discussion

The testing and validation results shown in Table 8.3 indicate that hybrid models combining several statistical techniques perform better than a model with a single technique. According to the results, the RA-based model achieved a MAPE of 21.35%, while the ANN-based model achieved better performance with a MAPE of 11.82%. Further, the final model that combined ANN and MCS techniques achieved even better performance with a MAPE of 3.53%.

According to the literature findings, only two hybrid models were developed by combining several techniques. However, both models were developed for water infrastructure projects. Ahiaga-Dagbui *et al*. (2013) developed a model combining ANN and fuzzy logic and achieved lowest error of +0.6% to +0.8%. Ahiaga-Dagbui and Smith (2014) developed another model combining RA, SVM and data mining techniques in another study. Their model showed an error between -3.83% and +2.33%. Though the models showed significantly better performance, they cannot be applied to road projects since they focused on water infrastructure.

Nevertheless, the models considered variables that need information from the post-design stage. In Chapters 6 and 8, the models developed using RA and ANN were compared with similar models in the literature. Table 8.4 summarises the performances of models developed for road projects identified in the literature.

Table 8.4. *Performance of similar cost models from literature*

	Reference	Modelling technique	Performance/ MAPE
1	Shr and Chen (2006)	RA	±5%
2	El-Maaty *et al*. (2017)	RA	+30.42%
3	Sodikov (2005)	RA	+30% to +36%
4	Adel *et al*. (2016)	ANN	Training - +4.51%; Evaluation- +5.8%; Validation - +16.0%
5	Cirilovic et al. (2014)	ANN	R-squared – 0.7469
6	Dimitriou et al. (2018)	ANN	Piers - 34%; Precast beams – 11.48%; Cast-in-situ – 13.94%; Cantilever – 16.12%
7	El-Kholy (2019)	ANN	+39.8%
8	Hegazy and Ayed (1998)	ANN	Back propagation – 10.4%; Simplex optimisation – 1%; Genetic algorithm – 21.8%
9	Jaafari et al. (2021)	ANN	0.038%

10	Mahalakshmi and Rajasekaran (2018)	ANN	8.46%
11	Roxas et al. (2019)	ANN	Average 20% error
12	Sodikov (2005)	ANN	+24% to +26%
13	Tijanic et al. (2020)	ANN	Training – 10.22%; Validation – 13.06%
14	Xue et al. (2020)	ANN	With bridge and tunnel data = 22.84% Without bridge and tunnel data = 40.62%
15	El-Sawalhi (2015)	SVM	-5% average error
16	El-Maaty *et al*. (2017)	Fuzzy logic	+40.37%

In Chapter 7, the literature review identified eight other RA-based cost models developed for road projects but are not in Table 8.4 (Ahmed, 2021; Cirilovic et al., 2014; El-Maaty et al., 2017; Kim et al., 2008; Lin and Techapeeraparnich, 2019; Mahamid, 2011; Ou and Swarthout, 1986; Zhang et al., 2017). However, our study's final model performed satisfactorily well compared to these models. As mentioned earlier, all the models mostly used variables that cannot be calculated during the pre-design stage. Further, the above models performed less well than our model. For instance, the error rates of the models identified in the literature were higher than 3.53%.

In contrast, out of the models listed in Table 8.4, only the model developed by Jaafari et al. (2021) accomplished an error rate of less than 3.53%. All the other models showed higher error rates. However, variables used in their model cannot be considered during the pre-design estimation. For example, variables such as 'the number of trees' and 'size of trees' can be decided after the initial design is completed. Nevertheless, the focus of the research is only on forest roads. Therefore, this model may not apply to other road types without investigation. On the other hand, the model developed in our study considers only general road construction variables that can be considered within any context or country. Therefore, the model can be easily adapted to other contexts as well.

Another notable issue about the models in Table 8.4 is that each contains variables related to information from post-design stages. Therefore, these models do not apply to conceptual cost estimation. The purpose of these models was to resolve estimation issues at the tender stage or the post-contract stage.

Chapters 6 and 7 emphasise that RA and ANN techniques perform poorly in risk identification and calculation. However, some of the models (Cirilovic et al., 2014; El-Kholy, 2019; El-*Maaty et al.,* 2017) contain variables that involve the future uncertainty of the project. The risk component of the project cannot be calculated as an exact amount with certainty; instead, it is a probability-based calculation. That was the reason our model combines ANN and MCS as one hybrid. However, the models mentioned above considered only one technique to calculate the project's technical and risk aspects. That is one of the significant drawbacks of using a single modelling technique for cost estimation.

In this study, the risk factors were mainly based on the study of Atapattu et al. (2023). The top factors were frequent design changes, poor planning and scheduling, poor and incomplete tender documentation, delays in design, mistakes/ errors in design and drawings, unforeseen ground conditions, inaccurate cost estimates, poor site management and supervision, poor project management, and inaccurate quantities take-off. The priority and severity of these factors were identified based on the NZ road projects study. Hence, applying the model in other contexts may require carrying out and identifying the crucial risk factors in the context of the proposed project.

In contrast to the risk factors, the independent variables were selected, allowing the model to be generalisable to other contexts. There were no variables considered to be specific only to NZ road projects. Generally, the geographical location is the primary variable that can affect the project context. However, this model adjusted this variable into 'distance from the nearest major city' to allow adaptability to other countries. The nearest major city would be where the material, labour plant supply, contractor and subcontractor setup are based. Therefore, this variable can impact the project cost more severely than the geographical location.

The following section discusses the conclusions and recommendations for further research.

8.13 Conclusion

Issues related to estimates and cost overruns have been a significant area of attention for researchers and industry experts. The conceptual estimate of the project is crucial since it is the basis for the funding and decision-making of the project. However, there has been a significant difference between the conceptual estimate and the project's final cost. That, in turn, severely affects the funding process as additional funds must be secured. The funding issue becomes even more severe when the public sector funds the project. Therefore, this research aimed to improve the reliability and accuracy of the current conceptual cost estimates in road construction.

The model developed in this study was a hybrid of ANN and MCS. ANN was utilised to calculate the technical aspects of the project as the base estimate. MCS is considered to calculate the probability and cost of the uncertainty. Overall, this hybrid model produced estimates with a 3.53% mean absolute percentage error, a significantly improved performance compared to the literature and the current cost overruns faced by road projects.

The main takeaway of this project is that at the conceptual stage, with less information available, it is more crucial to use modelling techniques than the traditional parametric methods. Machine learning and simulation models provide more robust estimation than traditional methods. With the model considering the risk component probability, capturing the lessons learnt and knowledge from completed projects is vital. Consequently, that knowledge can be transferred to the hybrid model to improve the accuracy and reliability further.

This study's initial estimate produced using the ANN contains a MAPE of 11.82%. Then, on top of that estimate, the risk component is added using the probability density function and simulation modelling. If the final estimate (ANN + MCS) is revealed to those making a tender or a selected contractor, they intend to achieve that budget target. That could result in experiencing additional

overruns from the expected budget. However, if only the ANN-based estimate is known to the contractor, they would be more careful when the actual cost comes near the budget limit. That is because it is unknown to the contractor that some additional funds are left before surpassing the actual budget limit.

Consequently, even if there were cost overruns compared to the ANN-based estimate, the client would still have funds left to cover the additional cost. Therefore, this study recommends that the final conceptual estimate combining ANN and MCS be only known to the consultant and the client. The contractor should only know the lower estimate without incorporating the risk contingency. Some projects may experience more positive and low adverse risks, resulting in a lower overall final estimate than the ANN-generated base estimate. In this situation, the final estimate can be communicated to the contractor as it is lower than the base cost. The key takeaway of this discussion is that the additional risk contingency must be kept for clients' records only.

This study considered forty-three cases for model development and sixteen for model validation. However, model reliability and accuracy can be improved by having more test cases as the model reliability is increased with more knowledge captured from previous projects. Further, this study considered only ANN as the technique for base estimate calculations. There may be other modelling techniques which provide more robust estimates. Hence, it is recommended that hybrids of other techniques be developed to see how their reliability compares to the ANN + MCS hybrid. Finally, the risk factors considered in the MCS were based on the study conducted by Atapattu et al. (2023) using NZ road projects, which may have different priorities in different contexts. Therefore, a proper risk analysis is recommended before the conceptual estimate of this model is used in other contexts.

8.14 Epilogue

This chapter studied combining two cost modelling techniques to increase the accuracy and reliability of the models developed in chapters 6 and 7. Consequently, the study developed a hybrid model combining the ANN model developed in Chapter 7 with the MCS modelling technique. As a result, the study achieved the aim of having a reliable model to use for the conceptual cost estimation during the pre-design stage. The next chapter discusses the overall findings of this research towards achieving the final research aim.

9 Discussion

9.1 Prologue

This chapter combines all the studies done in this thesis to make the research findings meaningful. From Chapter 3 to Chapter 8, separate studies were done with six journal publications. However, the six studies were done sequentially, answering each research question separately but moving towards the ultimate research aim of this thesis. Thus, the discussion chapter evaluates the findings and conclusions of each chapter towards achieving the overall research aim.

9.2 Risk factors affect the transportation infrastructure projects in NZ

Once the cost overrun was identified as a significant issue in NZ road projects, the next step was to investigate the major risk factors affecting the project cost. That is because these causes may be the reason for experiencing cost overruns. Moreover, if the factors are not accommodated in the conceptual estimation process, the solution must be a cost model to address the causes of cost overruns.

Through a systematic literature review, fifty-three risk factors that affect the project's final cost were identified. The factors were categorised into ten sub-groups: contractor-related, project management and contract administration-related, design and documentation-related, financial management-related, information and communication technology-related, labour management-related, material management-related, environment-related, Psychology-related, and political-related. Since the factors were identified through the literature, ranking them according to their impact on NZ-specific behaviour is necessary. Therefore, a questionnaire survey was carried out of Quantity Surveyors and Estimators with experience on TI projects in NZ. Then, the survey data

was analysed using quantitative analytical methods, and the factors were prioritised based on the severity of their impact on the project cost.

The most significant ten factors in the order of severity were frequent design changes, poor planning and scheduling, poor and incomplete tender documentation, delays in design, mistakes/ errors in design and drawings, unforeseen ground conditions, inaccurate cost estimates, poor site management and supervision, poor project management, and inaccurate quantity take-off.

Figure 9.1 shows the main risk factors affecting the project cost. The severity of these factors was prioritised based on the Quantity Surveyors perspective. In contrast, the results could have been different if the survey had been conducted considering all the construction professionals. Nonetheless, the survey was circulated only among the project's cost management professionals since this research aims to improve the conceptual cost estimation. Therefore, the identified factors should be based on the impact on the project cost rather than on the rest of the project characteristics. Most of the significant factors are connected to the issues in the pre-contract stage. Therefore, the study identified that measures should be taken to minimise the errors during the pre-contract stage.

The following Table 9.1 summarises the findings of this research against the findings of the literature. That was discussed in detail in Chapter 4.

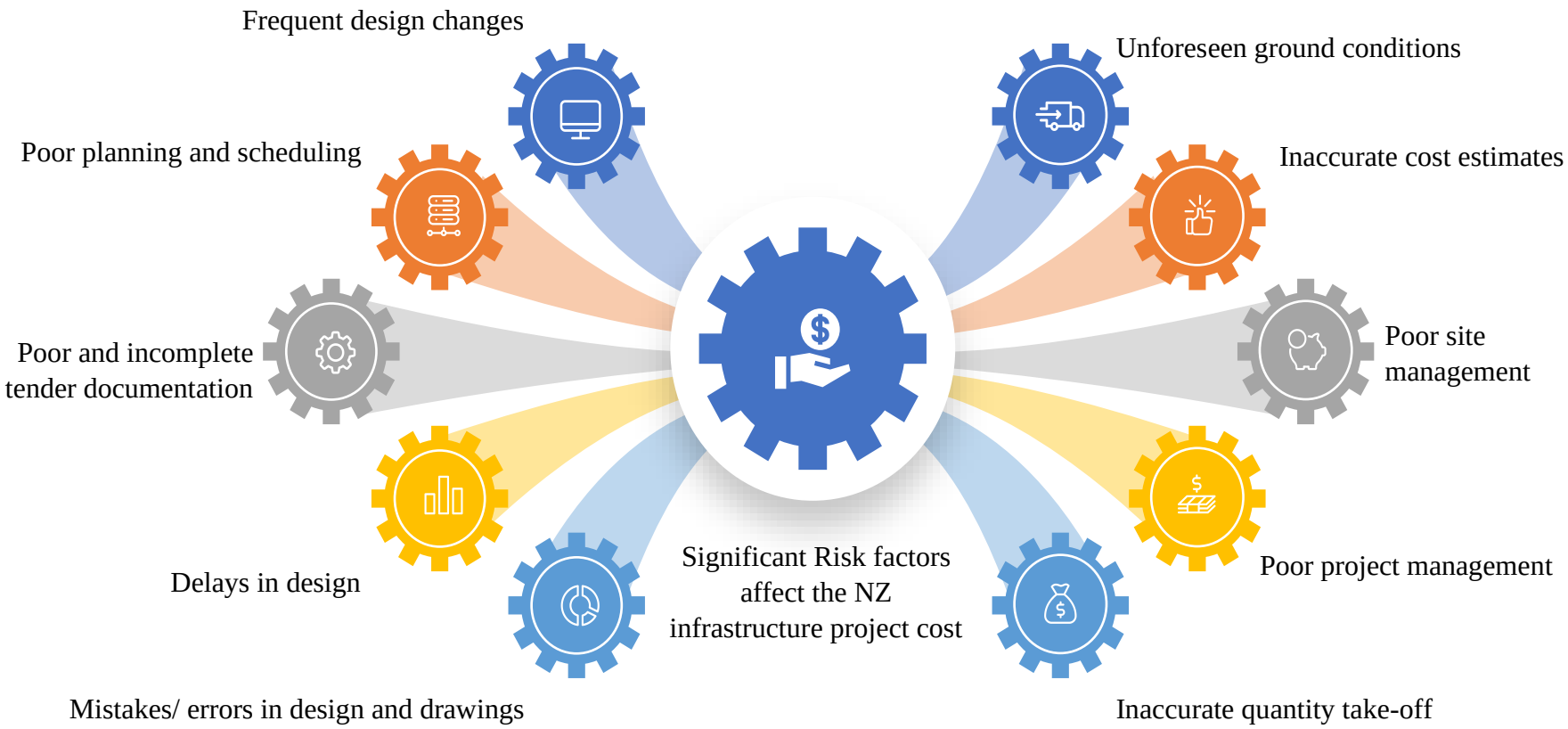

Figure 9.1. *Significant risk factors affect the NZ TI project cost (Source: Atapattu et al., 2023)*

Table 9.1. *Discussion of risk factors identified in the thesis against literature findings*

Factor	Significance according to the current research	Reference
Frequent design changes	'Lack of contractor experience', 'frequent design changes', 'economic stability' were the top three factors. To some extent their findings aligned with our study. However, their study found a lower ranking for 'frequent design changes' while our study identified it as a significant factor.	Ameh et al. (2010); Brunes and Lin (2014)
Poor planning and scheduling	Adequate project planning' is essential to avoiding cost overruns. However, they also argued that larger scale projects need the proper planning compared to smaller scale projects, to avoid cost overruns. According to our study 'poor planning and scheduling' achieved second rank. Therefore, it is very important to avoid cost overruns regardless of the project size. Andrić et al. (2019) emphasised that regardless of the project size the project planning is required to be carried out properly to avoid cost overruns.	Andrić et al. (2019); Contarelli et al. (2010); Flyvbjerg et al (2004); Harrera et al. (2020)
Poor and incomplete tender documentation	The findings are not directly applicable since the study was based on the factors affecting the tender price. However, 'incompetent tender document', 'design and construction complexity' and 'completeness of the project information' were ranked top three. Therefore, the factors are slightly connected to the pre-tender stage information.	Ji et al. (2014)
Delays in design	Although these factors were identified as risk factors that impact the project cost, none of the referenced studies recognised these as top significant factors. In contrast, this research identified these within the top ten factors. Based on the case study carried out in chapter 8 for the final model, the additional cost of these factors was in fact significant. Therefore, it is essential to consider them in the cost estimation.	Brunes and Lind (2014), Kumar (2020), Ameh et al. (2010), Abdul et al. (2013), Challal and Tkiouat (2012), Plebankiewicz (2018), Allahaim and Liu (2012), Britto and Perera, (2013)
Mistakes/ errors in design and drawings		Brunes and Lind (2014), Ji et al. (2014), Ameh et al. (2010), Abdul et al. (2013), Challal and Tkiouat (2012), Allahaim and Liu (2012), Britto and Perera (2013), Garry et al. (2006)
Unforeseen ground conditions		Brunes and Lind (2014), Adam et al. (2017), Ameh et al. (2010), Challal and Tkiouat (2012), Allahaim and Liu (2012), Park and Papadopoulou (2012)
Inaccurate cost estimates		Abdul et al. (2013), Adam et al. (2017), Ameh et al. (2010), Challal and Tkiouat (2012), Plebankiewicz (2018), Allahaim and Liu (2012)
Poor site management and supervision	'Price fluctuation', 'financial difficulties of the contractor', 'poor site management and supervision' were ranked among top factors. However, these are related to the post contract stage while our study suggested that the top-ranking factors are pre-design stage related.	Abdul et al. (2013)

Since these factors are qualitative risk factors, it is challenging to incorporate them into a technical variable-based traditional estimation process.

9.3 Cost overruns in NZ road projects

Cost overruns are a common global issue faced by all construction projects. Much research has been done to study the cost overruns in various construction projects. However, there is a lack of research into the NZ projects and no research on NZ road projects. Nevertheless, the magnitude of investment for the future of road projects in NZ is significant. Therefore, this study focussed on road projects to discover the issues related to cost estimation and overruns.

Researchers identified that the cost overrun issue has remained the same over the past seventy years in general construction projects. Further, the infrastructure projects face cost overruns of approximately 20% or more (Flyvbjerg et al., 2003). However, it was necessary to study the severity of this problem in NZ road projects before finding a solution. Therefore, data was collected from one hundred and six road projects completed during the past two decades, and the behaviour of the cost overruns was studied. According to Flyvbjerg et al.'s (2003) study, the MPCO of road projects was approximately 21%.

In contrast, a recent Odeck (2019) study observed that the MPCO of road projects increased by approximately 27%. On the other hand, Endut et al. (2009) studied all infrastructure projects together but public and private projects separately. According to their study, public projects faced 47% cost overruns while the private sector faced 37%. Since the road projects are mainly public, the cost overrun has become a severe issue. However, according to this research (chapter 4), the NZ road projects face approximately 20% cost overruns. Although the percentage is less compared to the literature findings, it is still a significant issue considering the time taken, the risk involved, and the usage of public funds.

Consequently, the study also identified that the cost overrun magnitude is not impacted by the year of project execution. In contrast, the project size and the duration are vital to the magnitude of the cost overrun. For this study, the project size was defined by the project's contract value, and the time was defined by the year of completion. Similar studies were carried out by Flyvbjerg et al. (2003) and Huo et al. (2018). The findings of this research were similar to Hu et al.'s (2018) research while contrasting findings were observed by Flyvbjerg et al. (2003). Subsequently, the study emphasised that a solution was needed to address the issues with cost estimation and cost overrun due to no improvement in the problem over the two decades studied.

Chapter 5 investigated the possible statistical methods to improve the conceptual estimation process. The following sub-section discusses the findings of Chapter 5.

9.4 Statistical techniques for better cost estimation process of infrastructure projects

A bibliometric literature review was conducted using the literature published on the statistical techniques used for cost estimation of construction projects. The analysis identified seven primary statistical techniques, namely, regression analysis (RA), artificial neural networks (ANN), support vector machine (SVM), Monte Carlo simulation (MCS), case-based reasoning (CBR), reference class forecasting (RCF), and fuzzy logic. These were the most frequently used techniques in cost forecasting modelling for construction projects. However, techniques such as CBR and RCF should have a robust cost database with many projects to produce reliable estimates. However, the other techniques can perform well with fewer projects, but identifying the correct independent variables is vital and can impact the model's performance.

Content analysis was carried out to evaluate the performance of each technique for infrastructure project estimation. The most suitable techniques for road projects were RA, ANN and SVM (Adel

et al., 2016; El-Sawalhi, 2015; Shr and Chen, 2006) because these techniques showed lower errors with higher performances. In addition, two models were identified and developed for water infrastructure projects with the lowest error rates. The speciality of these two models was that several statistical techniques were combined to improve reliability and accuracy. The techniques used for one model were ANN and fuzzy logic (Ahiaga-Dagbui et al., 2013), while the second model comprised RA and SVM (Ahiaga-Dagbui and Smith, 2014). Table 9.2 is an extract from Chapter 5, which discusses the performance of several cost models developed for infrastructure projects.

Table 9.2. *Cost models developed for infrastructure projects*

#	Modelling technique	Reference	Project type	Variables	MSE*
01	RA	Shr and Chen (2006)	Roads and highways	The final cost, awarded bid, days used, actual contract duration, and initial duration.	±5%
02	RA	El-Maaty *et al*. (2017)	Roads and highways	Inadequate project planning and execution, insufficient cost planning and monitoring, lack of communication between construction parties, price fluctuations, lack of proper technical study before the tender by the contractor, errors in project quantities measurements, slow decision-making process, equipment failures, lack of adequate field visits before tendering by the contractor, inappropriate use of project site, improper use of materials, inaccurate drawings and contract documents, material monopoly by suppliers, inflation, and rework.	+30.42%
03	RA	Sodikov (2005)	Roads and highways	Predominant work activity, project duration, pavement width, shoulder width, ground rise fall, average site clear, earthwork volume, surface class, and base material.	+30% to +36%
04	RA	Sonmez and Ontepeli (2009)	Urban railways	*Primary parameters* Percentage of tunnel section over the total length of the rail, percentage of the total length of elevated stations over the total rail length, percentage of the total length of at-grade stations over the rail length, percentage of the total length of cut-and-fill method over the main line length, supply and installation of the rails, and the number of underground stations. *Secondary parameters* Contract type, number of at-grade stations, number of elevated stations, the main line length, and percentage of the total length of depressed-open sections (ramps) to the total rail length.	+35.2%
05	ANN	Adel *et al*. (2016)	Roads and highways	Project scope, duration, year of construction, project region, mainline length, mainline classification.	Training - +4.51%; Evaluation- +5.8%;

					Validation - +16.0%
06	ANN	El-Kholy (2019)	Roads and highways	Inadequate project planning and execution, insufficient cost planning and monitoring, lack of communication between construction parties, price fluctuations, lack of proper technical study before the tender by the contractor, errors in project quantities measurements, slow decision-making process, equipment failures, lack of adequate field visits before tendering by the contractor, inappropriate use of project site, improper use of materials, inaccurate drawings and contract documents, material monopoly by suppliers, inflation, and rework.	+39.8%
07	ANN	Sodikov (2005)	Roads and highways	Predominant work activity, project duration, pavement width, shoulder width, ground rise fall, average site clear, earthwork volume, surface class, and the base material.	+24% to +26%
08	ANN	Sonmez and Ontepeli (2009)	Urban railways	*Primary parameters* Percentage of tunnel section over the total length of the rail, percentage of the total length of elevated stations over the total rail length, percentage of the total length of at-grade stations over the rail length, percentage of the total length of cut-and-fill method over the main line length, supply and installation of the rails, and the number of underground stations. *Secondary parameters* Contract type, number of at-grade stations, number of elevated stations, the main line length, and percentage of the total length of depressed-open sections (ramps) to the total rail length.	Model 1 - +49.8%; Model 2 - +33.3%
09	SVM	El-Sawalhi (2015)	Roads and highways	Road area, road surface type, base course type, base course thickness, interlock thickness, asphalt thickness, pipe diameter, manhole depth, cut and fill volume, curb length	-5% average error
10	Fuzzy logic	El-Maaty *et al*. (2017)	Roads and highways	Inadequate project planning and execution, insufficient cost planning and monitoring, lack of communication between construction parties, price fluctuations, lack of proper technical study before the tender by the contractor, errors in project quantities measurements, slow decision-making process, equipment failures, lack of adequate field visits before tendering by the contractor, inappropriate use of project site, improper use of materials, inaccurate drawings and contract documents, material monopoly by suppliers, inflation, and rework.	+40.37%
11	ANN and Fuzzy hybrid	Ahiaga-Dagbui *et al*. (2013)	Water infrastructure	Tendering strategy, site access, type of location, project type, Contractor's need, soil type, initial cost, and the initial duration.	+0.6% to +0.8%
12	RA, SVM and data mining hybrid	Ahiaga-Dagbui and Smith (2014)	Water infrastructure	Tendering strategy, procurement strategy, ground condition, soil type, delivery partner, scope, the purpose of the project, and location	-3.83% to +2.33%

*MSE – Mean Squared Error

Although it is evident that RA, ANN, SVM and hybrid models performed very well with the infrastructure projects, there are variables in these models that do not have enough information during the pre-design stage. Therefore, these models do not apply to the conceptual cost estimation. Consequently, this research applied these techniques to the NZ road projects to see the possibility of using these techniques to improve the current conceptual cost estimation practice. Subsequently, the following three sections discuss the model development for NZ road project conceptual estimation based on RA, ANN, and hybrid models.

9.5 Regression analysis as a cost modelling technique for NZ road projects

Under Chapter 6, the regression analysis (RA) technique was investigated in-depth to develop a cost estimation model using the concept for NZ road projects. Forty-three completed road projects completed between 2002 and 2022 were used to develop the model. RA-based models adopt four significant assumptions that will be tested later once the model is developed. They are linearity, normality, independence, and homoscedasticity. Data for the twelve independent variables were collected. The variables were road length, road width, distance from the nearest major city, the number of bridges, the approximate length of retaining walls, ground improvements area, pavement area, cut and fill area, expected project duration, expected year of completion, expected preliminary cost as a percentage of the project cost, and expected percentage change in construction cost index (CCI).

The backward elimination method was used to develop the model. In this method, every time RA was run, it eliminated the independent variable with an insignificant *p*-value. The best combination of the independent variables was model 6, with road length, road width, the number of bridges, pavement area, cut and fill area, expected preliminary cost as a percentage of the project cost, and expected percentage change in CCI. Later, the tests were run to test the four regression assumptions. It was observed that the model met all the regression assumptions.

On the other hand, model 6 achieved 0.593 of adjusted R-squared value, the highest value of all the eight models. Further, model 6 also achieved the lowest standard error of the estimate out of all eight models. Therefore, model 6 also passed the model fit test as well.

The next stage was to validate the model. This research used additional data from sixteen NZ road projects for model validation. Mean Absolute Percentage Error (MAPE) was used to measure the models' performance. All eight models were tested using the validation data set. Consequently, model 6 achieved 21.35% of the MAPE value. However, although the model achieved a good result among the eight models, MAPE is still considerably high. Therefore, concluding that the model can improve the current conceptual estimation practice is doubtful. Nevertheless, there might be other independent variables in the projects that have not been considered in this research. Table 9.3 compares similar RA models with the model developed in this study.

Table 9.3. *Findings discussion against similar RA models from literature*

Model	Comparison
Ou and Swarthout (1986)	The model cannot be used for conceptual cost estimation because some of the variables require tendering information, such as date of advertisement, lowest bid, percentage differences of the average bid/ lowest bid.
Kim et al. (2008)	Model showed lower error rate of 13.28% compared to our study. However, the cost data base was old in this model and needs upgrading. The model used projects from 1991 – 2001 and only twenty-seven cases were used, while our model used forty-three projects from 2002 – 2022.
Mahamid (2011)	The model reported that the MAPE ranged from 13% to 31%. However, the researcher reported that, out of the ten models with the highest accuracy they excluded the significant variables such as road length and width. However, in our study, model 6, which reported the lowest MAPE, considered all the significant variables during the conceptual stage. Regardless of the above, Mahamid (2011) models indicated higher adjusted R-squared values than our study.
Cirilovic et al. (2014)	The model predictions are based on the variables related to the country's economy rather than the technical attributes of the project. Project duration and road length were the only variables that defined the project characteristics. However, in our models, all the independent variables can be used to define the project. Considering the background of NZ's economy, the variables considered in Cirilovic et al. (2014)'s models may not significantly impact road projects over the technical aspects of the project.
El-Maaty et al. (2017)	The model showed an error more than 30%. Further to that the model was developed to predict the percentage of cost overrun. Therefore, it is not possible to use this as a conceptual cost model.
Zhang et al. (2017)	There were two models developed with error of only 7.1% and 7.6%. However, the models cannot be used to anticipate the conceptual cost estimate. For instance, there were variables such as contract price, construction spending, prime loan rate, and weather days. These factors are not easy to anticipate at the beginning of the project.
Lin and Techapeeraparnich (2019)	They also conducted similar research based on forty-four road projects in Thailand. Unlike most of the regression models discussed above, Lin and Techapeeraparnich (2019) took a similar approach to our study in finding the variables because the

	variables the model comprised of were road length and width, number of lanes, pavement type, earthworks, and miscellaneous work. The paper did not discuss the performance of the models in terms of error rates. However, in contrast to Dimitriou et al.'s (2018) findings, Lin and Techapeeraparnich (2019) observed that the regression model achieved better performance than the ANN-based model in terms of R statistics. On the other hand, compared to the R statistics of our models, Lin and Techapeeraparnich (2019) accomplished better R statistics in their model.
Ahmed (2021)	There were two models developed for roads and railways associated with tunnels. The model's accuracy was tested using the ratio of the predicted cost to the actual cost. The mechanised tunnelling roads/ railways-based model showed 0.782, while the conventional road and railways showed 0.768 as the accuracy ratio. But all eight models developed in this research (refer to chapter 6) achieved ratios closer to 1 compared to Ahmed's (2021) study results.

Based on the above discussion, although several cost models are developed using RA, most cannot be used for conceptual cost estimation as their target is the tender or post-contract stage—conversely, the model developed in this research aimed for conceptual cost estimation improvement. However, the RA-based model's performance was unsatisfactory, showing a mean error of 21.35%. According to the findings of Chapter 4, the cost overruns in NZ are approximately 20%, yet, in general, other countries also showed mean cost overruns of more than 20%. Therefore, chapter 7 investigated ANN as a modelling technique for conceptual cost estimation. The following section discusses the model developed using the ANN technique.

9.6 Artificial neural network as a cost modelling technique for NZ road projects

Since the RA-based model did not achieve significant results, as discussed in Chapter 7, the ANN technique was adopted to develop another conceptual cost model using the same data set. The independent variables for this model were also the same as the RA-based model. ANN is a technique similar to the human brain and the nervous system because the model uses previous experience and responds to complex problems. Typically, the ANN model contains several layers, mainly the input layer, one or two hidden layers, and the output layer. Input layers consist of the independent variables where the user can enter the data of the proposed project. Then, based on the inserted data, the ANN model runs the calculations using the previous data used for model

training and testing. These calculations happen in the hidden layers. Finally, the output layer gives the output for the proposed project, and in this research, the output was the final cost of the NZ road projects.

The study developed six ANN-based models by changing the number of variables, neurons in the hidden layers, and activation function (Tanh or sigmoid). During the model development, two main stages of testing were carried out: the training and testing phases. Out of all six models, model 2 achieved better results than the others. The R-squared values of the training and testing phases were 0.977 and 0.993, respectively. Further, the Root Mean Square Error (RMSE) value of each phase was 0.279 and 0.245, respectively. Although a few models have achieved higher R-squared values than model 2, the lowest RMSE value of the training phase was achieved by model 5.

However, to justify the selection of model 2, it is essential to validate the model. Therefore, similar to Chapter 6, the same validation data set was used to validate the ANN-based model. As expected, the lowest MAPE value, which is 11.82%, was achieved by model 2. Surprisingly, model 2 achieved the highest R-squared value for the validation phase, 0.877. All the other five models showed R-square values less than 0.8. Therefore, based on the model statistics, model 2 is a perfect model for the cost estimation of road projects. Table 9.4 compares the model performance with similar ANN-based models identified in the literature.

Table 9.4. *Findings discussion against similar ANN models from literature*

Model	Comparison
Hegazy and Ayed (1998)	Backpropagation training-based model – Error 10.4%. Simplex optimisation-based model – Error 1% Genetic algorithms optimisation – Error 21.8% However, the model was developed using Excel in 1998. Therefore, applicability of the model for today's construction industry is doubtful without re-testing.
Sodikov (2005)	Developed ANN-based model using data set from two different countries. However, the error rates were higher than the results of our study (which were 24% and 26%).
Elbeltangi et al. (2014)	Model achieved a much less error of 2.86%. Although the research was done in 2014, the data base used for the model development was between 2001 and 2005. Therefore, modern application of the model needs re-testing. Out of the independent variables, soil type and financial condition are the two variables that cannot be utilised for conceptual cost estimates.

Cirilovic et al. (2014)	Models used nineteen variables, which is a bit excessive. Mainly the variables such as road sector gasoline fuel consumption, transparency international corruption perception index were also used as variables. It is difficult to understand the connection between such variables and the construction cost of the road project. Further, there are several variables that have no information during the pre-design stage. At the end, the models with highest performance contained only oil prices, climate conditions, road sector gasoline fuel consumption, transparency/ international corruption perception index, and world governance index.
Adel et al. (2016)	The best model achieved 16% MAPE at the validation phase. The variables considered were also relevant to the pre-design stage. Therefore, the model can be used for conceptual estimates. Nevertheless, the model developed in our study achieved less error compared to this model.
Mahalakshmi and Rajasekaran (2018)	Their model was able to predict the outcome with a MAPE of 8.46%. This performance is much higher compared to our study. However, their model considered: topography, pavement type, soil condition, and drains. However, these were not considered in our model, because for the conceptual estimation stage, these variables' information is not available.
El-Kholy (2019)	All the models showed MAPE higher than 25%. It was highlighted in these models that out of the fifteen variables used, most of them are qualitative and risk-related factors. It was further emphasised that ANN cannot ascertain the project risks alone when risk factors are incorporated into the model as independent variables.
Jaafari et al. (2021)	Model showed a percentage of relative error index of 0.038%. Compared to most of the models developed for road projects, this model achieved an outstanding performance. However, this model considered variables such as number of trees, size of trees, cut and fill slopes, drainage, and culvert details. These variables cannot be used for conceptual estimate as there will be no details available at pre-design stage of the project.

Based on the literature comparison, the model developed in our research provides good insights into the current body of knowledge. Subsequently, the conceptual cost estimation models identified in the literature were either outdated, since some were developed a long time ago or contain variables that cannot be considered during the pre-tender stage, or some variables do not have an excellent connection to the project cost but yet identified as crucial variables, or the models showed a high rate of error. In contrast, our ANN model fills this research gap by providing a conceptual cost estimate model with a considerably low error rate.

The performance of the ANN-based model (model 2 – chapter 7) achieved better performance than the RA-based model (model 6 – chapter 6). Nevertheless, the MAPE is still higher than expected. The reason could be that none of the risk factors identified in Chapter 4 were considered in either model. Therefore, other independent variables can affect the project cost but are not considered here since the twelve variables considered were the maximum data that could be extracted from

the projects. Hence, the following section discusses the final model developed in this research to solve the above issue.

9.7 A conceptual cost estimation model for pre-design stage of NZ road projects – A hybrid of combining artificial neural network and Monte Carlo simulation

In Chapter 8, the investigation was conducted to incorporate the risk expectations of the project into the model. That is because the traditional costing methods do not consider these methods. MCS is an excellent technique to model the uncertainty of a project; however, the base estimate (the estimate before adding the risk) must be provided to MCS to generate the uncertainty. Therefore, in Chapter 8, ANN was combined with MCS. ANN produced the base estimate of the project using the model developed in Chapter 7, with a MAPE of 11.82%. Estimating the expected risk component is necessary to identify the risk. Therefore, the risks identified in Chapter 4 were used in Chapter 8 to estimate the risk component in the three measures: a low-risk estimate, a most likely estimate and a high estimate. Since the MCS calculates the risk based on the probability of occurrence, there can be unlimited possibilities for the risk component combination. However, two probabilities were considered and compared with the actual cost of the sixteen validation cases used in chapters 6 and 7. The possibilities were P50 (50% probability) and P95 (95% probability). Subsequently, the P50 estimate was the most reasonable and closest to the actual cost of the validation cases, and it achieved a MAPE of 3.53%.

Compared to the models developed using a single technique, as discussed and compared in the above sections 9.4, 9.5, and 9.6, the hybrid model developed in Chapter 8 shows significantly better results. Further, this model can be easily used for conceptual cost estimation of road projects. Chapter 5 only identified two hybrid models (Ahiaga-Dagbui et al., 2013; Ahiaga-Dagbui and Smith, 2014). One model combined ANN with fuzzy logic, and the other combined RA, SVM and

data mining. Both models achieved significantly high performance as errors were less than 1%. However, the models were developed for water infrastructure projects.

In contrast, this thesis developed a model using ANN to calculate the base estimate and combined it with MCS to calculate the uncertainty with a good performance level. According to the literature, our model is the first hybrid model combined with ANN and MCS to accommodate both the technical side of the project and the uncertainty side of the project into one model for road projects. Although the model was developed using NZ case study data, the variables were selected carefully to allow the model's adaptability in other contexts.

9.8 Epilogue

This chapter discussed the achievement of the overall research aim by combining the learning from each chapter. The next chapter discusses the conclusions of this thesis and recommendations for further research.

10 Conclusions and recommendations for further research

10.1 Prologue

This chapter presents the research overview, followed by a discussion on achieving the research objectives. The original contribution of this study to the construction financial management domain is also presented. The chapter then explains the study's limitations and recommends further research studies.

10.2 Research Overview

This research aimed to develop a cost model for the early design stage of road projects in NZ based on statistical analysis techniques. At this stage, there is limited information available on the proposed projects. Further, it is challenging for traditional estimation methods to provide reliable cost estimates. Therefore, the traditional cost estimation methods need upgrading. The research gap aims, objectives and questions were established in light of this. Each objective was investigated through one journal paper, comprising six journal papers. This thesis was structured into ten chapters: introduction, research methodology, literature review, data collection, analysis and findings, final model, discussion, conclusion, and recommendations.

Chapter 1 studied the background of the research topic, identified the research gap and established the research aim, objectives, and questions. This chapter outlines nine objectives to be met to achieve the overall aim of the research. Those objectives were in chronological sequence, and to meet each objective, a continuous step-by-step process was conducted, as discussed in the following sections.

Chapter 2 explained the research methodology adopted for this research. It started with exploring the nature of the research aim, research questions and objectives. Based on this, the chapter defined

the research philosophy applicable to this research. Subsequently, the selected philosophy identified the suitable methodology for the research.

Chapter 3 explored the possible risk factors affecting the cost of infrastructure projects. As background and research gap analysis suggests that the current estimation process needs improvement, this chapter investigates the causes or the factors that must be considered in the estimation process. Initially, a systematic literature review was conducted to identify the main factors globally accepted as the causes of cost overruns. Then, a questionnaire survey was conducted to validate and prioritise the NZ-specific factors. The chapter finally concluded that these significant risk factors should be incorporated into the cost model to improve reliability.

Chapter 4 investigated the cost overruns in road projects in NZ. Although Chapter 1 of the background study identified that cost overruns are a global phenomenon, it is essential to identify the severity of this problem within the NZ context. Actual data from the road project in NZ were collected and analysed to identify the issues of cost overruns. The chapter identified that cost overrun is a significant issue in NZ road projects and needs a solution regardless of the project size, such as small-scale, medium-scale, or large-scale.

Chapter 5 conducted a bibliometric literature review to identify the significant cost modelling techniques used for cost forecasting models. That is because Chapter 3 concluded that the current estimation process needs upgrading, while Chapter 4 recommended that the estimation process should be able to address the significant risks in the model. However, traditional estimation models do not consider risk factor analysis. Therefore, the findings of this chapter are highly significant as this will provide the background to connecting the research problem with the solution.

Chapters 6 and **7** developed two cost models using RA and ANN. These two techniques were identified in Chapter 5 as practical techniques in road project estimation. Several models were developed using RA and ANN in these two chapters using the actual cost data collected from NZ

road projects completed within the last decade. Later, the models' performances were evaluated, tested, and validated using additional case data and recommendations were made for further research. Based on the performance, it was observed that ANN can provide a more reliable estimate than RA within the considered parameters. Nevertheless, these two models did not consider the risk associated with the project.

Chapter 8 considered the risk component of the project within the estimation by combining MCS with ANN. The estimations produced through the ANN model developed in Chapter 7 were run through MCS to incorporate the risk component into the estimation. The RA-based model was not considered for this since the model's performance was lower than the ANN-based model. Finally, the recommendations were developed for further research based on the hybrid model, which is the overall output of this research.

Chapter 9 discussed all the chapters as contributors towards one specific research aim.

Chapter 10 is the last chapter of the thesis, providing conclusions of the research, contribution to academia and the industry and recommendations for future research.

The following section discusses the research content, the chapters and the achievement of the research objectives in chapter 1.

10.3 Research objectives achievement

This thesis aimed to improve the accuracy of the current early-design stage cost estimation practice used in NZ road projects. Four main research questions were formed in order to achieve the overall aim. Subsequently, those four research questions were subdivided into eight research objectives. The main chapters of this research, Chapters 3, 4, 5, 6, 7, and 8, focussed mainly on achieving these eight objectives.

The first research question (RQ1) was, "What are the primary risk factors affecting the final cost of the NZ road projects?" This research question was answered through two objectives. RO1 explored the risk factors affecting the cost of construction projects, while RO2 distinguished the significant risk factors specific to NZ road projects. RO1 was achieved through the systematic literature review. Fifty-three risk factors were identified under ten subcategories of overall construction projects. Later, a questionnaire survey was conducted to validate, prioritise, and identify the factors significant to the NZ road projects. Ten significant factors were identified, mainly concerning the qualitative risk factors rather than technical factors addressed in the current estimation practice. Therefore, this chapter recommended incorporating these risk factors into the estimation process.

The second research question (RQ2) was, "Is the cost overrun a significant issue in NZ road projects?". In order to answer the above question, two objectives were formulated. The third objective (RO3) was identifying the severity of the cost overrun issues in NZ road projects. In addition, RO4 was to investigate whether the cost overrun depends on the project size or the time. Both objectives were covered in Chapter 4. The background literature showed that cost overrun is a global phenomenon without any improvement over the past seventy years. Therefore, the actual cost data from the completed one hundred and six road projects in NZ for the past two decades were used to investigate this problem within the NZ context. Based on the data analysis, it was noted that cost overrun is significant in NZ road projects regardless of project size and time. Hence, it was concluded that the current estimation practice needs improvement.

The third research question (RQ3) was, "Which modelling techniques can be used to develop a prediction model for cost estimation of transportation infrastructure projects?" Similar to the previous two questions, there were two objectives in this question, and both were covered in chapter 5. RO5 was to examine the statistical analytical techniques used for developing cost

models for construction projects. Then, RO6 was to investigate appropriate conceptual cost modelling techniques to use at the pre-design stage to ensure more accurate out-turn cost prediction in transportation infrastructure projects. A bibliometric literature review was conducted to achieve both objectives. Seven statistical techniques were identified and used in the construction industry to develop cost models. These were regression analysis (RA), artificial neural network (ANN), support vector machine (SVM), Monte Carlo simulation (MCS), Case-based reasoning (CBR), reference class forecasting (RCF), and fuzzy logic. RA, ANN, SVM, and hybrid models combined several techniques to achieve higher performance with lower errors. It was recommended that these techniques be examined in-depth to develop better cost estimation practices.

The fourth and final research question (RQ4) was, "Can modelling techniques be used to develop a reliable cost estimation model for NZ road projects?" In order to answer the above question, three objectives were established. RO7 was to develop and validate the best cost model for conceptual estimation for the pre-design stage of road projects in NZ. Therefore, based on the findings from Chapter 5, chapter 6 carried out a case study on NZ road projects and developed a cost model using the RA technique. This objective was achieved through document analysis. Although the model could predict the cost, the Mean Absolute Percentage Error (MAPE) was 21.35%. Therefore, the reliability of the model is not significant. Hence, under the same objective (RO7), another experiment was conducted using ANN in chapter 7. Based on the analysis, the developed model achieved a lower error rate of 15.34% compared to RA. However, this performance is not yet satisfactory for the NZ road projects, considering the significance of cost overruns in such projects.

Subsequently, in Chapter 8, RO7 and RO8 were combined to examine if the model performance can be improved by combining several methods. In RO2, it was emphasised that project risks are significant cost estimation and that this component should be incorporated into the estimation

process. However, RA and ANN did not consider any risk factors. Further, in Chapter 5, it was revealed that those techniques did not perform well with risk-related variables. Hence, chapter 8 investigated MCS as a risk estimation technique that can combine with either the RA-based or ANN-based models. However, since the ANN-based model showed lower error than the RA-based model, this chapter only considered the ANN model in combination with MCS. The findings revealed that by combining ANN and MCS, the MAPE can be reduced from 15.34% to 3.53%. That is a significant achievement and contribution to knowledge in the domain of construction project estimation. Finally, the chapter also made recommendations for further improvement of this result.

Figure 10.1 summarises the above discussion and elaborates on the achievement of the objectives against the content of each chapter.

The following section discusses the contribution of this research to academia and the industry.

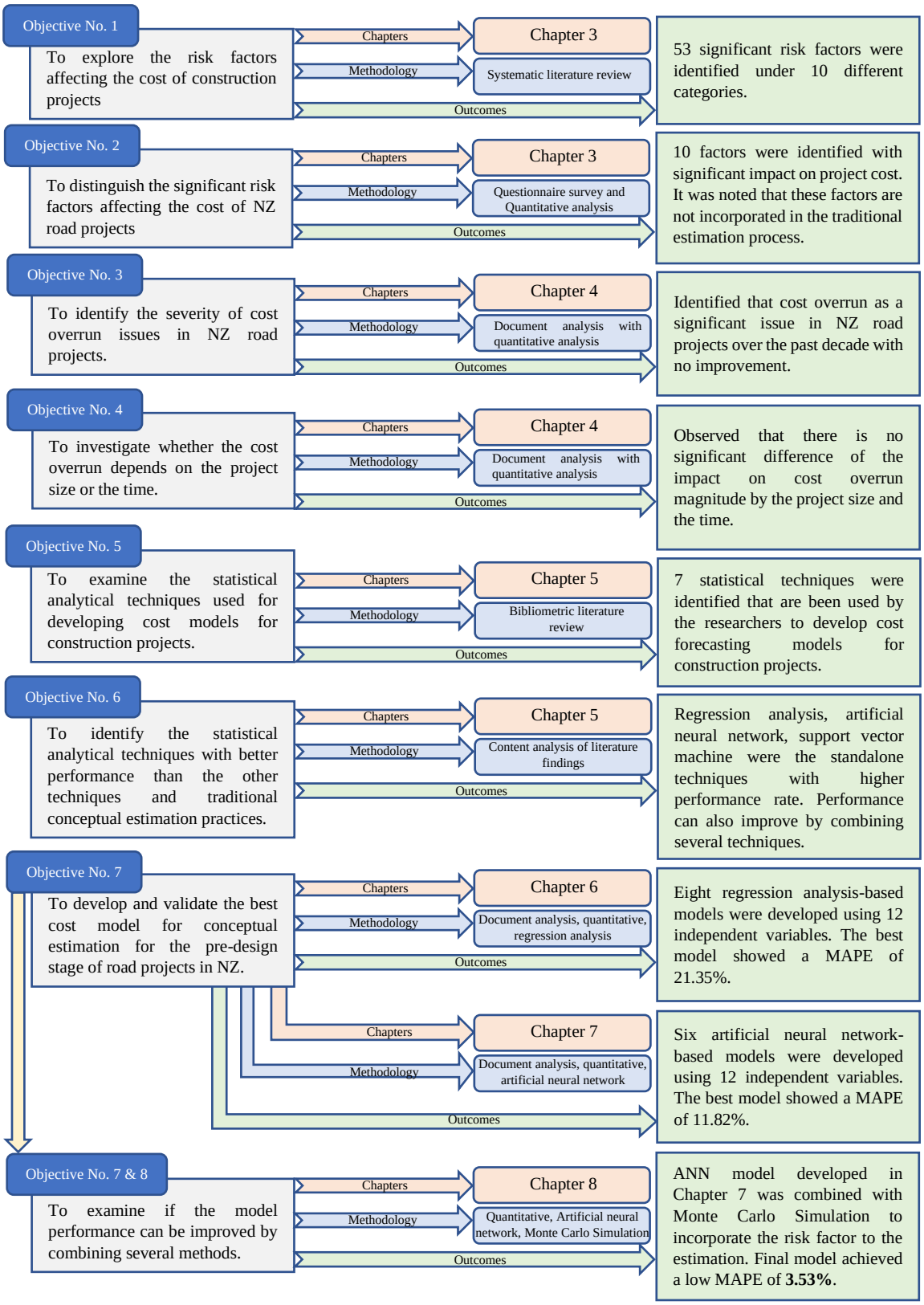

Figure 10.1. *Achievement of research objectives*

10.4 Research contribution

This thesis contributes to the existing body of knowledge on the conceptual cost estimation practice in road projects. The study focuses on improving the issues in the current traditional estimation process in NZ road projects by adopting cost modelling techniques into the cost model. The contributions of this research are outlined and discussed below.

10.4.1 Theoretical contribution

The significant theoretical contribution of this thesis is the development of the cost estimation model for the early design stage of road projects. The study findings establish the necessity of resolving the concerns of cost overruns. Further, it also emphasises that the current conceptual estimating techniques need upgrading. Based on past research, this study provided a list of risk factors which affect the TI project cost and possible statistical techniques for cost estimation. The theoretical findings of this research were further tested and validated through a questionnaire survey and document analysis to provide recommendations for the construction domain of the research. The research identified significant statistical techniques and compared the performances to identify the most robust techniques in predicting the reliability of the cost for construction projects. Therefore, the research contributes to the theory by comparing and suggesting reliable techniques for cost estimation and recommending ways to improve their performance and reliability.

Furthermore, according to the literature, this study is the first investigation conducted using ANN and MCS to explore the possibility of combining them as one model. Consequently, models have been developed to accommodate risk factors and technical factors. However, the modelling technique used in these studies was only one of the seven techniques identified in Chapter 5. Nevertheless, the techniques could perform better with technical and risk variables. In contrast, this study brings new knowledge to the existing studies by combining two techniques to calculate

technical and risk variables separately. In addition, this study aimed to overcome the drawbacks of each technique using the advantages of the other technique. Therefore, this thesis significantly contributes to the body of knowledge with this new hybrid model.

The following section discusses the research contribution to the industry.

10.4.2 Contribution to the industry

Firstly, this research provides an extensive list of risk factors from the Quantity Surveyor's perspective that are significant to the infrastructure projects. Further, the list was prioritised based on the severity of the impact on NZ road projects. This list can also be incorporated into the project risk analysis. The top factors to consider in TI projects are frequent design changes, poor planning and scheduling, poor and incomplete tender documentation, delays in design, mistakes/ errors in design and drawings, inaccurate cost estimates, unforeseen ground conditions, poor site management and supervision, poor project management, inaccurate quantity take-off. During the conceptual cost estimation, significant attention must be given to the risk involved in these factors.

Secondly, this research gives statistical evidence to industry experts that road projects in NZ face significant cost overruns regardless of the project size and the time of project execution. Further, this research also evidenced that there has been no significant improvement in the cost estimation practice. Hence, this research is a background study for the industry to start looking for a solution.

Thirdly, the research identified seven cost modelling techniques that researchers have used to develop cost estimation models. Therefore, industry experts can investigate these techniques to improve the current estimation practice. Moreover, the research shortlists the best techniques suitable for infrastructure projects, as this research aims to improve the estimation practice of road projects in NZ.

Fourthly, the final model developed for NZ road projects addresses less information availability and risk analysis issues. Therefore, the industry can directly use the model to improve their estimation accuracy and reliability at the early design stage. According to the findings, the NZ road construction industry is experiencing a continuous cost overrun of approximately 20%. However, the output of this research contributes to the industry by reducing the error in estimation from approximately 20% to below 4%.

Chapter 8 shows how the final model combines the ANN-based estimation and MCS-based risk component. The ANN-based estimation consists of technical variables only. The risk is considered only when it combines with the MCS. The estimate developed through ANN shows an 11.82% error. Then, the MCS adds the risk probability, which could be either an addition or a deduction. Generally, the conceptual cost estimation is communicated between the client and the cost consultant. If the overall budget is communicated to the Contractor, the tender prices will likely be closer to the budget. Sometimes, if the budget was not disclosed during the tender, it will be disclosed after the tenderer selection. In that case, the Contractor knows the client has funds for the allocated budget. Therefore, when contractors deal with risk mitigation, they will only consider the budget set at the beginning. By doing this, there is the possibility of experiencing further unexpected cost overruns. Therefore, when the budget is disclosed to the Contractor, it is recommended that the lower budget figure be disclosed based on one of the following situations.

1. Generally, if the risk experienced in the project is significantly positive (threats are more significant than opportunities) then the ANN-based estimate is less than the final estimate. In this circumstance, the ANN-based estimate should be disclosed to the Contractor as the budget. The remaining should be kept in the record only for the client as a contingency to address the risk situations.

2. Sometimes, projects experience adverse risks (opportunities are more significant than threats). Therefore, the ANN-based estimate can be higher than the final risk-based estimate. In this situation, it is advisable to disclose the risk-based final estimate to the Contractor as the budget limit. The remaining is the client's contingency.

Further, the findings of this research are recommended to be added as guidelines for the estimation and risk management manuals. The research identified ten significant risk factors for NZ road projects. The risk management team and the Cost management team should collaborate at the beginning of the project to discuss the possible impact of these risk events on the project. That will enable risk incorporation into the estimation and minimise unforeseen cost overruns.

The estimation guidelines of the NZ government agencies can be updated with the model developed in this research.

1. First, gathering data from the proposed project for the twelve variables identified in this model is required. Then, run the Artificial Neural Network model for those variables. The outcome of this would be the base cost estimate for the proposed project.

2. Take the outcome of the risk analysis. That should generate the possible risk component for the proposed project.

3. Combine the above two outcomes in the Monte Carlo Simulation model and identify the most feasible combination of Base cost estimation and the risk component. Generally, the P50 outcome is the estimation derived from the MCS model.

4. The outcome will have a 3.5% error. Further, it is also necessary to take note of the difference between the P50 Estimate and the P95 estimate because if any unforeseen risks in the proposed project were not identified during the feasibility stage, they would be covered by this additional reserved risk contingency.

Subsequently, the research provides further recommendations in the next section to improve the results of this research. Hence, the output of this research can be combined with those recommendations to improve the estimation practice further.

The following section discusses the research limitations and recommendations for further research.

10.5 Research limitations and recommendations for further research

Several limitations were identified and acknowledged during this research. The constraints encountered in accomplishing this thesis are listed in this section, and recommendations are provided for future research.

1. One of this study's limitations is identifying the main risk factors of cost overruns in NZTI projects. All identified factors were based on the systematic literature review. More factors could have been identified using other sources, such as interviewing the construction experts, surveying construction practitioners, or exploring the NZ construction industry's previously completed cases. Although much research is available in the literature, there may be NZ-specific factors that should have been considered in other countries. That can only be identified through actual data collection methods.

2. The second limitation is that the participants of the questionnaire survey conducted to prioritise and validate the risk factors identified were only Quantity Surveyors and Estimators. Therefore, the risk factors identified may have a Quantity Surveyor bias, as they would look at risk factors from the project cost perspective. However, if the survey was open to all construction professionals in NZ, the results may vary with different perspectives on the problem. Therefore, the reliability of the data would also improve.

3. The third limitation is that this research is focused on NZ road projects. However, initially, it was identified that the estimation process of TI projects in general needs improvement.

Therefore, this model can be further developed to accommodate other TI projects. Further, the vertical sector projects may also need such a statistical cost model, with further research in NZ as such research has been done in other countries.

4. The fourth limitation is that the output of this research is based on the investigation of three statistical techniques: regression analysis, artificial neural network, and Monte Carlo simulation. However, other techniques identified in the earlier chapters may also be helpful in developing a better model. On the other hand, other cost modelling techniques may not be identified in this research but can provide robust estimates.

5. The fifth limitation is that the results of the models developed in this research were based on the twelve independent variables identified. However, there may be other variables that were not identified but could significantly impact the project cost. Therefore, further investigations into projects using in-depth case studies and expert ideas can improve the models for more reliable estimates.

6. The sixth and final limitation is the number of cases used in this research. The outcome and the reliability of the outcome highly depend on the number of cases. That is because when the number of cases used for data to develop the model increases, the model's accuracy and performance will also increase. However, the models developed in the research were based on forty-three cases for model development and sixteen cases for model validation. However, models can be further developed by increasing the number of projects used in the model development.

10.6 Epilogue

Overall, the cost estimation at the early design stage is significant to the project decision-making process. Therefore, high accuracy and reliability of such estimates are required. This matter is even

more significant in road projects than vertical sector construction. However, the current estimation practice is not able to sustain this matter. Hence, this research addressed this gap through in-depth research into statistical techniques and cost models. The final hybrid model can be concluded as the final achievement of the study to be used for future research in NZ.

References

AACE International. (2020). Cost estimate classification system – TCM framework 7.3 – cost estimating and budgeting, available at https://web.aacei.org

Abdul, R. I., Memon, A. H., Karim, A. and Tarmizi, A. (2013). Significant factors causing cost overruns in large construction projects in Malaysia. *Journal of Applied Science,* 13(2)**,** 286-293.

Abebe, A. J., Guinot, V. and Solomatine, D. P. (2000). Fuzzy alpha cut vs Monte Carlo techniques in assessing uncertainty in model parameters. In *Proceedings of the 4th International Conference on Hydro-informatics.* Lowa City, USA.

Adafin J., Rotimi, O. B. and Wilkinson, S. (2020). An evaluation of risk factors impacting project budget performance in New Zealand. *Journal of Engineering, Design and Technology*, 19(1), 41-61.

Adafin. J. K., Wilkinson, S. J., Rotimi, J. O. B. and Odeyinka, H. A. (2016). Evaluating the budgetary reliability of design stage elemental cost plan in building procurement: New Zealand study. In: A. O. Windapo, S. J. Odediran, A. Adediran (Ed.), *Proceedings of the 9th CIDB postgraduate conference* (pp. 60-70); Feb 1-4; Cape Town (SA): University of Cape Town.

Adam, A., Josephson, P. E. B. and Lindahl, G. (2017). Aggregation of factors causing cost overruns and time delays in large public construction projects Trends and implications. *Engineering Construction and Architectural Management,* 24(3), 393-406.

Adel, K., Elyamany, A., Belal, A. M. and Kotb, A. S. (2016). Developing a parametric model for a conceptual cost estimate of highway projects. *International Journal of Engineering Science*, 6(7), 1728-1734.

Adel, K., Elyamany, A., Belal, A. M. and Kotb, A. S. (2016). Developing a parametric model for a conceptual cost estimate of highway projects. *International Journal of Engineering Science*, 6(7), 1728-1734.

Adeli, H. and Wu, M. (1998). Regularisation neural network for construction cost estimation. *Journal of Construction Engineering and management*, 124(1), 18-24.

Agnieszka, L. and Krzysztof, Z. (2018). Cost calculation of construction projects including sustainability factors using the case-based reasoning (CBR) method. *Sustainability,* 10(5), 1608-1608.

Ahiaga-Dagbui, D. D., Love, P. E., Smith, S. D. and Ackermann, F. (2017). Toward a systemic view to cost overrun causation in infrastructure projects: A review and implications for research. *Project Management Journal*, 48(2), 88-98.

Ahiaga-Dagbui, D. D. and Smith, S. D. (2014). Dealing with construction cost overruns using data mining. *Construction Management and Economics,*32(7-8), 682-694.

Ahiaga-Dagbui, D. D. and Smith, S. D. (2014). Dealing with construction cost overruns using data mining. *Construction Management and Economics,* 32(7-8), 682-694.

Ahiaga-Dagbui, D. D., Tokede, O., Smith, S. D. and Wamuziri, S. (2013). A neuro-fuzzy hybrid model for predicting the final cost of water infrastructure projects. *Proceedings of the 29th Annual Association of Researchers in Construction Management (ARCOM) Conference*, 2-4 September, Reading, UK, available at: http://www.arcom.ac.uk/-docs/proceedings/ar2013-0181-0190_Ahiaga-Dagbui_Tokede_Smith_Wamuziri.pdf

Ahmed, C. (2021). Early cost estimation models based on multiple regression analysis for road and railway tunnel projects. *Arabian Journal of Geosciences,* 14(11), 972.

Ahmed, S., Memon, A. H., Memon, N. A., Laghari, A. N., Akhund, M. A. and Imad, H. U. (2018). Common factors of cost escalation in the construction industry of Pakistan, *Engineering Technology and Applied Science Research,* 8**,** 3508-3511.

Ahn, J., Park, M., Lee, H. S., Ahn, S. J., Ji, S. H., Song, K. and Son, B. S. (2017). Covariance effect analysis of similarity measurement methods for early construction cost estimation using case-based reasoning. *Automation in Construction,* 81, 254-266.

Akinradewo, O., Aigbavboa, C., Oke, A., Coffie, H. and Ogunbayo, B. (2022). Unearthing causavative factors of cost overrun on Ghanaian road projects. *Baltic Journal of Road and Bridge Engineering*, 17(4), 171-188.

Al Amri, T. and Marey-Perez, M. (2020). Towards a sustainable construction industry: Delays and cost overrun causes in construction projects of Oman. *Journal of Project Management,* 5(2)**,** 87-102.

Al Hosani, I. I. A., Dweiri, F. T. and Ojiako, U. (2020). A study of cost overruns in complex multi-stakeholder road projects in the United Arab Emirates. *International Journal of System Assurance Engineering and Management*, 11(6), 1250-1259.

Alghonamy, A. (2015). Cost overrun in construction projects in Saudi Arabia: Contractors' perspective. *International Journal of Mechanical and Mechatronics Engineering,* 15(4)**,** 35-42.

Alhammadi, A. S. A. M. and Memon, A. H. (2020). Inhibiting factors of cost performance in UAE construction projects. *International Journal of Sustainable Construction Engineering and Technology,* 11(2)**,** 126-132.

Al-Hazim, N. and Salem Z. A. (2015). Delay and cost overrun in road construction projects in Jordan. *International Journal of Engineering and Technology*. 4(2), 288–293.

Ali, Z., Ahmad, I. and Hussain, Z. (2020). Analysis of critical causes of transaction cost escalation in public sector construction projects. *Pakistan Journal of Commerce and Social Science,* 14(4)**,** 838-865.

Alinaitwe, H., Apolot, R. and Tindiwensi, D. (2013). Investigation into the causes of delays and cost overruns in Uganda's public sector construction projects. *Journal of Construction in Developing Countries,* 18(2)**,** 33-47.

Allahaim, F. S. and Liu, L. (2012). Understanding major causes cost overrun for infrastructure projects: a typology approach. In *Proceedings of the 37th annual conference of the Australasian Universities Building Educators Association (AUBEA)*, 4-6 July, The University of New South Wales, Australia.

Alnuaimi, A. S., Taha, R. A., Al Mohsin, M. and Al-Harthi, A. S. (2010). Causes, effects, benefits, and remedies of change orders on public construction projects in Oman. *Journal of Construction Engineering and Management,* 136(5)**,** 615-622

Amadi, A. and Higham, A. (2019). Cognitive Mapping of Geotechnical Practices as Cost Overrun Drivers in Highway Projects. *Engineering Project Organisation Journal, 8*(1), 1-23.

Ameh, O. J., Soyingbe, A. A. and Odusami, K. T. (2010). Significant factors causing cost overruns in telecommunication projects in Nigeria. *Journal of Construction in Developing Countries,* 15(2)**,** 49-67.

Ammar, T., Abdel-Monem, M. and El-Dash, K. (2022). Risk factors causing cost overruns in road networks. *Ain Shams Engineering Journal*, *13*(5), 101720.

Andrić, J. M., Mahamadu, A. M., Wang, J., Zou, P. X. W. and Zhong, R. (2019). The cost performance and causes of overruns in infrastructure development projects in Asia. *Journal of Civil Engineering and Management*, 25(3), 203-214.

Anish, C., Kiruthiga, K., and Vinoth, S. (2019). Analysis of time delay and cost overrun in road construction. *International Journal of Innovative Technology and Exploring Engineering*, *8*(9 Special Issue 3), 901-907.

Annamalaisami, C. D. and Kuppuswamy, A. (2019). Reckoning construction cost overruns in building projects through methodological consequences. *International Journal of Construction Management*, 22(6), 1079-1089.

Ashtari, M.A.; Ansari, R.; Hassannayebi, E.; Jeong, J. (2022). Cost Overrun Risk Assessment and Prediction in Construction Projects: A Bayesian Network Classifier Approach. *Buildings*, 12(10), 1660.

Atapattu, C. N., Domingo, N. D. and Sutrisna, M. (2022). Statistical cost modelling for preliminary stage cost estimation of infrastructure projects", *IOP Conference Series: Earth and Environmental Science,* 1101, 052031.1-10.

Atapattu, C. N., Domingo, N. D. and Sutrisna, M. (2023). Significant factors affecting the New Zealand transportation infrastructure project cost – quantity surveyors' perception, *Built Environment Project and Asset Management,* 13(5), 756-777.

Auckland Transport, (2013). Auckland transport code of practice. Available on https://at.govt.nz

Augustin, N. H., Sauleau, E. A. and Wood, S. N. (2012). On quantile-quantile plots for generalised linear models. *Computational Statistics and Data Analysis*. 56(8), 2404-2409.

Awojobi, O. and Jenkins, G. P. (2016). Managing the cost overrun risks of hydroelectric dams: An application of reference class forecasting techniques. *Renewable and Sustainable Energy Reviews,* 63, 19-32.

Balali, A., Moehler, R. C. and Valipour, A. (2020). Ranking cost overrun factors in the mega-hospital construction projects using Delphi-SWARA method: an Iranian case study. *International Journal of Construction Management,* 22(13), 2577-2585.

Barlow, G., Tubb, A. and Riley, G. (2017). *Driving Business Performance - Project Management survey 2017*, Wellington, New Zealand: KPMG New Zealand.

Bayram, S. and Al-Jibouri, S. (2016). Application of Reference Class Forecasting in Turkish Public Construction Projects: Contractor Perspective. *Journal of Management in Engineering,* 32(3), 05016002.1-7.

Bonett, D. G. and Wright, T. A. (2015). Cronbach's alpha reliability: interval estimation, hypothesis testing, and sample size planning. *Journal of organisational* behaviour, 36(1), 3-15.

Bouayed, Z. (2016). Using Monte Carlo simulation to mitigate the risk of project cost overruns. International Journal of Safety and Security Engineering. 6(2), 293-300.

Bougie, R. and Sekaran, U. (2020). Research methods for business: A skill-building approach, John Wiley and Sons.

Britto, A. and Perera, B. (2013). Factors affecting the accuracy of pre-tender estimation of road construction in Sri Lanka. *FARU Journal,* 5(1), 243-254.

Brunes, F. and Lind, H. (2014). Explaining cost overruns in infrastructural projects: A new framework with applications to Sweden. *Construction Management and Economics*. 33(7), 554-568.

Cain, M. K., Zhang, Z. and Yuan, K. H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation. *Behaviour Research Methods,* 49, 1716-1735.

Calahorra-Jimenez, M., Alarcón, L. F., Torres-Machi, C., Chamorro, A. and Molenaar, K. (2020). Improving cost performance in design-bid-build road projects by mapping the reasons for cost overruns into the project phases. *Revista de la Construccion,* 19(2), 334-345.

Cantarelli, C. C., Flyvbjerg, B. and Buhl, S. L. (2012). Geographical variation in project cost performance: The Netherlands versus worldwide. *Journal of Transport Geography,* 24, 324-331.

Cantarelli, C. C., Flyvbjerg, B., Molin, E. J. E. and van Wee, B. (2010). Cost overruns in large-scale transportation infrastructure projects: Explanations and their theoretical embeddedness. *European Journal of Transport and Infrastructure Research,* 10(1)**,** 5-18.

Challal, A. and Tkiouat, M. (2012). The design of cost estimating a model of construction project: Application and simulation. *Open Journal of Accounting,*1(1), 15-26.

Chambers, J. C., Mullick, S. K. and Smith, D. D. (1971). *How to choose the right forecasting technique,* Harvard University, Cambridge, USA.

Cheng, M. Y. and Hoang, N. D. (2014). Interval estimation of construction cost at completion using least squares support vector machine. *Journal of Civil Engineering and Management,* 20(2), 223-236.

Cheng, Y. M. (2014). An exploration into cost-influencing factors on construction projects. *International Journal of Project Management,* 32(5)**,** 850-860.

Chinda, T. (2020). Factors affecting construction costs in Thailand: A structural equation modelling approach. *International Journal of Construction Supply Chain Management,* 10(3)**,** 115-140.

Cirilovic, J, Vajdic, N., Mladenovic, G. and Queiroz. C. (2014). Developing cost estimation models for road rehabilitation and reconstruction: case study of projects in Europe and Central Asia. *Journal of Construction Engineering and Management*. 140(3), 04013065.

Cirilovic, J., Vajdic, N., Mladenovic, G., and Queiroz, C. (2014). Developing cost estimation models for road rehabilitation and reconstruction: Case study of projects in Europe and Central Asia. *Journal of Construction Engineering and Management*, 140(3), 04013065.

City Rail Link Limited. (2022). Annual report 2022, available at https://www.cityraillink.co.nz/publications

Creedy, G. D., Skitmore, M. and Wong, J. K. (2010). Evaluation of risk factors leading to cost overrun in delivery of highway construction projects. *Journal of Construction Engineering and Management,* 136(5), 528-537.

Creswell, J. W. and Creswell, J. D. (2018). *Research design : qualitative, quantitative, and mixed methods approaches* (5th ed.): SAGE.

Creswell, J. W. and Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*: Sage publications.

Daniel, I. A. and Kumar, S. (2023). Cost overruns causes and factors in roadway and bridge construction projects, case study of Brundi. In P. Poluraju, S. H. Jeelani, K. D. C. Kumar, M. A. K. Reddy and K. H. Raja (Ed.), *Proceedings of ICASCM Conference on Advances in Sustainable Construction Materials.* 2759, 040009, AIP publishing.

Danisworo, B. and Latief, Y. (2019). Estimation model of Jakarta MRT phase 1 project cost overrun for the risk-based next phase project funding purpose. *IOP Conference Series: Earth Environmental Science*. 258(1), 012049.

Denyer, D. and Tranfield, D. (2009). Producing a systematic review. Buchanan, D. A. and Bryman, A. (Ed.), *The sage handbook of organisational research methods,* Sage Publications Inc., 671-689.

Dimitriou, L., Marinelli, M., Fragkakis, N. (2018). Early bill of quantities estimation of concrete road bridges: an artificial intelligence-based application. *Public Work Management Policy,* 23(2), 127-149.

Dimitriou, L., Marinelli, M. and Fragkakis, N. (2018). Early bill-of-quantities estimation of concrete road bridges: an artificial intelligence-based application. *Public Works Management and Policy*, 23(2), 127-149.

Doloi, H. (2013). Cost overruns and failure in project management: Understanding the roles of key stakeholders in construction projects. *Journal of Construction Engineering and Management,*139(3)**,** 267-279.

Donthu, N., Kumar, S., Mukharjee, D., Pandey, N. and Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285-296.

Durdyev, S. (2021). Review of construction journals on causes of project cost overruns. *Engineering Construction and Architectural Management*, 28(4)**,** 1241-1260.

Elbeltagi, E., Hosny, O., Abdel-Razek, R. and El-Fitory, A. (2014). Conceptual cost estimate of Libyan highway projects using artificial neural network. *International Journal of Engineering Research and Applications*, 4(8), 56-66.

Elbeltangi, E., Hosny, O., Abdel-Razek, R. and El-Fitory, A. (2014). Conceptual cost estimate of Libyan highway project using artificial neural network. Journal of Engineering Research and Applications, 4(8), 56–66.

El-Kholy, A. (2019). Exploring the best ANN model based on four paradigms to predict delay and cost overrun percentages of highway projects. *International Journal of Construction Management*, 21(7), 694-712.

El-Kholy, A. (2019). Exploring the best ANN model based on four paradigms to predict delay and cost overrun percentages of highway projects. *International Journal of Construction Management*, 21(7), 694-712.

El-Maaty, A. E. A., El-Kholy, A. M. and Akal, A. Y. (2017). Modelling schedule overrun and cost escalation percentages of highway projects using the fuzzy approach. *International Journal of Construction Management*, 21(7), 694-712.

El-Maaty, A. E. A., El-Kholy, A. M. and Akal, A. Y. (2017). Modelling schedule overrun and cost escalation percentages of highway projects using fuzzy approach. *Engineering, Construction and Architectural Management*, 24(5), 809-827.

El-Sawalhi, N. I. (2015). Support vector machine cost estimation model for road projects. *Journal of Civil Engineering and Architecture*, 9(9),1115-1125.

Emsley, M. W., Lowe, D. J., Duff, A. R., 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), 465-472.

Endut, I. R., Akintoye, A., Kelly, J. (2009). Cost and time overruns of projects in Malaysia. 21, 243-252, https://www.irbnet.de/daten/iconda/CIB10633.pdf.

Enshassi, A., Kumaraswamy, M. and Jomah, A. N. (2010). Significant factors causing time and cost overruns in construction projects in the Gaza strip: Contractors' perspective. *International Journal of Construction Management,* 10(1)**,** 35-60.

Eybpoosh, M., Dikmen, I. and Birgonul, M. T. (2011). Identification of risk paths in international construction projects using structural equation modelling. *Journal of Construction Engineering and Management,* 137(12)**,** 1164-1175.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A. and Pappas, G. (2007). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB Journal,* 22(2), 338-542.

Fayek, A. R. and Rodriguez Flores, J. R. (2010). Application of fuzzy logic to quality assessment of infrastructure projects at conceptual cost estimating stage. *Canadian Journal of Civil Engineering*, 37(8), 1137-1147.

Fellows, R. F. and Liu, A. M. (2015). *Research methods for construction*: John Wiley and Sons. West Sussex, UK.

Flyvbjerg, B. H., Holm, M. K. S. and Buhl, S. (2002). Underestimating costs in public works projects, error or lie. *Journal of the American Planning Association,* 68(3), 279-292.

Flyvbjerg, B. H., Holm, M. K. S. and Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport Reviews*. 23(1), 71-88.

Flyvbjerg, B. (2007). Cost overruns and demand shortfalls in urban rail and other infrastructure. *Transportation Planning and Technology,*30(1)**,** 9-30.

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.

França, A. and Haddad, A. (2018). Causes of construction projects cost overrun in Brazil. *International Journal of Sustainable Construction Engineering and Technology,* 9(1)**,** 69-83.

Freund, R. J., Wilson, W. J. and Sa, P. (2006). Regression analysis: statistical modelling of a response variable. London: Elsevier.

Fridgeirsson, T. V. (2016). Reference class forecasting in Icelandic transport infrastructure projects. *Transport Problems,* 11(2), 103-115.

Ganiyu, B. and Zubairu, I. (2010). Project cost prediction model using principal component regression for public building projects in Nigeria. *Journal of Building Performance,* 1(1), 21-28.

Garry, D. C., Martin, S. and Tony, S. (2006). *Risk factors leading to cost overrun in the delivery of highway construction projects*. (PhD), Queensland University of Technology.

Gordon, R. A. (2010). *Regression analysis for the social sciences*. New York (NY): Routledge.

Grech. V. and Calleja, N. (2018). WASP (Write a Scientific Paper) Parametric vs non-parametric tests. *Early Human Development,* 123, 48-49.

Grimsey, D. and Lewis, M. K. (2002). Evaluating the risks of public private partnerships for infrastructure projects. *International journal of project management,* 20(2), 107-118.

Günaydın, H. M. and Doğan, S. Z. (2004). A neural network approach for early cost estimation of structural systems of buildings. *International Journal of Project Management,* 22(7), 595-602.

Gunduz, M. and Maki, O. L. (2018). Assessing the risk perception of cost overrun through importance rating. *Technological and Economic Development of Economy,* 24(5)**,** 1829-1844.

Hallinger, P. and Kovačević, J. (2019). A bibliometric review of research on educational administration: Science mapping the literature, 1960 to 2018. *Review of Educational Research,* 89(3), 335-369.

Hammad, A. A. A., Ali, S. A., Sweis, G. J. and Bashir, A. (2008). Prediction model for construction cost and duration in Jordan. *Jordan Journal of Civil Engineering,* 2(3), 250-266.

Hanif, H., Khurshid, M. B., Lindhard, S. M. and Aslam, Z. (2016). Impact of variation orders on time and cost in mega hydropower projects of Pakistan. *Journal of Construction in Developing Countries,* 21(2)**,** 37-53.

Harrera, R. F., Snchez, O., Castaneda, K. and Porras, H. (2020). Cost overrun causavative factors in road infrastructure projects: a frequency and importance analysis. *Applied Sciences,* 10(16), 5506.

Hegazy, T., and Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management.*, 124(3), 210-218.

Heravi, G., and Mohammadian, M. (2021). Investigating cost overruns and delay in urban construction projects in Iran. *International Journal of Construction Management*, *21*(9), 958-968.

Herbsman, Z. (1986). Model for forecasting highway construction cost. *Transportation Research Record*, 1056, 47-54.

Hoffmann, J. P. (2022). *Linear regression models: applications in R*. New York (NY): Chapman and Hall/ CRC.

Huo, T., Ren, H., Cai, W., Shen, G. Q., Liu, B., Zhu, M. and Wu, H. (2018). Measurement and dependence analysis of cost overruns in mega-transport infrastructure projects: a case study in Hong Kong. *Journal of Construction Engineering and Management,* 144(3), 05018001.1-10.

Idrees, S. and Shafiq, M. T. (2021). Factors for time and cost overrun in public projects. *Journal of Engineering, Project, and Production Management,* 11(3)**,** 243-254.

Ika, L. A., and Pinto, J. K. (2022). The "re-meaning" of project success: Updating and recalibrating for a modern project management. *International Journal of Project Management*, 40(7), 835-848.

Ismail, M. Z. B., Ramly, Z. B. M. and Hamid, R. B. A. (2021). Systematic Review of Cost Overrun Research in the Developed and Developing Countries. *International Journal of Sustainable Construction Engineering and Technology*, 12(1), 196-211.

Israel, G. D. (1992). Determining sample size. *Program Evaluation and Organisational Development,* IFAS, University of Florida.

Jaafari, A., Pazhouhan, I., and Bettinger, P. (2021). Machine learning modelling of forest road construction costs. *Forests*, 12(9), 1169.

Jablonowski, C. J. and MacEachern, D. P. 2009. Developing probabilistic well construction estimates using regression analysis. *Journal Energy Exploration and Exploit*ation, 27(6), 439-452.

Jafarzadeh, R., Wilkinson, S., González, V., Ingham, J. M., and Amiri, G. G. (2014). Predicting seismic retrofit construction cost for buildings with framed structures using multilinear regression analysis. *Journal of Construction Engineering and Management*, 140(3), 04013062.

Ji, C. Mbachu, J. and Domingo, N. D. (2014). Factors influencing the accuracy of pre-contract stage estimation of final contract price in New Zealand. *International Journal of Supply Chain Management, 4*(2), 51-64.

Johnson C, Bosher L, Adekalan I, Jabeen H, Kataria S, Wijitbusaba A, Zerjav B, Arefian F. (2013). Private sector investment decisions in building and construction: increasing, managing and transferring risks. Background paper for United Nations National Strategy on Disaster Risk Reduction (UNISDR), Development Planning Unit, University College London, London, UK.

Johnson, R. M. and Babu, R. I. I. (2020). Time and cost overruns in the UAE construction industry: a critical analysis. *International Journal of Construction Management,* 20(5)**,** 402-411.

Kamal, A., Abas, M., Khan, D. and Azfar, R. W. (2019). Risk factors influencing the building projects in Pakistan: from perspective of contractors, clients and consultants. *International Journal of Construction Management,* 22(6), 1141-1157.

Kavuma, A., Ock, J. and Jang, H. (2019). Factors influencing Time and Cost Overruns on Freeform Construction Projects. *KSCE Journal of Civil Engineering,* 23(4)**,** 1442-1450.

Khan, R. A. and Umer, M. (2020). Impact of delays on the cost of a construction project- A cross-sectional study of the Pakistani construction industry. *Mehran University Research Journal of Engineering and Technology,* 39(4)**,** 815-825.

Kim, G. H., An, S. H. and Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Journal of Building and Environment*. 39, 1235-1242.

Kim2, S. Y., KIM, Y. M. and Luu, T. V. (2008). A cost estimation model for highway projects in Korea. In *Proceedings of the Korean Institute of Construction Engineering and Management* (pp. 922-925), Nov 7.

Kim, G. H., An, S. H. and Kang, K. I. (2004). Comparison of construction cost estimation models based on regression analysis, neural networks, and case-based reasoning. *Journal of Building and Environment*, 39(10), 1235-1242.

Kim, G. H., An, S. H., and Kang, K. I. (2004). Comparison of construction cost estimation models based on regression analysis, neural networks, and case-based reasoning. *Journal of Building and Environment*, 39(10), 1235-1242.

Kim, G. H., Shin, J. M., Kim, S. and Shin, Y. (2013). Comparison of school building construction costs estimation methods using regression analysis, neural network, and support vector machine. *Journal of Building Construction and Planning Research*, 1(1), 1-7.

Kim, G. H., Yoon, J. E., An, S. H., Cho, H. H. and Kang, K. I. (2004). Neural network model incorporating a genetic algorithm in estimating construction costs. *Journal of Building and Environment*, 39(11), 1333-1340.

Kim, H. J., Seo, Y. C. and Hyun, C. T. (2012). A hybrid conceptual cost estimating model for large building projects. *Automation in construction*, 25, 72-81.

Kim, S. and Shim, J. H. (2014). Combining case-based reasoning with genetic algorithm optimization for preliminary cost estimation in the construction industry. *Canadian Journal of Civil Engineering*, 41(1), 65-73.

Knight, A. and Ruddock, L. (2008). *Advanced research methods in the built environment*: Wiley-Blackwell.

KPMG (2017). Driving business performance - Project management survey 2017. available at: https://home.kpmg/nz/en/home/insights/2017/04/project-management-survey-2017.html.

Kumar, A. (2020). Examination of Cost Overrun in Highway Projects Using Artificial Neural Networks in Kerala. *International Journal of Innovation Science and Research Technology*, 5(3), 1382-1392.

Larsen, J. K., Shen, G. Q., Lindhard, S. M. and Brunoe, T. D. (2016). Factors affecting schedule delay, cost overrun, and quality level in public construction projects. *Journal of Management in Engineering*, 32(1), 04015032.

Laryea, S. and Hughes, W. (2006). The price of risk in construction projects. In: Boyd, D (Ed) *Proceedings of 22nd Annual ARCOM Conference* (pp. 553-561), 4-6 September, Birmingham, UK, Association of Researchers in Construction Management.

Latif, Q. B. A. I., Gopang, R. K. M. and Rahman, I. A. (2020). Substantial factors of construction management causes budget overrun in the construction industry of Oman. *International Journal of Sustainable Construction Engineering and Technology*,11(2)**,** 196-203.

Lee, J. K. (2008). Cost overrun and causes in Korean social overhead capital projects: Roads, rails, airports, and ports. *Journal of Urban Planning and Development*, 134(2), 59–62.

Lee, S., Jin, Y., Woo, S. and Shin, D. H. (2013). Approximate cost estimating a model of eco-type trade for river facility construction using case-based reasoning and genetic algorithms. *KSCE Journal of Civil Engineering*, 17(2), 292-300.

Lin, W. P. and Techapeeraparnich, W. (2019). Model for predicting the cost of rural road projects in Thailand. *IOP Conference Series: Materials Science and Engineering*, 652, 012004.

Lind, H. and Brunes F. (2015). Explaining cost overruns in infrastructure projects: a new framework with applications to Sweden. *Construction Management and Economics*, 33(7), 554-568.

Lipsey, M. W. and Wilson, D. B. (2001). *Practical meta-analysis*, Sage Publications Inc. New York, NY.

Love, P. E. D., Ahiaga-Dagbui, D. D. and Irani, Z. (2016). Cost overruns in transportation infrastructure projects: Sowing the seeds for a probabilistic theory of causation. *Transportation Research Part A: Policy and Practice*. 92, 184-194.

Lowe, D. J., Emsley, M. W. and Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of construction engineering and management*, 132(7), 750-758.

Mačková, D. and Bašková, R. (2014). Applicability of Bromilow´ s time-cost model for residential projects in Slovakia - Selected Scientific Papers. *Journal of Civil Engineering*, 9(2), 5-12.

Mahalakshmi, G., and Rajasekaran, C. (2019). Early cost estimation of highway projects in India using artificial neural network. In *Sustainable Construction and Building Materials: Select Proceedings of ICSCBM 2018* (pp. 659-672). Springer Singapore.

Mahamid, I. (2011). Early cost estimating for road construction projects using multiple regression techniques. *Australian Journal of Construction Economics and Building*, 11(4), 87-101.

Mahmud, A. T., Ogunlana, S. O., & Hong, W. T. (2021). Key driving factors of cost overrun in highway infrastructure projects in Nigeria: a context-based perspective. *Journal of Engineering, Design and Technology*, 19(6), 1530-1555.

Malkanthi, S. N., Premalal, A. G. D. and Mudalige, R. (2017). Impact of cost control techniques on cost overruns in construction projects. *Engineer*, L(94), 53-60

Maqsoom, A., Babar, Z., Shaheen, I., Abid, M., Kakar, M. R., Mandokhail, S. J. and Nawaz, A. (2021). Influence of construction risks on cost escalation of highway-related projects: Exploring the moderating role of social sustainability requirements. *Iranian Journal of Science and Technology-Transactions of Civil Engineering*,45(3)**,** 2003-2015.

Marigo, J. M., Blanco-Mesa, F., Gil-Lafuente, A. and Yager, R. R. (2017). Thirty years of the International Journal of Intelligent Systems: A bibliometric review. *International Journal of Intelligent Systems*,32(5), 526-554.

Marinelli, M., Dimitriou, L., Fragkakis, N., and Lambropoulos, S. (201). Non-parametric bill of quantities estimation of concrete road bridges superstructure: an artificial neural networks approach. In A. B. Raiden and E. Aboagye-Nimo (Ed.) *Proceedings 31st Annual ARCOM Conference (pp. 853-862)*, Lincoln, United Kingdom.

Masu, S., Gichunge, H. and K'Akumu, O. A. (2012). Component ratios of new building costs in Nairobi: a contractors' perspective. *Journal of Financial Management of Property and Construction*, 17(3), 222-234.

Mbachu, J. I. and Cross, C. (2015). Key drivers of discrepancies between initial and final costs of construction projects in New Zealand. *Project Management World Journal*, 4(9), 1-13.

Melaku, B. S., Tilahun, S., Yehualaw, M., Matos, J., Sousa, H., and Workneh, E. T. (2021). Analysis of Cost Overrun and Schedule Delays of Infrastructure Projects in Low Income Economies: Case Studies in Ethiopia. *Advances in Civil Engineering*, 2021, 4991204.

Memon, A. H. and Rahman, I. A. (2013). Analysis of cost overrun factors for small scale construction projects in Malaysia using PLS-SEM method. *Modern Applied Science,* 7(8)**,** 78-88.

Memon, A. Q., Memon, A. H. and Soomro, M. A. (2020). Contractor's perception on factors causing cost overrun in construction works of Pakistan. *International Journal of Sustainable Construction Engineering and Technology*, 11(3)**,** 84-92.

Ministry of Business, Innovations and Employment (MBIE). (2022). Building and Construction Sector Trends – Annual report 2022. Wellington.

Ministry of Transport New Zealand. (2020). New Zealand Government policy statement on land transport 2021/22 - 2030/31.

Mok, K. Y., Shen, G. Q. and Yang, J. (2015). Stakeholder management studies in mega construction projects: A review and future directions. *International Project Management*. 33(2), 446-457.

Narayanan, S., Kure, A. M. and Palaniappan, S. (2019). Study on Time and Cost Overruns in Mega Infrastructure Projects in India. *Journal of The Institution of Engineers (India): Series A*, 100(1), 139-145.

New Zealand Transport Agency (NZTA). (2020). Annual report 2017-2020. available at https://www.nzta.gov.nz

New Zealand Transport Agency (NZTA). (2020). Waka Kotahi Investment proposal 2020 – 31, available on https://www.nzta.govt.nz/assets/planning-and-investment/docs/waka-kotahi-investment-proposal-2021-31.pdf.

New Zealand Transport Agency (NZTA). (2018). One network road classification (ONRC) performance measured – A general guide. available at https://www.nzta.gov.nz

Ng, S. T., Mak, M. M. Y., Skitmore, R. M., Lam, K. C. and Varnam, M. (2001). The predictive ability of Bromilow's time–cost model. *Construction Management and Economics*, 19(2), 165-173.

Nicolaisen, M. S., Ambrasaite, I. and Salling, K. B. (2012). Forecasts: uncertain, inaccurate and biased? *Danish Journal of Transportation Research*, 1-10.

Nijkamp, P. and Ubbels, B. (1999). How reliable are estimates of infrastructure costs? A comparative analysis. *International Journal of Transport Economics*, 26(1), 23-53.

Odeck J. (2019). Variation in Cost Overruns of Transportation Projects: An Econometric Meta-regression Analysis of Studies Reported in the Literature. Transportation. 46(4), 1345-1368.

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

Panik, M. J. (2009). *Regression modelling: Methods, theory, and computation with SAS*. Florida: Chapman and Hall/ CRC.

Paraskevopoulou, C., and Boutsis, G. (2020). Cost overruns in tunnelling projects: Investigating the impact of geological and geotechnical uncertainty using case studies. *Infrastructures*, *5*(9), 5090073.

Park, Y. I. and Papadopoulou, T. C. (2012). Causes of cost overruns in transport infrastructure projects in Asia: Their significance and relationship with project size. *Built Environment Project and Asset Management*, 2(2)**,** 195-216.

Peleskei, C. A., Dorca, V., Munteanu, R. A., and Munteanu, R. (2015). Risk Consideration and Cost Estimation in Construction Projects Using Monte Carlo Simulation. *Management*, 10(2), 18544223.

Petruseva, S., Pancovska, V. Z., Zujo, V. and Vejzovic, A. B. (2017). Construction costs forecasting comparison of the accuracy of linear regression and support vector machine models. *Technicki Vjesnik*, 24(5), 1431-1438.

Plebankiewicz, E. (2018). Model of Predicting Cost Overrun in Construction Projects. *Sustainability*, 10(2)**,** 4387.

Principal Economics. (2022). Great Decisions are timely: Benefits from more efficient infrastructure investment decision-making. Report to Infrastructure New Zealand – October 2022. Available on https://infrastructure.org.nz/resources/reports.

Rahman, I. A., Memon, A. H. and Karim, A. T. A. (2013). Significant factors causing cost overruns in large construction projects in Malaysia. *Journal of Applied Sciences,* 13(2)**,** 286-293.

Randhawa, K., Wilden, R. and Hohberger, J. (2016). A bibliometric review of open innovation: Setting a research agenda. *Journal of Product Innovation Management,* 33(6), 750-772.

Rosenfeld, Y. (2014). Root-cause analysis of construction-cost overruns. *Journal of Construction Engineering and Management,* 140(1), 04013039.

Ryan, T. P. (2009). *Modern regression methods*. New Jersey (NJ): Wiley and Sons Inc.

Salmi, A., David, P., Blanco, E., and Summers, J. D. (2016). A review of cost estimation models for determining assembly automation level. *Computers and Industrial Engineering*, 98, 246-259.

Saunders, M., Lewis, P., and Thornhill, A. (2016). *Research methods for business students* (7th ed.): Pearson education.

Seeley, I. H. (1996). *Building economics: appraisal and control of building design cost and efficiency* (4th ed.). London (UK): Bloomsbury.

Shabniya, V., and Dilruba, K. M. (2017). A review on construction cost forecasting techniques. *International Research Journal of Engineering and Technology*, 4(5), 3333-3339.

Shah, R. K. (2016). An exploration of causes for delay and cost overrun in construction projects; A case study of Australia, Malaysia and Ghana. *Journal of Advanced College of Engineering and Management,* 2(1), 41-55.

Shaikh, F. A. (2020). Financial mismanagement: A leading cause of time and cost overrun in mega construction projects in Pakistan. *Engineering Technology and Applied Science Research,*10(1)**,** 5247-5250.

Shanmugam, M., Amaratunga, R. and Zainudeen, N. (2006). Simulation modelling of cost overruns in building projects", In *Proceedings of the 6th International Postgraduate Research*

Conference in the Built and Human Environment, 6-7 April, Delft University of Technology and TNO, Netherlands, available at: http://usir.salford.ac.uk/id/eprint/9884/.

Shanmuganathan, N. and Baskar, G. (2015). Ranking of delay factors cause time and cost overruns in construction projects in Tamil Nadu. *International Journal of Applied Engineering Research*, 10(24)**,** 44445-44453.

Shehu, Z., Endut, I. R., Akintoye, A. and Holt, G. D. (2014). Cost overrun in the Malaysian construction industry projects: A deeper insight. *International Journal of Project Management*, 32(8), 1471-1480.

Shr, J. F. and Chen, W. T. (2006). Functional model of cost and time for highway construction projects. *Journal of Marine Science and Technology*, 14(3), 127-138.

Silva, I. N. D., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., and Alves, S. F. D. R. (2017) *Artificial neural networks: A practical course*, Springer, Switzerland.

Siraj, N. B., and Fayek, A. R. (2019). Risk identification and common risks in construction: Literature review and content analysis. *Journal of Construction Engineering and Management*, 145(9), 03119004.

Skitmore, M. and Marston, V. (2005), *Cost Modelling*. Taylor and Francis, London, UK.

Skitmore, R. M. and Ng, S. T. (2003). Forecast models for actual construction time and cost. *Journal of Building and Environment,* 38(8), 1075-1083.

Snee, R. D. (1983). Regression diagnostics: identifying influential data and sources of collinearity. *Journal of Quality Technology*, 15(3), 149-153

Sodikov, J. (2005). Cost estimation of highway projects in developing countries: an artificial neural network approach. *Journal of the Eastern Asia Society for Transportation Studies,* 6, 1036-1047.

Sohu, S., Abd Halid, A., Nagapan, S., Fattah, A., Latif, I., and Ullah, K. (2017). Causative factors of cost overrun in highway projects of Sindh province of Pakistan. *IOP Conference Series: Materials Science and Engineering*, 271(1), 012036.

Sohu, S., Ansari, A. A. and Jhatial, A. A. (2020). Most common factors causing cost overrun with its mitigation measure for Pakistan construction industry. *International Journal of Sustainable Construction Engineering and Technology*, 11(2)**,** 256-261.

Sonmez, R. and Ontepeli, B. (2009). Predesign cost estimation of urban railway projects with parametric modelling. *Journal of Civil Engineering and Management*, 15(4), 405-409.

Steininger, B. I., Groth, M. and Weber, B. L. (2021). Cost overruns and delays in infrastructure projects: the case of Stuttgart 21. *Journal of Property Investment and Finance*,39(3)**,** 256-282.

Tabachnick, B. G., and Fidell, L. S. (2001). Using Multivariate Statistics, 4th Edition, Needham Heights, MA: Allyn and Bacon.

Tah, J. H. M., Carr, V. and Howes, R. (1998). An application of case-based reasoning to the planning of highway bridge construction. *Engineering, Construction and Architectural Management*, 5(4), 327-338.

Tan, J., and Zhao, J. Z. (2019). The rise of public–private partnerships in China: an effective financing approach for infrastructure investment? *Public Administration Review*, 79(4), 514-518.

Tarek, H. and Ayed, A. (1998). A neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management*, 124(3), 210-218.

The Association for the Advancement of Cost Engineering – International. (2020). Cost Estimate Classification System – As applied in Engineering, Procurement, and Construction for the Process Industries: TCM Framework 7.3 – Cost Estimating and Budgeting.

Themsen, T. N. (2019). The processes of public megaproject cost estimation: The inaccuracy of reference class forecasting. *Financial Accountability and Management*,35(4), 337-352.

Touran, A. and Lopez, R. (2006). Modelling cost escalation in large infrastructure projects. *Journal of Construction Engineering and Management*, 132(8), 853-860.

Ullah, K., Abdullah, A. H., Nagapan, S., Sohu, S. and Khan, M. S. (2018). Measures to mitigate causative factors of budget overrun in Malaysian building projects. *International Journal of Integrated Engineering*, 10(9)**,** 66-71.

Vahdani, B., Mousavi, S. M., Mousakhani, M., Sharifi, M. and Hashemi, H. (2012). A neural network model based on support vector machine for conceptual cost estimation in construction projects. *Journal of Optimization in Industrial Engineering,*6(10), 11-18.

Van Eck, N. J. and Waltman, L. (2014). CitNetExplorer: A new software tool for analysing and visualising citation networks. *Journal of Informatrics,* 8(4), 802-823.

Vargha, A. and Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioural Statistics,* 23(2), 170-192.

Vickerman, R. (2007). Cost—Benefit analysis and large-scale infrastructure projects: State of the art and challenges. *Environment and Planning B: Planning and Design*, 34(4), 598-610.

VU, T. Q., Pham, C. P., Nguyen, T. A., Nguyen, P. T., Phan, P. T. and Nguyen, Q. L. H. T. T. (2020). Factors influencing cost overruns in construction projects of international contractors in Vietnam. *Journal of Asian Finance Economics and Business,* 7(9), 389-400.

Wang, B., Yuan, J and Ghafoor, K. Z. (2021). Research on construction cost estimation based on artificial intelligence technology. *Scalable Computing: Practice and Experience,* 22(2), 93-104.

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), 470-478.

Wanjari, S. P. and Dobariya, G. (2016). Identifying factors causing cost overrun of the construction projects in India. *Sadhana - Academy Proceedings in Engineering Sciences,* 41(6), 679-693.

Won-Gil, H., Sangyong, K. and Jung-Kyu, J. (2019). Improved similarity measure in case-based reasoning: a case study of construction cost estimation. *Engineering, Construction and Architectural Management,* 27(2), 561-578.

Xue, X., Jia, Y. and Tang, Y. (2020). Expressway project cost estimation with a convolutional neural network model. I*EEE Access*, 8, 217848-217866.

Yin, R. K. (2009), *Case study research: design and methods* (4th ed.), Sage Publications Inc, California.

Young, D. S. (2017). *Handbook of regression methods*. New York (NY): Chapman and Hall/ CRC.

Zhang, Y., Minchin, R. E. and Agdas, D. (2017). Forecasting completed cost of highway construction projects using LASSO regularised regression. *Journal of Construction Engineering and Management*, 140(10), 04017071.

Zhao, L., Mbachu, J. and Zhang, H. (2019). Forecasting residential building costs in New Zealand using a univariate approach. *International Journal of Engineering Business Management,* 11, 1-13.

Zhao, L., Wang, B., Mbachu, J., and Liu, Z. (2019). New Zealand building project cost and its influential factors: a structural equation modelling approach. *Advances in Civil Engineering*, 2019, 1-16.

Zhao, X., Zuo, J.,Wu, G. and Huang, C. (2019). A bibliometric review of green building research 2000-2016. *Architectural Science Review,* 62(1), 74-88.

Zima, K. (2015). The case-based reasoning model of cost estimation at the preliminary stage of a construction project. *Procedia Engineering*, 122, 57-64.

Appendix A - Ethics approval

Date: 29 July 2020

Dear Chinthaka Atapattu

Re: Ethics Notification - **4000023063** - **Cost model for cost overruns in infrastructure projects in New Zealand**

Thank you for your notification which you have assessed as Low Risk.

Your project has been recorded in our system which is reported in the Annual Report of the Massey University Human Ethics Committee.

The low risk notification for this project is valid for a maximum of three years.

If situations subsequently occur which cause you to reconsider your ethical analysis, please contact a Research Ethics Administrator.

Please note that travel undertaken by students must be approved by the supervisor and the relevant Pro Vice-Chancellor and be in accordance with the Policy and Procedures for Course-Related Student Travel Overseas. In addition, the supervisor must advise the University's Insurance Officer.

A reminder to include the following statement on all public documents:
"This project has been evaluated by peer review and judged to be low risk. Consequently, it has not been reviewed by one of the University's Human Ethics Committees. The researcher(s) named in this document are responsible for the ethical conduct of this research.

If you have any concerns about the conduct of this research that you want to raise with someone other than the researcher(s), please contact Professor Craig Johnson, Director - Ethics, telephone 06 3569099 ext 85271, email humanethics@massey.ac.nz."

Please note, if a sponsoring organisation, funding authority or a journal in which you wish to publish requires evidence of committee approval (with an approval number), you will have to complete the application form again, answering "yes" to the publication question to provide more information for one of the University's Human Ethics Committees. You should also note that such an approval can only be provided prior to the commencement of the research.

Yours sincerely

Research Ethics Office, Research and Enterprise
Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand **T** 06 350 5573; 06 350 5575 **F** 06 355 7973
E humanethics@massey.ac.nz **W** http://humanethics.massey.ac.nz

Human Ethics Low Risk notification

Professor Craig Johnson
Chair. Human Ethics Chairs' Committee and Director (Research Ethics)

Research Ethics Office, Research and Enterprise
Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand **T** 06 350 5573; 06 350 5575 **F** 06 355 7973
E humanethics@massey.ac.nz **W** http://humanethics.massey.ac.nz

Appendix B - Questionnaire Survey Sample

School of Built Environment
Massey University – Auckland Campus
Albany 0632
Auckland, New Zealand

Factors affecting the cost overruns of transportation infrastructure projects in New Zealand

Project description

Cost overrun is considered as one of the most critical issues during the execution of construction projects in New Zealand. According to the figures published by the New Zealand Transport Agency (NZTA), average cost overruns in infrastructure projects is 27%. It is emphasized that the traditional estimating methods failed to provide accurate estimates although cost databases such as QV Cost Builder is useful sometimes. The government is planning to invest $12 billion on infrastructure over the next four years, including $86 million on transport.

This survey focuses on identifying the factors that affect the cost overruns of infrastructure projects in New Zealand. This survey forms part of my PhD research project that aims to develop a model to calculate the Preliminary estimate of the project minimizing the overruns custom made to the New Zealand infrastructure projects. The outcome of the research will help builders prepare more accurate cost estimates.

Project Procedures and Data Management

The survey will take around 10 minutes to complete. Data will be stored securely under strict access and password protection. Access to this information is only available to the researcher and supervisors. The data will be stored in a desktop computer and secure cloud storage. The project findings will be published in conferences and journals. The participants will be notified of the publications upon request.

Ethics committee Approval Statement

This research has been reviewed and approved by the Massey University Human Ethics Committee: Northern, Application 4000017232. If you have any concerns about the conduct of this research, please contact A/Prof David Tappin, Chair, Massey University Human Ethics Committee: Northern, telephone 64 9 414 0800 x 43384, email humanethicsnorth@massey.ac.nz.

Thank for participating with the survey.

Do you wish to continue with the survey?	Yes	☐	No	☐
Do you have experience in infrastructure projects?	Yes	☐	No	☐

SECTION 01 – GENERAL INFORMATION

Instructions: Please tick where appropriate

01.	Designation of the respondent									
	Project Manager		Quantity Surveyor		Architect		Builder		Other (please specify) …….	

02.	Years of experience in the construction industry (New Zealand + overseas)							
	Less than 5 years		From 5 to 10 years		From 10 to 20 years		More than 20 years	

03.	Years of experience in the **New Zealand** construction industry							
	Less than 5 years		From 5 to 10 years		From 10 to 20 years		More than 20 years	

04.	Years of experience in the **New Zealand Infrastructure Projects**							
	Less than 5 years		From 5 to 10 years		From 10 to 20 years		More than 20 years	

05.	Academic qualifications						
	Doctorate		Master's Degree		Bachelor's Degree		Diploma or any other technical	

06.	Are you a member of any professional Institute?				
	Yes		No		If yes, please write the post nominals ……

07.	What type of project/ s you have been working for?					
	Building projects		Infrastructure projects		Services projects	

08.	What type of client/ s you have been working with?					
	Public Sector		Private Sector		PPP (Public Private Partnerships)	

SECTION 02 – FACTORS AFFECTING THE COST OVERRUNS IN NEW ZEALAND INFRASTRUCTURE PROJECTS

Instructions: According to your experience; please identify (carefully) the degree of effect of each factor as a cause of cost overruns in **infrastructure projects in New Zealand.**

Very high effect = 5; High effect = 4; Neutral = 3; Low effect = 2; Very low effect = 1

No	Causes	Degree of the effect				
		1	2	3	4	5
	Section 2:1 – Contractor related causes					
01	Poor site management and supervision					
02	Incompetent sub-contractors					
03	Poor planning, scheduling, monitoring & controlling					
05	Lack of experience in infrastructure					
06	Inaccurate time and cost estimates					
07	Mistakes during construction/ Rework					
08	Improper construction methods					
	Section 2:2 – Design and documentation related causes					
09	Mistakes and frequent changes in design					
10	Extra works					
11	Incomplete tender drawings and specifications					
12	Delays in design and approvals					
	Section 2:3 – Financial management related causes					
13	Financial difficulties of the client or contractor					
14	Financial difficulties of the contractor					
15	Poor cashflow forecast					
16	Poor financial management					
17	Payment delays					
18	Contractual claims					
	Section 2:4 – Information and communication technology related causes					
19	Poor team communication & coordination					
20	Slow information flow among parties					
	Section 2:5 – Labour management related causes					
21	Poor labour productivity					
22	Shortage labours					
23	Shortage of technical personnel					
24	Absence of workers					
25	High cost of labours					
26	Lack of incentives					

	Section 2:6 – Material management related causes					
27	Price fluctuation					
28	Shortages of materials					
29	Late delivery of materials and equipment					
30	Equipment availability and failure					
	Section 2:7 – Project management and contract administration related causes					
31	Tendering strategy selection					
32	Procurement method selection					
33	Poor risk assessment					
34	Poor project management					
35	Change in the scope of the project					
36	Delays in decision making/ approvals					
37	Project complexity					
38	Project duration					
39	Project location					
40	Inaccurate quantity take-off					
	Section 2:8 – Organizational related causes					
41	Unsuitable management structure					
42	Poor organizational management procedures					
	Section 2:9 – Environmental related causes					
43	Unpredictable harsh weather conditions					
44	Problems with neighbours					
45	Unforeseen ground conditions					
46	Improper public traffic control methods					
	Section 2:10 – Psychological related causes					
47	Optimism bias among local officials					
48	Cognitive bias of people					
49	Cautious attitude towards risk					
50	Deception					
	Section 2:11 – Political related causes					
51	Government initiatives					
52	Deliberate cost under estimation					
53	Manipulation of forecasts					

Appendix C - Statements of Contribution for Publications

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Chinthaka Niroshan Atapattu
Name and title of main supervisor:	Dr Niluka Domingo
In which chapter is the manuscript/published work?	Chapter 03
What percentage of the manuscript/published work was contributed by the student?	80%

Describe the contribution that the student has made to the manuscript/published work:

The candidate has written the original draft of manuscript including conceptualisation, methodology, data collection and analysis, visualisation, and editing.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  Atapattu, C.N., Domingo, N., and Sutrisna, M. (2023). What significant risk factors affect the final cost of road projects in NZ – Quantity Surveyor's perception, Built Environment Project and Asset Management, Vol. 13 No. 5, pp. 756-777. DOI: https://doi.org/10.1108/BEPAM-07-2022-0105.

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Chinthaka Atapattu Digitally signed by Chinthaka Atapattu DN: cn=Chinthaka Atapattu, c=NZ, email=Chinthaka.Atapattu@op.ac.nz Date: 2023.08.21 20:55:49 +12'00'	Main supervisor's signature:	Niluka Digitally signed by Niluka DN: cn=Niluka, c=NZ, o=Massey University, ou=School of Built Environment, email=n.d.domingo@massey.ac.nz Date: 2023.08.29 09:55:43 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Chinthaka Niroshan Atapattu
Name and title of main supervisor:	Dr Niluka Domingo

In which chapter is the manuscript/published work?	Chapter 04
What percentage of the manuscript/published work was contributed by the student?	80%

Describe the contribution that the student has made to the manuscript/published work:

The candidate has written the original draft of manuscript including conceptualisation, methodology, data collection and analysis, visualisation, and editing.

Please select one of the following three options:

- ○ **The manuscript/published work is published or in press**
  Please provide the full reference of the research output:
- ◉ **The manuscript is currently under review for publication**
  Please provide the name of the journal:
  Developments in the Built Environment
- ○ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Chinthaka Atapattu Digitally signed by Chinthaka Atapattu DN: cn=Chinthaka Atapattu, c=NZ, email=Chinthaka.Atapattu@op.ac.nz Date: 2023.08.21 20:58:27 +12'00'	Main supervisor's signature:	Niluka Digitally signed by Niluka DN: cn=Niluka, c=NZ, o=Massey University, ou=School of Built Environment, email=n.d.domingo@massey.ac.nz Date: 2023.08.29 09:54:46 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Chinthaka Niroshan Atapattu
Name and title of main supervisor:	Dr Niluka Domingo

In which chapter is the manuscript/published work?	Chapter 05
What percentage of the manuscript/published work was contributed by the student?	80%

Describe the contribution that the student has made to the manuscript/published work:

The candidate has written the original draft of manuscript including conceptualisation, methodology, analysis, visualisation, and editing.

Please select one of the following three options:

- (●) **The manuscript/published work is published or in press**

  Please provide the full reference of the research output:

  Atapattu, C. N., Domingo, N. and Sutrisna, M. (2023). A bibliometric review of statistical modelling techniques for cost estimation of infrastructure projects. Smart and Sustainable Built Environment, (Ahead of print) DOI: https://doi.org/10.1108/SASBE-01-2023-0005

- ( ) **The manuscript is currently under review for publication**

  Please provide the name of the journal:

- ( ) **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Chinthaka Atapattu Digitally signed by Chinthaka Atapattu DN: cn=Chinthaka Atapattu, c=NZ, email=Chinthaka.Atapattu@op.ac.nz Date: 2023.08.21 20:56:20 +12'00'	Main supervisor's signature:	Niluka Digitally signed by Niluka DN: cn=Niluka, c=NZ, o=Massey University, ou=School of Built Environment, email=n.d.domingo@massey.ac.nz Date: 2023.08.29 09:54:57 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Chinthaka Niroshan Atapattu
Name and title of main supervisor:	Dr Niluka Domingo
In which chapter is the manuscript/published work?	Chapter 06
What percentage of the manuscript/published work was contributed by the student?	80%

Describe the contribution that the student has made to the manuscript/published work:

The candidate has written the original draft of manuscript including conceptualisation, methodology, data collection and analysis, visualisation, and editing.

Please select one of the following three options:

- ○ **The manuscript/published work is published or in press**
  Please provide the full reference of the research output:
- ◉ **The manuscript is currently under review for publication**
  Please provide the name of the journal:
  Journal of Financial Management of Property and Construction
  Manuscript ID: JFMPC-08-2023-0052
- ○ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Chinthaka Atapattu Digitally signed by Chinthaka Atapattu DN: cn=Chinthaka Atapattu, c=NZ, email=Chinthaka.Atapattu@op.ac.nz Date: 2023.08.21 20:56:55 +12'00'	Main supervisor's signature:	Niluka Digitally signed by Niluka DN: cn=Niluka, c=NZ, o=Massey University, ou=School of Built Environment, email=n.d.domingo@massey.ac.nz Date: 2023.08.29 09:55:08 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Chinthaka Niroshan Atapattu
Name and title of main supervisor:	Dr Niluka Domingo

In which chapter is the manuscript/published work?	Chapter 07
What percentage of the manuscript/published work was contributed by the student?	80%

Describe the contribution that the student has made to the manuscript/published work:

The candidate has written the original draft of manuscript including conceptualisation, methodology, data collection and analysis, visualisation, and editing.

Please select one of the following three options:

○	**The manuscript/published work is published or in press** Please provide the full reference of the research output:
◉	**The manuscript is currently under review for publication** Please provide the name of the journal: Journal of Construction Engineering and Management Manuscript ID: COENG-14398
○	**It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Chinthaka Atapattu Digitally signed by Chinthaka Atapattu DN: cn=Chinthaka Atapattu, c=NZ, email=Chinthaka.Atapattu@op.ac.nz Date: 2023.08.21 20:57:30 +12'00'	Main supervisor's signature:	Niluka Digitally signed by Niluka DN: cn=Niluka, c=NZ, o=Massey University, ou=School of Built Environment, email=n.d.domingo@massey.ac.nz Date: 2023.08.29 09:55:20 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.

MASSEY UNIVERSITY
TE KUNENGA KI PŪREHUROA
UNIVERSITY OF NEW ZEALAND

GRADUATE RESEARCH SCHOOL

STATEMENT OF CONTRIBUTION
DOCTORATE WITH PUBLICATIONS/MANUSCRIPTS

We, the student and the student's main supervisor, certify that all co-authors have consented to their work being included in the thesis and they have accepted the student's contribution as indicated below in the Statement of Originality.

Student name:	Chinthaka Niroshan Atapattu
Name and title of main supervisor:	Dr Niluka Domingo

In which chapter is the manuscript/published work?	Chapter 08
What percentage of the manuscript/published work was contributed by the student?	80%

Describe the contribution that the student has made to the manuscript/published work:

The candidate has written the original draft of manuscript including conceptualisation, methodology, data collection and analysis, visualisation, and editing.

The manuscript is significantly connected with chapter 6 and 7. Therefore, this manuscript will be submitted to the journal once the papers based on chapter 6 and 7 are published to avoid major repeating sections.

Please select one of the following three options:

- ○ **The manuscript/published work is published or in press**
  Please provide the full reference of the research output:
- ○ **The manuscript is currently under review for publication**
  Please provide the name of the journal:
  Sustainable Cities and Society
- ◉ **It is intended that the manuscript will be published, but it has not yet been submitted to a journal**

Student's signature:	Chinthaka Atapattu Digitally signed by Chinthaka Atapattu DN: cn=Chinthaka Atapattu, c=NZ, email=Chinthaka.Atapattu@op.ac.nz Date: 2023.08.21 20:57:51 +12'00'	Main supervisor's signature:	Niluka Digitally signed by Niluka DN: cn=Niluka, c=NZ, o=Massey University, ou=School of Built Environment, email=n.d.domingo@massey.ac.nz Date: 2023.08.29 09:55:31 +12'00'

This form should be placed at the beginning of each relevant thesis chapter.