Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author’s right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

To request permissions please use the Feedback form on our webpage. http://researchspace.auckland.ac.nz/feedback

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.

Note: Masters Theses

The digital copy of a masters thesis is as submitted for examination and contains no corrections. The print copy, usually available in the University Library, may contain corrections made by hand, which have been requested by the supervisor.
Seismic Retrofit Cost Modelling of Existing Structures

Reza Jafarzadeh

A thesis submitted in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Construction Engineering and Management
School of Engineering
Department of Civil and Environmental Engineering
The University of Auckland, New Zealand

Supervised by

Associate Professor Suzanne Wilkinson

Co-supervised by

Associate Professor Jason M. Ingham and Dr. Vicente Gonzalez

October 2012
ABSTRACT

Attempts to predict construction cost represent a problem of continual concern and interest to both researchers and practitioners. Such an attempt was carried out in this study to predict the construction cost associated with seismic retrofitting of existing structures. This prediction is a fairly new field of research in the civil engineering and construction industry, which has received little attention thus far despite its increasing importance and the large investment being directed towards seismic retrofit projects. In this research, retrofit net construction cost (RNCC) and its influential variables served as the dependent and independent variables, respectively. The total number of independent variables considered was fourteen, almost half of which have never been studied in the literature. These variables, together with RNCC, were tailored to develop parametric and non-parametric retrofit cost estimating (RCE) models, using the multi-linear regression (MLR) analysis and the artificial neural network (ANN) methodologies, respectively. The database used in this development was composed of 158 data points, each pertaining to a particular earthquake-prone public school with framed structures (i.e., concrete or steel structure). Of the total 158 data points in the research database, 75% (i.e., 121 samples) were randomly separated out to constitute the development or training dataset, and the remaining 25% (i.e., 37 samples) were held to constitute the hold-out or test dataset. The former dataset was mainly used to develop RCE models and evaluate their predictive accuracy ability, while the latter dataset was mainly used to evaluate the generalization ability of the developed RCE models.

The backward elimination (BE) regression technique was employed to properly explore the extent of the influence of independent variables on the RNCC, and consequently those variables that made a statistically significant contribution to the prediction of the RNCC were identified. Of the fourteen independent variables examined in this research, seven variables addressing the building characteristics in addition to the site characteristics appeared to be significant predictors of the RNCC (i.e., p<0.05). Using the BE technique, fourteen different regression models were developed. Of these regression models, the model with the inclusion of all seven statistically significant variables was found to reveal the highest predictive accuracy ability. The causal relationships between the RNCC and each of its independent variables in this regression model were described. The generalization ability of a range of regression models, each with a specific set of statistically significant variables,
was also evaluated. The results of this evaluation indicated that the most parsimonious regression model having only one independent variable (i.e., building area) showed the highest generalization ability. This indication lead to the establishment of a double-log regression model that can be simply, yet, reliably used for the general purpose of predicting the retrofit construction cost.

In this research, intelligent non-parametric RCE models were developed for the first time in the literature by means of the ANN methodology. Using the feed-forward multi-layer perceptron (MLP) architecture, the sigmoid logistic activation function, and the back-propagation learning algorithm, a novel two-stage procedure was proposed for the successful development of ANN models. Upon the completion of this procedure, for each of the last seven MLR models derived from the implementation of the BE regression technique, the best correspondent ANN model with the highest generalization ability was developed. The application of the best ANN models to the training and test datasets resulted in the finding that the last ANN model, including all seven statistically significant variables, maintained the highest accuracy and generalization abilities. In addition, the sensitivity of the ANN methodology with respect to the variation of different parameters, which were considered in the proposed development procedure, was investigated and practical solutions for better exploitation of this methodology were suggested.

Finally, the accuracy and generalization abilities of the ANN models were respectively compared to those of their MLR counterparts, using the same sets of independent variables. The results obtained from these comparisons illustrated the general superiority of the ANN methodology over the MLR analysis. This superiority was more pronounced when a greater number of independent variables were taken into account, and when a greater number of data points were included in the analysis.
ACKNOWLEDGEMENT

I am deeply grateful to my supervisor Associate Professor Suzanne Wilkinson for her trust on me and offering the opportunity to do a PhD overseas. I would like to express my sincere thanks to my first co-supervisor Associate Professor Jason M. Ingham for his second to none support and critical work reviews throughout my doctoral research. I am forever grateful to him for not only his unwavering enthusiasm and boundless optimism but also for his excellent mentoring during my writing of this thesis as well as publications that have come out of it. Thanks are also extended to my second co-supervisor Dr. Vicente Gonzalez for his technical advice, assistance, and wealth of knowledge to me. I would like to express my gratefulness to Professor Keith Molenaar from University of Colorado Boulder and Professor Martin Skitmore from Queensland University of Technology for their constructive review and comments on this work during the thesis examination.

This research would not have been possible without the financial support provided by the Farasaz Industrial Group Inc. This financial support is gratefully acknowledged as it strengthened my desire to continue and complete this research on time.

Sincere gratitude is extended to Mr. Teymour Honarbakhsh, Professor Gholam Reza Ghodrati, Professor Ali Akbar Aghakouchak, and Professor Mohammad Taghi Ahmadi, who provided diverse opportunity to expand my knowledge and offered insightful guidance tailored to my research needs. My special thanks go to all firms and organisations which generously offered their time, shared their experience, provided data, and transformed this research from a theoretical proposal into a grounded life experience. They deserve my utmost respect and appreciation, although their names have to remain in confidence.

I would like to thank all my colleagues in the construction management research group (Mohammad Reza Zare, Mostafa Babaian Jelodar, Mohammad Alkali, Yan Chang, Fei Ying, Kelvin Zuo, Temitope Egbelakin, Tingtin Liu) for their continuous support and their friendship throughout the years working together. I also want to express my gratitude to all my Iranian friends here in New Zealand who became my extended family during these years.

Last and most importantly, I would like to unreservedly thank my beloved wife, Mona Damavandi, whose constant love, encouragement, support and patience, particularly during
the long hours of writing, were *sine qua non* in allowing me to complete this thesis. As you have recently embarked on your doctorate study, I will attempt with all my heart to make this study an enjoyable journey for you as you did for me in the best way. I am deeply indebted to my mother, Sorraya Abbasi, for her sacrifice and compassion. I have no suitable words that can fully describe her perpetual passion and care towards me, my sisters, Solmaz and Sanaz, and my sweetheart niece, Lilía.
DEDICATION

This thesis is proudly dedicated to:

My adored country, Iran

(for its enriched culture and fascinating beauty)

The soul of my father, Rahim Jafarzadeh

(for his everlasting strength and spirit)

My dearest mother, Sorraya Abbasi

(for her endless sacrifice and compassion)

My beloved wife, Mona Damavandi

(for her immutable love and inspiration)

My most esteemed teacher of life, Mohammad Javad Naghshineh

(for his unflagging generosity and encouragement)
LIST OF CONTENTS

ABSTRACT .............................................................................................................................. I

ACKNOWLEDGEMENT ........................................................................................................ III

DEDICATION .......................................................................................................................... V

LIST OF CONTENTS ........................................................................................................... VII

LIST OF TABLES .................................................................................................................. XV

LIST OF FIGURES .............................................................................................................. XIX

CHAPTER 1: INTRODUCTION ............................................................................................ 1

1.1 Problem Statement ........................................................................................................ 1

1.2 Importance of Estimating Retrofit Cost ....................................................................... 2

1.2.1 Socio-economic Feasibility of Taking Retrofit into Account .................................... 3

1.2.2 Funding Retrofit Programs ..................................................................................... 4

1.2.3 Setting Optimal Retrofit Compliance Interval ....................................................... 4

1.2.4 Improving the Accuracy of Decision-making Analysis Outcomes ......................... 5

1.2.5 Framing Rational Retrofit Policy and Regulation ................................................... 5

1.3 Acceptable Thresholds for Retrofit Cost ...................................................................... 6

1.4 Research Principal Objectives ...................................................................................... 7

1.5 Motivation of the Research ......................................................................................... 8

1.6 Significance of the Research ....................................................................................... 8

1.7 Tender Price vs. Construction Cost ............................................................................. 9

1.8 Thesis Organization .................................................................................................... 10

CHAPTER 2: LITERATURE REVIEW .................................................................................. 13
2.1 Introduction .................................................................................................................... 13
2.2 Estimating Accuracy ...................................................................................................... 13
2.3 Level of Estimating Accuracy ...................................................................................... 16
2.4 Cost Modelling ............................................................................................................ 21
2.5 Cost Modelling Approaches ....................................................................................... 24
    2.5.1 Parametric Approach ........................................................................................... 24
    2.5.2 Non-parametric Approach .................................................................................. 31
2.6 Retrofit Cost Estimating (RCE) Models .................................................................... 37
    2.6.1 Potangaroa RCE Model ....................................................................................... 37
    2.6.2 FEMA Typical RCE Models ............................................................................... 38
    2.6.3 Other RCE Models ............................................................................................. 42
2.7 Limitations of the Reviewed RCE Models ................................................................. 45
2.8 Cross-Validation Technique ...................................................................................... 47
2.9 Conclusion .................................................................................................................. 51

CHAPTER 3: METHODOLOGY ....................................................................................... 53
3.1 Introduction .................................................................................................................. 53
3.2 Retrofit Cost Components and Definitions ............................................................... 53
3.3 Effective Factors Contributing to Retrofit Cost ......................................................... 60
3.4 Data Collection .......................................................................................................... 70
    3.4.1 New Zealand Survey .......................................................................................... 70
    3.4.2 Iran Survey ....................................................................................................... 75
3.5 Cost Database ............................................................................................................. 77
3.6 Data Reliability .......................................................................................................... 79
CHAPTER 4: DATA CHARACTERISTICS ................................................................. 85

4.1 Introduction ........................................................................................................ 85

4.2 Retrofit Net Construction Cost (RNCC) ............................................................ 85

4.3 Building Characteristics ..................................................................................... 89

4.3.1 Building Area .................................................................................................. 89

4.3.2 Building Total Number of Stories .................................................................. 90

4.3.3 Building Age ................................................................................................... 91

4.3.4 Building Weight .............................................................................................. 93

4.3.5 Building Configuration ................................................................................... 94

4.3.6 Structural Type ................................................................................................ 96

4.3.7 Diaphragm Type ............................................................................................. 100

4.3.8 Foundation Type ............................................................................................. 102

4.3.9 Non-structural Equipment and Component (Infill Wall Mitigation) ............ 102

4.4 Site Characteristics ........................................................................................... 103

4.4.1 Seismicity ....................................................................................................... 103

4.4.2 Soil Type ......................................................................................................... 105

4.5 Social Characteristics ....................................................................................... 107

4.6 Retrofit Policy Characteristics ........................................................................ 107

4.6.1 Compliancy with the First Building Seismic Design ..................................... 107

4.6.2 Seismic Guideline Used for Retrofitting ....................................................... 109

4.6.3 Retrofit Measure ............................................................................................ 109
CHAPTER 5: DEVELOPING REGRESSION MODELS FOR ESTIMATING RETROFIT NET CONSTRUCTION COST (RNCC) IN FRAMED STRUCTURES

5.1 Introduction

5.2 Multi-linear Regression (MLR) Analysis

5.3 Variables to Include

5.4 Functional Form

5.5 Development and Testing Datasets

5.6 Development of MLR Models

5.6.1 Criteria for Selecting the Most Appropriate MLR Model

5.6.1.2 Minimum Absolute t-value

5.6.1.3 Residual Mean Squares (RMS) of Errors

5.6.1.4 Akaike Information Criteria (AIC)

5.6.1.5 Modified Akaike Information Criteria (AICc)

5.6.1.6 Bayes Information Criteria (BIC)

5.6.2 The Selected MLR Model for Estimating RNCC

5.7 Detection of the Proposed MLR Model Violations

5.7.1 Checking Linearity and Normality Assumption

5.7.1.1 Normal Probability Plot of the Standardized Residuals

5.7.1.2 Scatter Plot of the Standardized Residuals against the Predictor Variables

5.7.1.3 Scatter Plot of the Standardized Residuals against the Predicted Values
5.7.2 Outliers .................................................................................................................. 144
  5.7.2.1 Outliers in the Dependent Variables ............................................................... 145
  5.7.2.2 Outliers in the Independent Variables ............................................................. 146
5.7.3 Multicollinearity .................................................................................................... 148
  5.7.3.1 The Variance Inflation Factor (VIF) ............................................................... 148
  5.7.3.2 Principal Component Approach ...................................................................... 149
5.7.4 Homoscedasticity ............................................................................................... 151
5.7.5 Autocorrelation ................................................................................................... 151
5.8 Causal Relationships in the Proposed MLR Model ..................................................... 152
  5.8.1 The Relation between Building Area and RNCC .................................................. 153
  5.8.2 The Relation between Building Number of Stories and RNCC ............................ 155
  5.8.3 The Relation between Building Weight Indicator and RNCC .............................. 157
  5.8.4 The Relation between Seismicity and RNCC ........................................................ 158
  5.8.5 The Relation between Soil Type and RNCC ........................................................ 159
  5.8.6 The Relation between Plan Irregularity and RNCC .............................................. 161
  5.8.7 The Relation between Structural Type and RNCC ............................................... 162
5.9 Performance on the Test Dataset ................................................................................ 163
5.10 Interaction Term ......................................................................................................... 167
5.11 Conclusion .................................................................................................................. 171

CHAPTER 6: ARTIFICIAL NEURAL NETWORK (ANN) METHODOLOGY ...... 175
  6.1 Introduction ............................................................................................................... 175
  6.2 Biological Neuron vs. Artificial Neuron ................................................................... 175
  6.3 Historical Background of ANN ................................................................................ 177
### LIST OF TABLES

Table 2-1: Number of data points in each building group (FEMA 156, 1994) .......................40

Table 2-2: Typical hard retrofit costs (FEMA 227, 1992) .......................................................43

Table 3-1: Cost studies with a database smaller than that of the current research ...............79

Table 4-1: Statistical summary of metrically scaled variables in the database (N=158) ......113

Table 4-2: Statistical summary of variables in the database that are not metrically scaled (N=158) ..................................................................................................................................114

Table 5-1: Terms and Definitions of Variables .....................................................................119

Table 5-2: Inclusion of the qualitative variables in the regression analysis ......................120

Table 5-3: Regression models developed using the BE procedure ....................................127

Table 5-4: Numbering of the regression models developed using the BE procedure ..........128

Table 5-5: Summary of the key statistics of the regression models developed using the BE procedure ..................................................................................................................128

Table 5-6: Results of the leave-one-out cross validation technique for all candidate models ..............................................................................................................................130

Table 5-7: Coefficients, predictions, MSE and R² values determined by implementing the leave-one-out cross validation technique for “model number seven” ..........................131

Table 5-8: Models minimum |t| and corresponding t0.05 (n-p) values ..................................134

Table 5-9: Results of different statistics for evaluating regression models developed using the BE procedure ........................................................................................................137

Table 5-10: Summary of model selection criteria and the models suggested by each criterion ..................................................................................................................................139

Table 5-11: Multicollinearity control by taking principle component approach into account ..........................................................................................................................151
Table 6-1: Recommendations for setting number of hidden neurons ........................................ 189
Table 6-2: ANN learning algorithms (Boussabaine, 1996) ............................................................ 195
Table 7-1: Details of variables considered for ANN modelling .................................................. 224
Table 7-2: ANN models to be developed in this research .......................................................... 225
Table 7-3: Number of input and output neurons in ANN models ............................................. 228
Table 7-4: Initial number of hidden neurons in ANN models ..................................................... 229
Table 7-5: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “one hidden layer” .................. 237
Table 7-6: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “two hidden layers” ............... 240
Table 7-7: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “three hidden layers” ............. 243
Table 7-8: The most desirable set of networks (of different architectures) for each of ANN models studied .................................................................................................................. 246
Table 7-9: The best architectures of ANN models along with their corresponding most suitable learning parameters .............................................................................................................. 252
Table 7-10: Results of iteration of the number of hidden neurons on the performance of the best primary network developed for the last ANN model (model number “seven”) (MLP network with two hidden layers; LR= 0.5; M= 0.9) ................................................................. 255
Table 7-11: The best numbers of neurons in the hidden layer(s) of ANN models .................... 257
Table 7-12: Characteristics of the final best ANN models .......................................................... 265
Table 7-13: Suitability of the initial values considered for the number of hidden neurons... 266
Table 7-14: Comparison of the networks sensitivity with respect to learning parameters and the number of hidden neurons .......................................................................................... 268
Table 7-15: Comparison of overlearning rates within the last best primary model (MLP network with two hidden layers; LR=0.5; M=0.9) .............................................................. 270
Table 7-16: Summary of the worst overlearning rates and their corresponding numbers of hidden neurons

Table 8-1: Characteristics of the MLR models and the final best ANN models

Table 8-2: R² values of the MLR and ANN models over the training dataset

Table 8-3: Actual and estimated values of RNCC and the absolute percentage errors produced by the last set of MLR and ANN models in the training dataset

Table 8-4: Comparison of the MLR and ANN models with respect to different measures of absolute percentage errors of RNCC in the training dataset

Table 8-5: Desirability of ANN modelling technique over MLR technique in the training dataset

Table 8-6: Results of the χ² test utilized to assess the null hypothesis in different models that their corresponding absolute percentage errors of RNCC in the training dataset follow a normal distribution

Table 8-7: Results of the non-parametric Mann-Whitney U test utilized to test the null hypothesis that there is no significant difference between the absolute percentage errors of RNCC produced by the ANN models and the errors produced by their respective MLR models in the training dataset

Table 8-8: R² values of the MLR and ANN models applied to the test dataset

Table 8-9: Actual and estimated values of RNCC and the absolute percentage errors produced by the last set of MLR and ANN models in the test dataset

Table 8-10: Comparison of the MLR and ANN models with respect to different measures of absolute percentage errors of RNCC in the test dataset

Table 8-11: Desirability of ANN modelling technique over MLR technique in the test dataset

Table 8-12: Results of the χ² test utilized to assess the null hypothesis in different models that their corresponding absolute percentage errors of RNCC in the test dataset follow a normal distribution
Table 8-13: Results of the parametric t-test utilized to test the null hypothesis that there is no significant difference between the absolute percentage errors of RNCC produced by the ANN models and those produced by their respective MLR models in the test dataset (All sets of models excluding the third and the fifth models) .................................................................. 326

Table 8-14: Results of the non-parametric Mann-Whitney U test utilized to test the null hypothesis that there is no significant difference between the absolute percentage errors of RNCC produced by the ANN models and those produced by their respective MLR models in the test dataset (The third and the fifth sets of models) ................................................................................. 326
LIST OF FIGURES

Figure 2-1: Factors affecting cost estimating accuracy (Ogunlana and Thorpe, 1991) .......... 16
Figure 2-2: Risk factor in forecast of costs (Mak and Raftery, 1992) ................................. 18
Figure 2-3: Chronology of cost estimate accuracy over the project development progress. Source: Barnes (Ref. in Skitmore and Cheung, 2007) ......................................................... 20
Figure 2-4: The Freiman curve. Source: Daschbach and Agpar (Ref. in Ogunlana, 1989) .... 21
Figure 3-1: Retrofit procedure .......................................................................................... 54
Figure 3-2: Retrofit cost framework .................................................................................. 59
Figure 3-3: Potential factors which may have influence on retrofit project cost ............... 69
Figure 3-4: Online questionnaire structure ........................................................................ 74
Figure 4-1: RNCC histogram and lognormal fitting (N=158) ............................................. 88
Figure 4-2: RNCC per unit area histogram and lognormal fitting (N=158) ......................... 88
Figure 4-3: Building distribution as a function of building area (N=158) ......................... 89
Figure 4-4: Building area histogram and gamma fitting (N=158) ..................................... 90
Figure 4-5: Building distribution as a function of number of stories (N=158) ................. 91
Figure 4-6: Building distribution as a function of building age (N=158) .......................... 92
Figure 4-7: Building age histogram and gamma fitting (N=158) ...................................... 93
Figure 4-8: Building distribution as a function of building weight indicator (N=158) ....... 94
Figure 4-9: Building distribution as a function of plan irregularity (N=158) ..................... 95
Figure 4-10: Building distribution as a function of vertical irregularity (N=158) ............... 96
Figure 4-11: An example of the buildings with concrete moment resisting frames ......... 98
Figure 4-12: An example of the buildings with one-way steel braced frames ............... 98
Figure 4-13: An example of the buildings with two-way steel braced frames

Figure 4-14: An example of the buildings with bare steel frames

Figure 4-15: Building distribution as a function of structural type (N=158)

Figure 4-16: Building distribution as a function of diaphragm type (N=158)

Figure 4-17: Building distribution as a function of foundation type (N=158)

Figure 4-18: Building distribution as a function of infill walls mitigation (N=158)

Figure 4-19: Seismicity map of Iran (Standard 2800, 3rd revision (BHRC, 2005))

Figure 4-20: Building distribution as a function of seismicity (N=158)

Figure 4-21: Response coefficient of buildings with respect to different soil types (Standard 2800, 3rd revision (BHRC, 2005))

Figure 4-22: Building distribution as a function of soil type (N=158)

Figure 4-23: Building distribution as a function of compliancy with the provisions of the first in-place seismic design code (Standard 2800, 1st edition (BHRC, 1988)) (N=158)

Figure 4-24: An example of the buildings to be retrofitted by adding new concrete shear walls

Figure 4-25: An example of the buildings to be retrofitted by adding steel braced frames

Figure 4-26: Building distribution as a function of retrofit measure (N=158)

Figure 5-1: Variation of RMS statistic within models under study

Figure 5-2: Variation of AIC, BIC, AICc statistics within models under study

Figure 5-3: Normal probability plot of the standardized residuals

Figure 5-4: Best fitted normal distribution for the standardized residuals of the proposed MLR model

Figure 5-5: Standardized residuals vs. predictor X1

Figure 5-6: Standardized residuals vs. predictor X2
Figure 5-7: Standardized residuals vs. predictor X₄ ......................................................... 143
Figure 5-8: Standardized residuals vs. predictor X₆ .......................................................... 143
Figure 5-9: Standardized residuals vs. predictor X₇ .......................................................... 143
Figure 5-10: Standardized residuals vs. predictor X₈ ......................................................... 143
Figure 5-11: Standardized residuals vs. predictor X₁₁ ......................................................... 143
Figure 5-12: Standardized residuals vs. predicted values of dependent variable Y .......... 144
Figure 5-13: Index plot of Standardized Residuals .............................................................. 145
Figure 5-14: Index plot of Leverage Values ................................................................. 147
Figure 5-15: Index plot of Cook’s Distances ................................................................. 148
Figure 5-16: Distribution of Ln (RNCC) in the development dataset as a function of
Ln (Area) ............................................................................................................................ 154
Figure 5-17: The effect of area on the mean value of RCI (RCI) ........................................ 155
Figure 5-18: Distribution of Ln (RNCC) in the development dataset as a function of
Ln (Number of Stories) ..................................................................................................... 156
Figure 5-19: The effect of number of stories on the mean value of RNCC ....................... 157
Figure 5-20: Distribution of Ln (RNCC) in the development dataset as a function of
Ln (Building Weight Indicator) ..................................................................................... 158
Figure 5-21: Distribution of Ln (RNCC) in the development dataset as a function of
seismicity .......................................................................................................................... 159
Figure 5-22: Distribution of Ln (RNCC) in the development dataset as a function of soil
type .................................................................................................................................. 160
Figure 5-23: Distribution of Ln (RNCC) in the development dataset as a function of plan
irregularity ....................................................................................................................... 161
Figure 5-24: Distribution of Ln (RNCC) in the development dataset as a function of structural
type .................................................................................................................................. 162
Figure 7-8: The best learning parameters along with their respective R² values for “training” and “test” datasets (networks with “three hidden layers”)

Figure 7-9: Comparison of the “accuracy” ability of the networks with the best learning parameters within each model

Figure 7-10: Comparison of the “generalization” ability of the networks with the best learning parameters within each model

Figure 7-11: Characteristics of the best primary networks along with their respective R² values for “training” and “test” datasets

Figure 7-12: Comparison of R² values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “one”)

Figure 7-13: Comparison of R² values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “two”)

Figure 7-14: Comparison of R² values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “three”)

Figure 7-15: Comparison of R² values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “four”)

Figure 7-16: Comparison of R² values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “five”)

Figure 7-17: Comparison of R² values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “six”)

Figure 7-18: Comparison of R² values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “seven”)

249

250

251

253

258

259

260

261

262

263

264
Figure 7-19: Comparison of the worst overlearning rates with respect to the number of input variables .................................................................................................................................................272

Figure 7-20: Characteristics of the final best ANN models along with their respective $R^2$ values for “training” and “test” datasets .........................................................................................................275

Figure 8-1: Comparison of $R^2$ values of the MLR and ANN models over the training dataset ...............................................................................................................................................299

Figure 8-2: Comparison of MAPE values of the MLR and ANN models over the training dataset ................................................................................................................................................306

Figure 8-3: Best fitted normal distribution in the training dataset (MLR model number one) ..........................................................................................................................................................312

Figure 8-4: Best fitted normal distribution in the training dataset (MLR model number two) ..........................................................................................................................................................312

Figure 8-5: Best fitted normal distribution in the training dataset (MLR model number three) ..........................................................................................................................................................312

Figure 8-6: Best fitted normal distribution in the training dataset (MLR model number four) ..........................................................................................................................................................312

Figure 8-7: Best fitted normal distribution in the training dataset (MLR model number five) ..........................................................................................................................................................313

Figure 8-8: Best fitted normal distribution in the training dataset (MLR model number six) ..........................................................................................................................................................313

Figure 8-9: Best fitted normal distribution in the training dataset (MLR model number seven) ..........................................................................................................................................................313

Figure 8-10: Best fitted normal distribution in the training dataset (ANN model number one) ..........................................................................................................................................................314

Figure 8-11: Best fitted normal distribution in the training dataset (ANN model number two) ..........................................................................................................................................................314
Figure 8-12: Best fitted normal distribution in the training dataset (ANN model number three) .................................314

Figure 8-13: Best fitted normal distribution in the training dataset (ANN model number four) .................................314

Figure 8-14: Best fitted normal distribution in the training dataset (ANN model number five) .................................315

Figure 8-15: Best fitted normal distribution in the training dataset (ANN model number six) .................................315

Figure 8-16: Best fitted normal distribution in the training dataset (ANN model number seven) ........................................315

Figure 8-17: Comparison of $R^2$ values of the MLR and ANN models over the test dataset. 318

Figure 8-18: Comparison of MAPE values of the MLR and ANN models over the test dataset .................................322

Figure 8-19: Best fitted normal distribution in the test dataset (MLR model number one) .................................328

Figure 8-20: Best fitted normal distribution in the test dataset (MLR model number two) .................................328

Figure 8-21: Best fitted normal distribution in the test dataset (MLR model number three) .........................................328

Figure 8-22: Best fitted normal distribution in the test dataset (MLR model number four) ........................................328

Figure 8-23: Best fitted normal distribution in the test dataset (MLR model number five) ........................................329

Figure 8-24: Best fitted normal distribution in the test dataset (MLR model number six) .........................................329

Figure 8-25: Best fitted normal distribution in the test dataset (MLR model number seven) .........................................329

Figure 8-26: Best fitted normal distribution in the test dataset (ANN model number one) .........................................330

Figure 8-27: Best fitted normal distribution in the test dataset (ANN model number two) .........................................330

Figure 8-28: Best fitted normal distribution in the test dataset (ANN model number three) ........................................330

Figure 8-29: Best fitted normal distribution in the test dataset (ANN model number four) .........................................330
Figure 8-30: Best fitted normal distribution in the test dataset (ANN model number five) ..331

Figure 8-31: Best fitted normal distribution in the test dataset (ANN model number six) ...331

Figure 8-32: Best fitted normal distribution in the test dataset (ANN model number seven)
................................................................................................................................................ 331

Figure A-1: Home page of designed web site (www.retrofitcost.net) ................................... 380

Figure A-2: Fundamental questionnaire (essential information for cost modelling) ............. 381

Figure A-3: Supplemental questionnaire (complementary information for cost modelling) 383
CHAPTER 1

INTRODUCTION

1.1 Problem Statement

It has been said that “disasters do not happen to well-prepared communities” (United States Congressional Budget Office (CBO), 2007). Nevertheless, requiring improvements in a community’s existing building stock is among the most difficult types of public policy decision, requiring careful engineering and economic analysis along with societal considerations (United States Government Accountability Office (GAO), 2007). Investors, city planners, and owners of buildings in geographic regions subject to seismic hazards are frequently faced with the decision of whether or not to retrofit existing structures in order to lower their potential losses due to earthquakes.

A report published by the California Seismic Safety Commission (1999) demonstrates that effective retrofit programs can greatly reduce the potential for casualties, decrease the amount of direct and indirect damage, and minimize the disruption caused by, and funds necessary for response and recovery operations. An effective retrofit program should have certain specifications such as being adequate to meet the retrofit objectives, and the cost of implementation of retrofit measures should not be prohibitive. Major retrofit programs have been conducted on the basis that “retrofit can be cost-effective in protecting people and property from natural disaster” (CBO, 2007). Fearing that the retrofit cost may be much higher than could reasonably be financed, many communities may elect to manage the problem after the event, by repairing damage when it occurs. However, leaving communities vulnerable to extensive direct and indirect losses may not be a proper management technique, particularly for those communities located in moderate to high active seismic zones. The extent of retrofit for an existing building can be mainly examined by comparing the direct and indirect cost incurred by an earthquake with the cost arising from the implementation of the retrofit measures to the building. Retrofit cost is, therefore, one of the major factors having a crucial role in the retrofit suitability of any building (FEMA 227 and 228, 1992).

Retrofit cost modelling is highly sophisticated in nature, and is a multidisciplinary task that requires taking advantage of different engineering areas of expertise ranging from geotechnical and structural engineering to engineering economics. The retrofit cost is
therefore subject to variations arising from a variety of sources associated with the various engineering areas stated above. In addition to the identification of these sources, other major problems that need to be carefully addressed in retrofit cost modelling are how to get the best information from the sparse past data and how to make the most use of analytical tools for establishing reliable models to estimate the retrofit cost. Unless solutions to these problems are found, a sound retrofit cost model cannot be developed.

Addressing the above concerns, this dissertation seeks to examine the relationship between the retrofit cost and its potential predictors and develop accurate and robust models for predicting the retrofit cost of public school buildings with framed structure (i.e., concrete and steel frame structures). This study sets out to develop such models with the aid of two mathematical approaches: a traditional parametric modelling approach using multi-linear regression (MLR) analysis, and a novel non-parametric modelling approach using artificial neural networks (ANNs). The relative predictive ability of these modelling approaches is also compared in this dissertation.

1.2 Importance of Estimating Retrofit Cost

Cost estimation is a major task in many engineering projects (Smith and Mason, 1997). Cost and time are the most important factors in determining the project performance (Love et al., 2005; Stoy and Schalcher, 2007; Stoy et al., 2007). These factors have been typically used as key criteria for examining project performance in the construction industry (Love et al., 2005). Adequate estimation of cost is a key factor in construction projects (Kim et al., 2004). A construction project begins with an initial cost estimate (Hwang, 2009). The accuracy of this estimate is essential to the success of the construction project (Hwang, 2009; Kim et al., 2004; Kim et al., 2005; Stoy and Schalcher, 2007). Cost estimation is also an important task for the management of construction projects (Adeli and Wu, 1998). Many decisions for managing such projects depend on predicted cost information (Hwang, 2009). The predicted cost, together with duration, forms a basis for budgeting, planning, monitoring and even litigation purposes (Skitmore and Ng, 2003). Accurate cost estimation is therefore crucial to boost the quality of construction management.

Like other construction activities, an accurate estimation of the retrofit cost is highly desirable in the early stage of the development of retrofit projects. In the early phase when a
new program is considered, cost estimate analyses are used to support or refute the program (Shtub and Versano, 1999). Several researchers have previously commented on the importance of early cost estimates. Smith (1995) observed that the first estimate that is published for review has a particularly crucial role to play because it is the basis for the release of funds for further studies or estimates, and because it becomes the marker against which subsequent estimates are compared. Moreover, early estimates are important because of the need to know, for the purposes of economic appraisal, the capital cost of the project. These ideas are further supported by Trost and Oberlender (2003), who indicated that early cost estimates are critical to the initial decision making process for the construction of capital projects. Early estimates, even when grossly inaccurate, often become the basis upon which all future estimates are judged (Soutos and Lowe, 2005).

The ability to accurately predict the retrofit cost has many advantageous implications for building owners, decision and policy makers, and other authorities dealing with retrofit projects. For instance, building owners have a particular interest in cost issues from the initial concept to the completion of their retrofit projects due to budgeting and programming concerns. Without the correct analysis, building owners may be subject to financial burden. Early identification of the retrofit cost also provides the opportunity for authorities to take sound actions for an effective management of retrofit projects. The following are the areas where the importance of the retrofit cost estimation comes into effect.

1.2.1 Socio-economic Feasibility of Taking Retrofit into Account

If retrofit costs were always inexpensive, the social and political costs and controversies would largely disappear (FEMA 356, 2000). Retrofit programs need to be planned such that the financial requirements of the building owners are met. From the building owners’ point of view the decision to retrofit or demolish is largely based on the retrofit economic feasibility (Smyth et al., 2004). According to retrofit experts, for most home owners cost is the primary factor in deciding whether or not to proceed with the retrofit work (Smyth et al., 2004). An accurate forecast of the retrofit cost helps the building owners with this critical decision. By comparing the construction cost of different retrofit measures incurred by full and partial retrofitting measures, Smyth et al., (2004) found that if the owners have a choice between a set of retrofit options they may choose the lowest cost option even though this option may not be as effective as those that are more expensive. This finding was
also supported by Hopkins (2005), who stated that the owners’ decision is more dominated by lower cost retrofit solutions than the effectiveness of the solutions to decrease the risk level. These studies clearly highlight the fact that the retrofit cost can have an appreciable influence on owners’ decision on retrofitting and indicate that owners would be likely to opt for selecting cheaper retrofit solutions. Consequently, for planning retrofit programs, retrofit managers should know about the cost consequences of retrofit implementation in order to design innovative and suitable options to facilitate a final decision in favour of participation in retrofitting.

1.2.2 Funding Retrofit Programs

The ability to develop an ambitious yet realistic budget is heavily dependent on the ability to obtain basic information about the project, including the project cost (Khosrowshahi and Kaka, 1996). Cost estimation is an indispensable tool in this regard, providing a reasonable basis to better allocate available limited resources for projects. This estimation is also essential for retrofit projects as it enables authorities to set a more meaningful plan and budget for such projects. For instance, after the 1994 Northridge earthquake, advisability of the retrofit cost was of much concern for The City Council of Los Angeles as to whether to begin retrofit programs for certain building types (Holmes, 1996). In addition, irrational cost estimation can lead to misleading, sometimes hazardous decisions with regard to earthquake-prone structures, especially in high seismicity regions. For example, in February 1996, CALTRANS announced that the original retrofit cost estimate for the San Francisco Bay Bridge was being revised upward from $300-400 million to over $1,000 million which may eventually prove to be unaffordable (Holmes, 1996). Therefore, reliable and valid cost estimates are necessary to prevent the negative impact of cost overruns on retrofit projects.

1.2.3 Setting Optimal Retrofit Compliance Interval

Setting an optimal retrofit compliance interval is another issue which is highly affected by the retrofit cost. If the compliance interval is too short, many owners experience financial hardship, and their political opposition can block the adoption of a mandatory ordinance. Conversely, if the interval is too long, a deadly earthquake may strike before buildings are retrofitted. The effect of the retrofit cost on the adopted compliance interval was
demonstrated in a study conducted by Dean (1993), who compared compliance intervals for cities with different seismicity and found that the retrofit financial hardship is an increasing function of the retrofit cost and a decreasing function of the compliance intervals in that the shorter the compliance time, the higher the demand for retrofit in a given year, so the higher the price charged by overworked contractors.

1.2.4 Improving the Accuracy of Decision-making Analysis Outcomes

It is technically possible to retrofit any building, regardless of its condition. However, the design of a suitable retrofit program depends on the benefits and costs of different loss reduction measures to the relevant interested parties. Retrofit cost is an inseparable and integral element of every risk assessment and decision-making analysis such as benefit-cost analysis and life cycle cost analysis. The former analysis enables determination of the most cost-effective retrofit solution, while the latter analysis is a useful tool for determining the optimal level of retrofit (Kappos and Dimitrakopoulos, 2008). Therefore, realistic and accurate retrofit cost estimation is highly required for performing a valid decision-making analysis.

1.2.5 Framing Rational Retrofit Policy and Regulation

Retrofit programs have high economic impact to communities. The cost to communities to implement and maintain stringent retrofit policies can considerably limit the amount of retrofit activities. Communities, particularly those that depend on new development for economic growth, may be hesitant to impose strict retrofit regulations as these regulations can potentially limit the amount of development that is allowed to occur in hazard-prone areas. Retrofit cost estimation is vital to improve decision-makers’ and planners’ understanding of the likely condition of their inventory, to make most use of retrofit programs in reducing the seismic hazard. For instance, retrofit feasibility can be framed in terms of the maximum allowable cost of a seismic retrofit such that if a retrofit project will be more expensive than its maximum allowable cost, the retrofit is not economically feasible. Similarly, as the expense of the retrofit compared to its maximum allowable cost decreases, the economic feasibility of the retrofit increases. The next section provides more discussion on this issue.
1.3 Acceptable Thresholds for Retrofit Cost

Retrofit activities could be ideally implemented as extensively as possible, but every retrofit activity has a cost that must be considered in the world of limited resources. Consequently, while plans for improving disaster preventive performance are desirable, too high an implementation cost is unacceptable. According to a recent document published by the United States Government Accountability Office (GAO, 2007), about one-third of over 25,000 unreinforced masonry (URM) buildings, having been inventoried in the high seismic regions of California, have not been retrofitted, primarily because of the high cost of retrofitting. The same was also reported in New York, where retrofitting thousands of earthquake-prone buildings would be impractical and economically unrealistic (GAO, 2007).

Retrofit cost should not be so high, especially when compared to present cost of a new building. In order to facilitate this comparison, the retrofit cost is expressed as a ratio of building replacement value (BRV). Certain thresholds for this ratio may be set to determine more objectively whether a building should be retrofitted or replaced. These thresholds imply that if the cost of retrofitting exceeds a certain percentage of the building replacement value, the final recommendation may favour replacement. This recommendation is warranted because demolishing an old building and replacing it with a new building with a longer period of usage is more desirable than retrofitting the building.

An early example of the thresholds set for the retrofit cost is found in Smith’s (1979) work. Smith set a cost criterion for retrofitting ordinary buildings in New Zealand, in that, the cost of strengthening should not be more than about one-third of the cost of a replacement building. Later, Potangaroa (1985) performed an economic comparison of retrofitting and rebuilding in the U.S. on the basis of the dominant interest rates at the time of study and useful building lives. Potangaroa concluded that the economic breakeven point between retrofitting and rebuilding is in the range of 60% to 75% of the building replacement cost.

A series of thresholds has been also proposed in the literature for buildings with different functionalities. According to a document disseminated by the California Seismic Safety Commission (1999), the maximum allowable retrofit cost for office buildings and hospitals in the California State Building Seismic Program was determined at 60% and 80% of the estimated replacement cost, respectively. This document also reported on the California State Allocation Board provision on retrofitting school buildings in the Berkeley Unified School District. Based on this provision, the total expenditure on modernizing and
structurally retrofitting a school was constrained to 75% of the school’s replacement value. The modernization would involve, by law, making all the schools compliant with current fire codes, child welfare regulations, and disabled accessibility standards. More recently in 2004, the Ministry of Education in Canada (MEd) undertook an assessment of existing schools located in high-risk seismic zones of the British Columbia province to determine the potential risk of structural damage or failure that could result from a significant seismic event. It was determined that renovations and seismic retrofit can proceed at a cost that does not exceed 70% of replacement cost. Renovations would be limited to the repair of existing systems or removal of hazardous materials that are disrupted in the retrofit process (British Columbia’s Ministry of Education, 2005).

1.4 Research Principal Objectives

Attempts to predict cost represent a problem of continual concern and interest to both researchers and project managers. This research presents such an attempt in a field which is unique in the engineering and construction industry and has received little attention in the literature. This research aims to investigate the cost associated with retrofitting existing buildings. The principal objectives of this research are as follows:

- To identify major factors that have a potential influence on the retrofit cost
- To explore the extent of the contribution of these factors to the retrofit cost and consequently to identify those factors that have significant impact on the retrofit cost
- To develop reliable, yet, easy to use retrofit cost estimating (RCE) models by employing a parametric cost modelling approach (i.e., multi-linear regression (MLR) technique)
- To develop more advanced RCE models by employing an innovative non-parametric cost modelling approach (i.e., artificial neural network (ANN) technique)
- To test and compare the prediction accuracy of the parametric RCE models with that of their associated non-parametric RCE models, using the same development or training dataset
- To test and compare the generalization ability of the parametric RCE models with that of their associated non-parametric RCE models, using the same unforeseen hold-out or test dataset
1.5 Motivation of the Research

In many earthquake-prone countries, there has been an increasing need to improve the seismic performance of existing structures as owners face huge problems with aging buildings. This improvement requires an in-depth consideration of the economic implications of retrofit programs. The challenge to seismically vulnerable countries is therefore lodging an economically justifiable policy to optimize the outcome of their retrofit programs. The successful development of this policy is highly dependent on the availability of an accurate estimation of the retrofit cost. The initial motivation for conducting the current research was the paucity of predictive models that can provide reliable estimation of the retrofit cost. No models that would support a retrofit cost forecast have been available to date for many earthquake-prone countries including Iran and New Zealand. There remains, therefore, an urgent need for accurate and timely retrofit cost estimating (RCE) models that can better assist building owners, decision and policy makers, and other authorities in establishing a sound retrofit program. In addition, little is known about which variables deserve to be paid more attention when developing RCE models. Currently, these variables and the extent of their influence on the retrofit cost are not known in many earthquake-prone countries, especially with respect to particular specifications, regulations and other requirements that these countries impose in their current retrofit practice.

1.6 Significance of the Research

As stated earlier, one of the major challenges faced in retrofit projects addressing earthquake-prone buildings is the ability to successfully and accurately predict their incurred cost at early stages of the retrofit planning phase. The general importance of this issue was demonstrated earlier in this chapter. The significance of this research can be detailed in the following categories:

- To assist decision-makers to develop a more effective retrofit strategy and policy at the early stages of development of retrofit projects
- To assure allocation of adequate budget for successful completion of a retrofit project
- To facilitate efficient management of retrofit projects by providing a precise basis for the budgetary estimates and also by assisting in the proper utilization of available limited resources for retrofit projects
- To help clients to become aware of their financial commitments when retrofitting their buildings so that they can make informed decisions on whether or not to proceed with retrofit
- To provide more clarity regarding the economic implications of a retrofit program so that project managers can better plan retrofit projects and so that clients can better arrange their funding requirements
- To provide reliable input to any retrofit decision-making analysis such as benefit-cost analysis or life cycle cost analysis
- To provide a reliable benchmark for further research in this area and to facilitate international comparison of the retrofit cost

1.7 Tender Price vs. Construction Cost

The price offered by a contractor to complete a construction project is more than the costs associated with construction works involved in the project. According to Newton (1991), the bidding process is usually held to comprise two stages: (1) the basic cost estimate – which is intended to establish the construction cost to the contractor, and (2) the mark-up – which is an amount added to the basic cost to cover items such as overheads, profit, risk and variations in market conditions. The basic cost and mark-up together comprise the bid price.

The construction cost estimate represents the amount to be paid for materials, labour, equipment, and other resources required for performing the work involved in the contract (Alex et al., 2010). McCaffer (1976) suggested that the construction cost estimate is a scientifically prepared estimate whose accuracy is a major factor influencing the outcome of a bidding competition. In comparison to this estimate, the mark-up is a less scientifically prepared figure which reflects the contractor’s profit expectations, his knowledge of the risks involved in taking the contract, and his judgement of the market (McCaffer, 1976). McCaffer’s suggestion was supported by Ogunlana (1989), who identified that the mark-up is applied as a purely commercial decision requiring consideration of many factors such as the state of the market, the client type, the desirability of winning the contract, and risk assessment. Ogunlana (1989) then concluded that, as the mark-up is applied due to considerations not necessarily related to actual construction on site, the contractor’s tender may often give a distorted view of the construction cost. Therefore, using the tender price as a proxy for the construction cost may lead to biased predictions for the latter, and vice versa.
1.8 Thesis Organization

With the above brief introduction, the rest of the thesis is organized as follows:

Chapter 2 presents background on the concept of cost modelling and identifies the analytical approaches used in the literature to develop cost models. Current applications of these approaches in various fields of study including engineering, construction, and management are reviewed. This chapter also presents an overview of available studies performed on the subject of retrofit cost modelling. The limitations of these studies are demonstrated.

Chapter 3 presents the methodology developed to achieve the objectives of this research. The chapter describes the hierarchical taxonomy of the retrofit cost components. This description is followed by identification of those variables that may influence the total retrofit cost. This chapter presents the procedure and tools employed to collect the required data on the retrofit cost and its influential variables. The data collection effort in both New Zealand and Iran are described. The effort in the latter country resulted in the creation of a large cost database with an exceptional reliability, while the effort in the former country was hampered. The reasons behind this failure are also described.

Chapter 4 determines the overall characteristics of the variables in the research cost database. These variables are divided into two general categories: quantitative variables and qualitative variables. The descriptive statistical summary of the variables in each category is presented.

Chapter 5 describes the development of a series of parametric retrofit cost estimating (RCE) models, using multi-linear regression (MLR) analysis. A backward elimination (BE) technique is employed to define the most critical variables that have a significant impact on determination of the retrofit construction cost. The causal relationships between these variables and the retrofit construction cost are also described. This chapter also evaluates the generalization ability of a range of MLR models, each with a specific set of statistically significant variables. According to this evaluation, this chapter offers for the first time in the literature a simple, yet, reliable double-log regression model that can be generally used for estimating the construction cost of different retrofit projects.

Chapter 6 describes artificial neural network (ANN) modelling technique and provides a historical background of its development. This chapter provides a comprehensive
discussion on the ANN principal elements, together with the paradigms upon which ANN models are developed. These paradigms are mainly defined by three components including network architecture, transfer or activation function, and learning algorithm. In this chapter, a special emphasis is given to the multilayer feed-forward network, sigmoid logistic function, and the back-propagation learning algorithm as the network architecture, activation function and the learning algorithm, respectively. These are the paradigms utilized in this study.

Chapter 7 examines, for the first time in the literature, application of the ANN modelling technique to develop intelligent non-parametric RCE models. This chapter provides a novel procedure that allows for the successful development of ANN models with higher generalization potential. Using this procedure, for each of the last seven MLR models derived from the implementation of the BE technique, the best correspondent ANN-based RCE model is developed. In addition, the effects of different ANN architectures (including the number of hidden layers, the number of hidden neurons, and the number of input neurons), and different learning parameters (including the learning rate, and momentum) on the predictive performance of the ANN models are examined.

Chapter 8 deals with an important question that has remained unanswered in the literature. This question is whether or not the ANN technique can perform better than the conventional MLR technique in estimating the retrofit cost. This chapter provides an overview of existing studies that compare the prediction strength of the ANN technique with that of conventional statistical techniques. Following this overview, the predictive performance of the MLR models developed in this study is compared to that of their respective ANN models. Accuracy and generalization abilities are considered as the principles for this comparison. The results from this comparison are then discussed.

Finally, Chapter 9 presents the conclusions derived from this study and provides suggestions for future research.
CHAPTER 2
LITERATURE REVIEW

2.1 Introduction

This chapter presents the results of an extensive literature search on different aspects associated with this study. The chapter starts with issues related to cost estimating accuracy and provides a description of the concept of cost modelling. This description is followed by a definition of the analytical approaches used in the literature to develop cost models. The application of these approaches is demonstrated through an extensive overview of the previous cost modelling studies performed in various disciplines of science. Two types of studies are mainly reviewed. Those studies using the parametric modelling approach are reviewed first, and then applications using a novel non-parametric modelling approach are demonstrated. The chapter concludes with a review of available sparse studies on the issue of retrofit cost modelling. Highlights are given regarding each study’s methodology and findings. The limitations of these studies are also described.

2.2 Estimating Accuracy

Estimation is of fundamental importance in all forms of sciences. The purpose of estimation is to make predictions and to predict (Cheung et al., 2008). Prediction accuracy is of obvious importance to users of forecasts, because forecasts are used to guide decisions (Diebold and Mariano, 1995). Prediction accuracy is also of obvious importance to producers of forecasts, whose reputations (and fortunes) rise and fall with forecast accuracy (Diebold and Mariano, 1995). As noted by Ogunlana (1989), a construction work suffers the effects of uncertainty more than most human undertakings and construction cost estimating is perhaps the most error-prone construction activity. Two reasons were offered by Ogunlana (1989) to explain the error-prone nature of construction cost estimating. First, this estimation depends on historical cost data. In construction projects, historical data may not fully reflect current conditions and, therefore, may not be a reliable indication for the future. Second, cost estimating attempts to predict future human actions in a world where things are never static.
Ashworth and Skitmore (1983) maintain that a vital consideration associated with any method of cost estimating is the accuracy with which anticipated costs can be predicted. In most studies, accuracy in terms of bias and consistency (Skitmore et al., 1990) is considered to be the major attribute involved, with an assumed need to minimize both bias and inconsistency of estimates (Cheung et al., 2008). Skitmore (1991) and Ogunlana and Thorpe (1987) have provided summaries of many of the past studies in terms of bias and consistency. According to Skitmore (1991), the former is the average of differences between actual bid prices, while the latter is the degree of variation around this average. A complete summary of parameters which have been shown in the literature to affect bias and consistency is provided in the study of Gunner and Skitmore (1999). According to this study, parameters that are associated with systematic change in bias comprise: (1) building function; (2) type of contract; (3) conditions of contract; (4) contract sum; (5) price intensity; (6) contract period; (7) number of bidders; (8) good/bad years; (9) procurement basis; (10) sector; (11) number of priced items and number of drawings; and (12) price forecast. Gunner and Skitmore (1999) classified parameters associated with systematic changes in consistency into 7 groups, including (1) building function; (2) contract work type; (3) contract sum; (4) price intensity; (5) contract period; (6) good/bad years; and (7) price forecast.

Several researchers have long expressed their concern about cost estimating inaccuracies by recognising that the accuracy achieved in cost estimating has been less than desirable (e.g., Ashworth and Skitmore, 1983; Ogunlana and Thorpe, 1987). Ogunlana (1989) attributed low accuracy to the nature of historical cost data, estimating method and the expertise of the estimator. Ashworth and Skitmore (1983) suggested that accuracy could be improved by exploiting the intuition of experienced cost estimators. Skitmore’s (1985) experiment with experts and novices showed significant differences between the accuracy of predictions made by the two groups. Skitmore (1985) concluded that experts are more relaxed and confident, more concerned with the market situation, and more able to recall costs of projects undertaken. According to Skitmore and Cheung (2007), experienced contractors also have a similar intuition which, in view of the large amount of uncertainty involved in predicting future costs at the bidding stage, is an important characteristic for their survival. Akintoye and Fitzgerald (2000) argued that the major causes of inaccuracy in cost estimates are the lack of practical knowledge of the construction process by those responsible for the estimating function, insufficient time to prepare cost estimates, poor tender documentation and the wide variability of costs incurred by subcontractors. Bowen and Edwards (1996)
determined this inaccuracy to be a consequence of potential communication problems between clients and cost estimators. As noted by Cheung et al., (2008), these problems are caused by different understandings of the cost message between clients and cost estimators and are an important aspect concerning clients’ satisfaction with cost advice given by cost estimators. Rather surprisingly, another key source of inaccuracy in construction cost estimates is associated with this satisfaction. As noted by Skitmore and Cheung (2007), construction industry folklore is that its consultants always aim to be ‘on the safe side’ when making construction cost forecasts on the basis that clients are much more likely to be concerned with underestimates than overestimates. Although giving accurate estimates seems to be essential (Cheung et al., 2008), researchers realised that clients prefer overestimated rather than underestimated forecasts (Skitmore and Cheung, 2007). This preference stems from the fact that underestimation would cause cost overrun, which may then necessitate a project to be re-designed and re-tendered, and in the worst case, to be abandoned (Cheung et al., 2008). In order to avoid cost overruns, several researchers found that there is a tendency for cost estimators to deliberately overestimate the actual costs involved (Cheung et al., 2008; Shash and Al-Khaldi, 1992). Despite this finding, the results of a recent study by Skitmore and Cheung (2007) indicated that overestimating, although often the case, is not always practised. McCaffer's (1976) analysis of estimating performance on some Belgian projects showed a distinction between the levels of underestimation in building contracts and road contracts. Also, Skitmore (1985) found a tendency to underestimate industrial projects and a tendency to overestimate commercial projects. Based on this finding, Skitmore (1985) concluded that overestimating is related to the complexity of projects and increased uncertainty levels involved. Another crucial factor influencing estimating accuracy is the state of the market, as demonstrated by Skitmore (1987). According to Skitmore (1987), contractors are willing to undertake less attractive projects, sometimes at a loss, in periods of low market activity. Skitmore's (1987) analyses showed that this factor may result in estimate differences up to 25% between low intensity and high intensity competition.

In summary, it is noted that cost estimating is an imprecise art (Ogunlana, 1989) as there are various factors affecting estimating accuracy. Following a thorough review of the related literature, Ogunlana and Thorpe (1991) classified these factors into three broad categories, being (1) project-based, (2) immediate environment-based, and (3) external environment-based categories. Factors included in the first category are type, size, and duration of a project, together with the level of information available. The second category
encompasses factors including number of bidders, ability of the estimator, local construction practices (i.e., geographical location of the project), resource availability, site access conditions, and price movements in the immediate environment. Finally, factors addressed by the last category are industry structure, state of the market, and price movements in the external environment. A schematic summary of these categories and their contributing factors is illustrated in Figure 2.1.

Figure 2-1: Factors affecting cost estimating accuracy (Ogunlana and Thorpe, 1991)

2.3 Level of Estimating Accuracy

Construction work is fraught with uncertainty and contains an element of risk (Ogunlana, 1989). Mak and Raftery (1992) asserted that forecasters should adopt estimating methods which explicitly deal with risk. Risk is amongst the factors identified by Skitmore et al., (1990) that contribute to the determination of construction price forecasts. According to Skitmore et al., (1990), forecasts provided by low risk forecasters are likely to be of poorer
accuracy when compared to the forecasts provided by high risk forecasters. Mak and Raftery (1992) demonstrated that clients tended to be risk averse in forecasting project return (conservative) and to be risk seeking in estimating project cost (speculative), which leads to the differences in personal risk perceptions and also generates the systematic biases in forecasting both project cost and return. In addition, de Neufville et al., (1977) indicated that contractors behave differently when dealing with small and large projects, and when operating in good years or bad, so that they are most risk averse toward larger projects in lean years and bid relatively lower. Cheung et al., (2008) investigated the satisfaction level of clients and cost estimators towards the accuracy of estimating. This investigation led to the finding that both clients and cost estimators are risk averse as they tolerate overestimates more than underestimates. This risk aversion behaviour was also demonstrated by Han et al., (2005) to be prevalent amongst clients as they dislike losses more than they like equivalent gains. Clients are more tolerant of an overestimated forecast because an underestimated forecast increases the risk of redesigning, retendering, and even abandoning a project (Cheung et al., 2008). As explained by Skitmore and Cheung (2007), overestimating is a means of risk reduction for cost estimators to be on the safe side so that they can protect and enhance their reputation with clients. In the construction industry, as asserted by Howard et al., (1997), risk aversion would prompt a contractor to charge a higher price to perform work subject to random income variation, even if the average deviation is expected to be zero in the long run. Howard et al., (1997) concluded that a risk averse contractor would prefer to have a smaller certain income than a slightly larger income on average that is subject to unpredictable and uncontrollable variations. Moreover, there will be a natural tendency to overestimate forecasts in cases where an organisation applies negative sanctions to managers who exceed budgets and takes no action of praise or reward when projects come in or under budget (Mak and Raftery, 1992). In such a case, estimators tend to be conservative (i.e., risk averse), reporting a rather high figure for project cost. Figure 2-2 illustrates project cost forecasts, showing the areas within which conservative and risk-seeking forecasts lie. According to this figure, a conservative or self-preserving estimate of project cost lies in the upper range of possible outcomes, while any estimate from the lower end of the range is regarded as a risk-seeking estimate. As noted by Mak and Raftery (1992), the choice of 50% as the cut-off point in Figure 2-2 is arbitrary, so that a client may choose a different cut-off value to reflect its own corporate risk attitude.
All of the studies reviewed above indicate that risk plays a central role in estimating practice. Concern has long been expressed over the level of accuracy of this practice, and therefore the reliability and usefulness of the forecasts in guiding design decisions (Gunner and Skitmore, 1999). Given this concern, a series of studies have been undertaken to define the levels of accuracy achieved in construction cost prediction and to determine a reasonable estimating accuracy level. Rubey and Milner (1966) proposed that contractors should estimate with error considerably less than 10% of their final cost. Park (1973) suggested that semi-detailed estimates prepared from rough quantity take-offs and suppliers’ quotation should be accurate to within ±10% of actual cost, while detailed estimates prepared from complete engineering specifications, drawings and site surveys should be accurate to within ±5% of actual cost. Barnes (1974) suggested that an accuracy of ±33% is likely in most estimates made for feasibility purposes. Barnes (1974) further suggested that, with contract estimates having an accuracy of between 5-7%, it is unlikely that design estimates can be better than 15-20% (when prepared by professional estimators) or 20-30% (when prepared by novice estimators). Following these suggestions, Barnes (1974) established a relation between estimating accuracy and the development of project design, indicating that there is an improvement in accuracy of cost estimates from limits of -40% and +20% at the commencement of feasibility studies to -20% and +10% at the commencement of detailed design. Keating (1977) reported that experienced estimators in process engineering contracts can achieve accuracy within 10-25% of total installed cost when provided with complete process design and equipment specifications. Marr (1977) proposed five levels of accuracy for adequate prediction of construction cost, being (1) 20-40% for planning control estimate;
(2) 15-30% for budget control estimate; (3) 10-20% for schematic stage control estimate; (4) 8-15% for preliminary drawing stage control estimate; and (5) 5-10% for construction drawing stage control. Ashworth and Skitmore (1983) and Ogunlana and Thorpe (1987) pointed out that suitable accuracy of cost forecasting in the early design stage would be of the order of 15% to 20%, improving perhaps to around 13% to 18% at the detailed design stage immediately prior to receiving tenders. Skitmore (1985) reported on the influence of expertise on the estimating accuracy. The results showed that mean error increases from -12.71% to +16.98% for non-experts and from -0.57% to 5.17% for experts, depending on the amount of information supplied. More recently, Lowe et al., (2006) noted that the accuracy of previous studies using traditional methods of cost estimation is typically in the order of 25%. A more comprehensive list of studies on the topic of level of estimating accuracy can be found in Ogunlana (1989).

The studies reviewed above indicate that there is a divergence of opinion on the issue of what level of accuracy is reasonable for cost estimation. Despite this divergence, all appear to agree that estimating accuracy improves as a project proceeds through its course of activities, as portrayed in Figure 2.3. In this figure, as explained by Skitmore and Cheung (2007), point A denotes the point at which there is a probability of 1:3 that the current cost estimate is 10% higher than the ultimate actual cost; point B is the actual cost; point C is the median of current estimates; and point D is the point at which there is a 1:3 probability that the current estimate is 20% lower than the actual cost. Taking these points into account, it becomes apparent that the inaccuracy in cost estimates narrows from the feasibility stage to the settlement stage, suggesting that estimating accuracy is highly dependent on the stage at which the estimate is made. According to Ogunlana (1989), it seems generally acceptable to quote figures between 10% and 20% for early estimates and of 5-10% for later stage estimates. Considering the results of Skitmore's (1985) study, it may also be argued that the deficit in error originates from estimates made by inexperienced estimators as professional estimators produce figures of less than 10%.
To conclude the discussion above, it is imperative to demonstrate that the importance of estimating accuracy is aligned with financial consequences involved. This importance may be well understood by Freiman’s curve illustrated in Figure 2.4. As can be observed in this figure, the final project cost is minimised when the estimate of cost is most accurate. This observation implies that financial consequences associated with cost estimation would be less dire, as the difference between the estimate and the actual value decreases. On the contrary, a serious problem may arise in situations where the estimate is considerably different from the actual value. Such a high difference can result in unfavourable financial consequences whose intensity varies depending on how low or high the estimate has been made. In order to eliminate the risk of an inaccurate estimate on project success, appropriate strategies need to be employed. These strategies may include risk mitigation measures (e.g., allocating sufficient time and resources to deliver a precise estimate) and risk transference measures (e.g., taking out a suitable insurance policy to cover the risk).
2.4 Cost Modelling

Estimation of the cost of a construction project is an important task for the management of construction projects (Adeli and Wu, 1998). The objective of cost estimating is to generate an indication of a project’s likely construction costs (Ashworth, 2004) in order to assist the client in setting a budget, predicting the tender price, and managing the design so that it meets the budget (Lowe et al., 2006). Construction clients require early and accurate cost advice, prior to site acquisition and the commitment to build, to enable them to assess the feasibility of the proposed project; this is a key task, performed by construction contract price forecasters (Lowe et al., 2006).

In reality, construction cost is not a deterministic function of construction cost governing attributes (Bode, 2000), but instead construction cost estimates fluctuate around the real construction cost values with an unknown pattern. Although it is not known if this fluctuation is really random, researchers invariably assume randomness to be the case in their modelling and analysis. Recognizing the variation pattern in construction cost data is a task that is complex and difficult to achieve by human or alternative simple computing tools, particularly when dealing with a large volume of data (Alex et al., 2010). Because of the uncertainties involved in mark-up estimation (see Chapter 1), the complexity of this task is
even more pronounced for the construction price offered by a contractor to perform a specified work for a client. Studies by Ellis & Turner (1986), Proctor et al., (1993) and more recently Lowe et al., (2006) have indicated that clients are often dissatisfied with the initial cost advice that they were provided with by their construction professionals. This dissatisfaction is mainly due to the fact that the cost estimators perform their services in a highly subjective manner (Adeli and Wu, 1998; Skitmore and Ng, 2003). Such a subjective analysis is subject to human errors and varying results depending on who the cost estimator is and possible litigation consequences (Adeli and Wu, 1998).

According to Fortune and Cox (2005), the formulation of a reliable cost forecast is at the core of professional practice, and irrespective of its origin it is necessary to ensure that the strategic cost advice provided to clients is of an acceptable quality level to enable them to make informed decisions on their value of money. In order to understand how cost advice can have its quality improved, it is necessary to first consider the cost estimation process itself. This process was identified by Skitmore (1991) in his key work on pre-tender design forecasts as being more than a mere technical calculation. Skitmore (1991) asserted that the dominating presence of uncertainty militates against the production of accurate estimates by numerical analysis alone. Skitmore was amongst the first to recognise that the process of cost estimation must include both the technical formulation and calculation of the advice and the application to it of the cost forecaster’s expertise or judgement. Bowen (1995) identified that there were intra-personal judgements made by forecasters on issues such as cost data and its relevance to project location and project complexity and that cost estimation processes were combined with the selection and application of appropriate cost models to produce quality early stage cost advice. Later, Fortune and Cox (2005) argued that cost practitioners may need to follow a four-stage process when preparing their cost advice, including (1) to assess the available project related information; (2) to select an appropriate cost model; (3) to consider results of the selected model; and (4) to apply their professional judgement so as to arrive at a forecast of the overall project cost. Given these studies, it seems reasonable to postulate that there should be an improvement in cost forecast performance commensurate with the level of expertise possessed by the forecaster. The effect of expertise should normally be derived from two sources (Ogunlana and Thorpe, 1991): (1) experts should be better than non-experts at sorting through information and deciding which is relevant for making forecasts, and (2) experts should be more knowledgeable than non-experts about the substance and process of estimating. As mentioned earlier, Skitmore’s (1985) experiment
with experts and novices showed significant differences between the performances in the two groups. A similar experiment was also performed by Ogunlana (1989) and the results showed distinctions between experienced estimators and novices. Ogunlana and Thorpe (1991) reported that experienced estimators are more confident in selecting information for estimating and more consistent in the reasons considered for the choice of historical cost data.

Although being beneficial, subjective judgement may lead to inconsistency and inaccuracy in the estimation process (Yau and Tsoi, 1998). Researching the reliability and value of various cost estimating methods, Fortune and Lees (1996) found that forecasts made by up to date methods were considered to be more valuable than forecasts made by using subjective cost estimating method. This finding would indicate that there is a potential for new cost estimating methods at the early stage of project planning (Soutos and Lowe, 2005). Subjective judgement, especially during the preliminary stages of conceptual design, is also prone to uncertainties that arise due to the lack of preliminary information, lack of sorted records of past costs, missing data, and a lack of appropriate cost estimating method (Harding et al., 1999). Lack of efficiency is another problem associated with the subjective cost estimating method. Such method does not enable any general truth to be identified about the relationships that exist between the cost and significant cost predictors, or provide any models which can forecast the cost (Lowe et al., 2006). Consequently, automating the process of cost estimating is a critical need not only to remove the subjective questionable human factors, but also to improve the efficiency as much as possible (Adeli and Wu, 1998). In trying to aid this critical need, the concept of cost modelling appeared (Soutos and Lowe, 2005).

The main intent of cost modelling is to simulate a current or future situation in such a way that the solutions posed in the simulation will generate results, which may be analysed and used in the decision-making process of design (Soutos and Lowe, 2005). Ferry et al., (1999) define cost modelling as a symbolic representation of a system, expressing the content of that system in terms of the factors, which influence its cost. According to De la Garza and Rouhana (1995), cost estimating is essentially a computational process that attempts to predict the cost of a future project, even though not all of the parameters and conditions are known when the cost estimate is prepared. This definition is also supported by Seeley (1996) who noted that the primary function of an early cost estimate is to produce a forecast of the probable cost of a future project, before the project has been designed in detail and contract particulars are prepared. It is therefore a genuine belief and hope that cost models would be
able to offer consultancies and contractors the opportunity to provide improved cost advice to clients (Soutos and Lowe, 2005).

2.5 Cost Modelling Approaches

The nature of cost is known to be uncertain (Newton, 1991). According to Hwang (2009), practitioners and researchers have long recognized the uncertainty in construction cost estimates and the need to improve the prediction capability of construction industry costs. To deal with this uncertainty, a range of modelling approaches is available today to cost estimators. Cost modelling approaches vary depending upon the available information, the nature of the project, and the time available to prepare the estimate (De la Garza and Rouhana, 1995). The choice of which cost modelling approach to adopt is also determined by its ease of application, familiarity, speed, and more importantly a satisfactory level of accuracy and reliability (Ashworth, 2004). Ashworth and Skitmore (1983) maintain that a vital consideration with any method of cost estimating is the accuracy with which anticipated costs can be predicted. While it is widely held that a perfect estimate is not possible and even the best possible estimate will always contain a number of key risks, the goal of cost estimating is a practicable level of accuracy (Smith, 1995). A detailed summary of cost modelling approaches can be found in a review by Newton (1991) and in more recent studies by Seeley (1996) and Fortune and Lees (1996). In the following sections, two approaches used in the literature to solve cost-related problems are presented.

2.5.1 Parametric Approach

Parametric cost estimating models are useful at the early stages of a project’s life cycle when little information is known about the project’s scope (Hegazy and Ayed, 1998). According to De la Garza and Rouhana (1995), parametric cost estimating methods have been used in the past in applications involving the determination of order of magnitude or preliminary cost estimates in a short period of time and with no or little specific information about the system that is being evaluated. These methods use historical cost data from previous projects to establish a cost estimating relationship capable of generating cost estimates for future projects of the same type. Parametric cost estimating methods often lead to a mathematically fitted function called a cost estimating relationship, which is a functional
model that mathematically describes the cost as a function of one or more cost predictors (Ostwald, 1984). In parametric costing, the cost engineer needs to predetermine a functional relationship between the cost and its predictors based on his or her experience (Bode, 2000). Such relationships are predominantly formulated by applying regression analysis to the available past historical data. In their survey, Mason et al., (1994) found that professional cost estimators regularly use regression analysis to develop their cost models. Regression analysis was also recognised by Fortune and Lees (1996) as a non-traditional modelling technique available to cost estimators. Despite this recognition, and despite the previous finding by Mason et al., (1994), Fortune and Less (1996) concluded that traditional techniques of early cost advice (e.g., judgement, unit rate, elemental analysis, approximate quantities) are more popular with practitioners than are non-traditional techniques (e.g., regression analysis). However, Fortune and Lees (1994) stated that as the size of an organization increases, the more likely that organization would be to use alternative costing techniques. The overwhelming use of traditional types of forecasting models was also demonstrated in a recent study by Fortune and Cox (2005). The findings of this study revealed that traditional cost estimating techniques continue to be in widespread use irrespective of organisational type.

Researchers from many scientific disciplines have been proposing regression models to solve a variety of problems. Early examples of the use of regression analysis as a parametric cost estimating tool are provided by Hadfield (1974), Cochran (1976), Putnam (1978), and McCaffer et al., (1984). Sigurdsen (1992) used linear regression analysis to determine the cost of carbon steel pipes as a function of pipe diameter, number of elbows, and flange rating. De la Garza and Rouhana (1995) adopted the database and variables used by Sigurdsen (1992) to present a non-linear regression model for the same purpose. In another study in this field, Shtub and Versano (1999) studied four types of steel pipes to predict their respective bending costs by means of regression analysis. Shtub and Versano (1999) also described the paramount role of sound cost estimations in the manufacturing of new products and emphasized that the decision to replace a system or a product is mainly based on the estimated cost of its future operation and maintenance. Shtub and Dar-El (1989) and Shtub and Zimerman (1993) used the regression technique for estimating the expected cost of assembly systems for manufacturing any product that consists of two or more parts. In another study, Bode (2000) developed regression models to support cost estimation of manufacturing personal computers at an early stage of product development. According to
Smith and Mason (1997), the regression technique has been also used to support many cost estimation relationships in diverse applications including design for manufacturability, maintenance scheduling in power plants, capital and operating cost equations in southwestern U.S. mining operations, software development costs, equipment and tooling configurations in plastic moulding, query costs in databases, and the costing of pressure vessels for new chemical production.

In addition to the fields stated above, the use of regression analysis in the construction industry is also well established in the literature. This type of analysis has facilitated the development of parametric models for estimating construction cost. A parametric cost estimating model consists of one or more functions, known as cost estimating relationships, between the construction cost as the dependent variable and the construction cost governing factors as the independent variables. According to Hegazy and Ayed (1998), cost estimating relationships are traditionally developed by applying regression analysis to historical project information. Kim et al., (2004) asserted that regression cost models have been used since the 1970s for estimating cost because these models have the advantage of a well-defined mathematical basis as well as incorporating measures of how well a model matches a given dataset. This assertion was also supported by a recent study by Shehab et al., (2010), stating that regression analysis has been successfully utilised in the development of many cost estimating models. Wilmot and Mei (2005) noted that regression cost models are often developed in the linear format. Application of non-linear regression is more complex because a user must have detailed knowledge about the appropriate non-linear relationship (Christensen, 1996), which is rarely the case in practice. Consequently, the user is required to carrying out many simulations with different types of non-linear functions until the best function is found. This requirement is even more pronounced when the cost impact of cost-related attributes is not known (Bode, 2000). A review of the early applications of regression analysis to construction cost forecasting is presented by Skitmore and Patchell (1990). The importance of models to estimate the cost of buildings has been highlighted by Ferry et al., (1999). The following are some of the studies being recently carried out by researchers in various construction areas to generate cost estimating models using regression analysis.

Khosrowshahi and Kaka (1996) identified the need for the development and use of regression models for appraising construction cost. They studied a number of variables which act individually and in combination to influence the construction cost and introduced a simply-applied yet reliable regression model for the estimation of total construction cost for
housing projects in the U.K. Elhag and Boussabaine (1999) utilized gross floor area together with project duration to estimate the tender price of office buildings using the regression technique. Emsley et al., (2002) described the development of regression models to predict the construction cost of buildings, based on input data from building projects in the UK. Following this study, Soutos and Lowe (2005) applied regression analysis to a database comprised of elemental cost breakdowns to investigate the way that a series of building characteristics affect the cost of building elements such as substructure, superstructure, finishes, and services. Building characteristics, in addition to construction duration, were also employed by Kim et al., (2004) for developing regression models to estimate the construction cost of residential buildings in Seoul, Korea. More recently, Stoy and Schalcher (2007) performed a similar study to support early cost estimates for residential properties in Germany via the use of regression analysis. In this study, relevant predictors of the building construction cost were identified and included in the proposed regression model. These predictors were the median floor height, the share of the ancillary area for services, the construction duration, and the compactness of the building.

The predictive regression models reviewed above were developed to estimate the construction cost at the contract stage. In many cases, however, the actual construction cost can be different at the end of the projects than was expected at the beginning (i.e., the contract cost) (Love et al., 2005; Skitmore and Ng, 2003). Rework and variation orders issued by the owner, consultant, or claimed by the contractor are the primary reasons for such deviations (Abu Hammad et al., 2010; Love et al., 2005; Williams, 2002). According to Williams (2002), models that can provide predictions of the magnitude of the final cost would allow owner organizations to better plan and budget for the actual costs of construction projects. Williams et al., (1999) found that simple regression models produce reasonable predictions of the final project cost for competitively bid highway construction projects in New Jersey, USA, and in the UK. The same conclusion was also derived from other studies by Williams (2002) and Williams et al., (2005), in which regression models were developed to predict the completed cost of highway projects constructed by the New Jersey Department of Transportation and the Texas Department of Transportation, respectively. The ability of regression models as a means of capital budgeting was also demonstrated in a study by Williams (2003). The results showed that an owner organization with a large number of projects in a given period can input the sum of the low bids of a group of projects and find a reasonable prediction of the ultimate total cost for the group. This prediction can serve to
enhance the budgeting of construction agencies. In another study, Skitmore and Ng (2003) utilized the contract details of 93 Australian construction projects to develop several regression models for predicting the actual construction cost. Such details included the contract time, the contract cost, client sector, project type, contractor selection method and contractual arrangement. The results of this study indicated that the regression models could assist clients and contractors to predict the actual construction cost based on the contract details as mentioned. The same conclusion was also drawn in a more recent study by Abu Hammad et al., (2010) using real historical data of 113 public building projects in Jordan. A regression model was proposed in this research to estimate the actual project cost based on a series of predictors, including floor area, project type, project budgeted cost, and planned project duration.

The studies stated above were all concerned with the application of regression analysis to the problem of cost estimating in new construction projects. Regression analysis has been also tried to solve cost-related problems in other construction domains. For instance, Attalla and Hegazy (2003) showed that regression analysis is a viable tool for predicting cost deviation in reconstruction projects. Chen and Huang (2006) described the development of regression models to predict the reconstruction cost of schools damaged during the 1999 Chi-Chi Earthquake in Taiwan. Yang and Su (2007) regressed the construction cost of trenchless and open-cut sewer rehabilitation projects on two parameters, being the diameter and the burial depth of pipes. In an effort to address the limitation of the latter study, Shehab et al., (2010) used regression analysis to estimate the cost of water and sewer network repair projects by processing more cost influential parameters.

Improved budgeting decisions through predicting the change of resource prices is another area explored by some researchers using regression analysis. The ability to predict trends in prices can potentially enable contractors to incorporate expected price fluctuations into their bidding strategy, leading to more accurate bids (Williams, 1994). Williams (1994) developed regression-based models to predict change in construction cost indexes. This work was recently followed by Hwang (2009) who developed dynamic regression models for forecasting the change of construction costs as a response to economic conditions in the market. These two studies illustrated the use of regression analysis to develop a simple, yet effective means of predicting a construction cost index.
All the studies reviewed above confirmed the popularity of regression analysis to develop cost estimating relationships. Despite this popularity, regression analysis is limited in its estimating ability (Wang et al., 2000). This limitation is caused by certain assumptions made during development of the regression models. As stated earlier, the general mathematical function of the cost estimating relationship has to be defined before the regression analysis can be performed. Finding a properly fitted approximation function is extremely important; otherwise, the predicted cost will have a substantial error (Adeli and Wu, 1998). This requirement accounts for the major disadvantage of regression-based cost models (Bode, 2000) as there is no specific or clearly-defined approach available to find a mathematical function capable of describing the cost models. Defining an appropriate regression functional form becomes more complex as the dimensionality of the variables to be considered in the cost problem grows (Smith and Mason, 1997). Another consideration when using regression analysis is that cost estimating relationships are often modelled according to a single mathematical form, such as linear or nonlinear. In many problems, however, the cost predictors used in the cost modelling often have a different mathematical correlation with the cost (De la Garza and Rouhana, 1995). For example, the change in cost may be a linear function of changes in a particular cost predictor, and a non-linear function for other predictors. These considerations may result in low accuracy of the regression-based cost models.

In addition to the drawbacks stated above, the ordinary least squares (OLS) method (used in the development of MLR models) was demonstrated to be inappropriate for predicting construction cost. The main reason for this inappropriateness is that this method is likely to result in an asymmetric loss function. Skitmore and Cheung (2007) define the loss function as a measure of prediction accuracy in the form of the cost of making errors. Therefore, as asserted by Skitmore and Cheung (2007), by allocating costs to prediction errors, the loss function characterises the accuracy of a prediction. It is well known in the statistics literature that the choice of this function is critical for model estimation and evaluation (Christoffersen and Jacobs, 2004). For construction cost prediction, loss functions that have been mainly utilised with the predominant use of the ordinary least squares (OLS) technique include the relative error \( \frac{Y_{predicted} - Y_{Actual}}{Y_{Actual}} \), relative absolute error \( \frac{|Y_{predicted} - Y_{Actual}|}{Y_{Actual}} \), percentage error \( 100 \times \frac{Y_{predicted} - Y_{Actual}}{Y_{Actual}} \), ratio error \( \frac{Y_{predicted}}{Y_{Actual}} \), and the log ratio error \( Ln \frac{Y_{predicted}}{Y_{Actual}} \), with their mean and standard deviation being taken as...
measures of bias and inconsistency, respectively. According to Skitmore and Cheung (2007), these functions are arbitrary and possibly result in biased forecasts. Typical measures of goodness of construction cost predictions (e.g., mean, standard deviation, coefficient of variation, mean square errors) can be valid only if the loss function determined from the deviations between predicted and actual values is directly proportional to the square of their values. Skitmore and Cheung (2007) used two approaches to test this validity. The first approach was to analyse the deviations that occur in practice, based on the assumption that cost predictors are sufficiently skilled to have inbuilt the loss functions into their predictions. In the first approach, 10 datasets collected from Australia (2 datasets), Belgium (2 datasets), Hong Kong (2 datasets), Singapore (1 dataset), United Kingdom (1 dataset), and United States (2 datasets) were analysed. The second approach was to try to obtain the loss function directly from clients. In the second approach, the use of a postal questionnaire survey was explored in Hong Kong to elicit clients’ perceptions on the quality construction cost predictions. The results of the first approach mitigated against any general view that projects tend to be overestimated but did clearly show asymmetric under/overestimates for the statistics used. In particular, the distribution of under and overestimates were each found to approximate to separate normal density functions (but with different parameters), with the variance of overestimates trending downwards with increasing project size (value), while the variance of underestimates remained constant. The second approach resulted in an ordinal-based approximated loss function, varying between building types. The most asymmetric loss function was found for commercial and residential projects, while the loss function of school and industrial projects was less asymmetric. The second approach described above was essentially based on the attitude of clients towards estimating errors. Cheung et al., (2008) extended this approach to also include the estimators’ perceptions on the quality of estimates that they produced. The respondents were asked to indicate the extent to which they had experienced underestimates and overestimates. The results demonstrated a more extreme asymmetrical distribution between underestimates and overestimates than was shown in the study of Skitmore and Cheung (2007). It was also found that both clients and estimators are risk averse, being more tolerant of overestimates than of underestimates. In addition, pairwise comparison test results revealed that clients and estimators have similar tolerance in general for estimating errors for the set of commercial and residential projects, and for the set of school and industrial projects. Finally, by comparing the disutility of overestimates to that of underestimates, it was concluded that clients have a stronger preference for overestimates for the former set of projects than for the latter.
2.5.2 Non-parametric Approach

The non-parametric approach is of great potential to support cost estimating problems. In non-parametric costing, no assumption about the shape of the cost approximation function has to be made before carrying out the estimation (Bode, 2000). The cost engineer therefore need not get involved in a task which is of utmost importance in parametric costing; that is, predetermining a priori a sufficiently accurate cost function. Offering an alternative approach for cost estimating, the artificial neural network (ANN) technique is a new breed of non-parametric estimators that has recently evolved based on artificial intelligence.

ANN models have been introduced into civil engineering applications as an alternative to traditional approaches (Chen and Huang, 2006); however, the application of ANNs to construction is a relatively new research area (Boussabaine, 1996; Kim et al., 2004). In the construction management field, ANNs can be seen as components of larger systems which make use of expert-given rules or statistical inference techniques as required (Boussabaine, 1996). Such systems, in turn, are able to provide decision support for experts, help decision makers perform at a higher level, assist in the training of inexperienced personnel and help scenario planning (i.e., what if?) by managers (Boussabaine, 1996). ANNs are primarily used within the construction management field for difficult tasks involving intuitive judgement or requiring the detection of data patterns that elude traditional analytical techniques (Hegazy and Ayed, 1998). In a similar way, the application of ANNs with respect to cost estimation has been particular to complex projects as the early-stage cost estimation is still largely based on a combination of simple models and professional judgement (Alex et al., 2010). Nevertheless, the use of ANNs during the conceptual phase of project development can reduce inaccuracies in cost estimation and, as a result, provide more accurate cost estimates. These estimates are crucial for planning and feasibility studies (Sodikov, 2005), while also enabling users to better understand the financial aspects of construction management (Alex et al., 2010). It is therefore believed that ANNs are useful tools that can provide a practical solution to cost estimating problems in an accurate and objective way (Shtub and Versano, 1999) and that ANNs have the potential to eventually be regarded as suitable and as practical as spread sheets are today for cost engineers (McKim, 1993b).

In recent years, several authors have discussed potential applications of ANNs within the fields of construction management in general, and construction cost in particular. Moselhi
et al., (1991, 1992) provided a detailed description of the fundamentals of ANNs, along with their potential applications in civil and construction engineering. Boussabaine (1996) further reviewed the most important applications of ANNs in construction management topics and provided guidance and tips for the successful development of ANNs in those topics. The potential applications of ANNs to problems particularly associated with cost estimation were investigated by De la Garza and Rouhana (1995).

One of the first applications of ANN methodology in the field of cost engineering was McKim’s (1993a) ANN model to predict project cost overruns. In another study, McKim (1993b) discussed some aspects of ANNs as a new computer technology that may emerge as a tool to be widely used by practicing cost professional in their day-to-day functions. Following this discussion and in order to provide an illustrative example, McKim (1993b) described the development of an ANN model to estimate the cost of pumps. Shtub and Zimmerman (1993) demonstrated the potential value of using ANNs to estimate the expected cost of six major types of assembly system. Moselhi et al., (1993) used ANNs to estimate the expected cost of carbon steel pipes. Adeli and Wu (1998) formulated a regularized neural network and then presented an ANN model for estimating the cost of construction projects. The model was applied to estimate the cost of reinforced-concrete pavements as an example. Liu (1998) demonstrated application of the ANN technique in capital project risk analysis by developing ANN models to predict cost and time variations at the front end stages of projects. Hegazy and Ayed (1998) used the ANN technique in order to effectively manage construction cost data of eighteen actual highway projects (i.e., bridge, highway, others) constructed in Newfoundland, Canada. Elhag and Boussabaine (1999) applied the ANN technique to the problem of estimating the tender price of school buildings. Later, Elhag and Boussabaine (1999) extended their work to predict the tender price of office buildings, using the ANN technique.
Shtub and Versano (1999) employed the ANN technique as a response to the increasing intensity of cost-based competition in the manufacturing industry, developing an ANN model to predict the cost of steel pipe bending. As one of the last studies from the late 20th century, Al-Rashidi (1999) showed that the drilling cost of new wells can be effectively predicted by ANN models.

Over the last decade, interest in using ANN models within the field of cost estimation has increased even more. Bode (2000) used pilot cost data from a manufacturing company to investigate the applicability of ANNs for estimating a product cost in early phases of the product design. Attalla and Hegazy (2003) investigated the challenging environment of reconstruction projects and described the development of an ANN model for predicting cost deviation in such projects. Williams (2002) developed ANN models to predict the completed cost of competitively bid highway projects constructed by the New Jersey Department of Transportation in the USA. This study was followed by another study by Williams et al., (2005) in which ANN models were developed to predict the completed cost of Texas highway projects. Georgy and Barsoum (2005) established an ANN model by which the school construction costs in Egypt could be reasonably estimated. In another attempt within the school sector, Chen and Huang (2006) developed ANN models to predict the cost of school reconstruction projects in central Taiwan which were in need of being quickly rebuilt after the Chi-Chi earthquake. More recently, Shehab et al., (2010) employed the ANN technique to develop cost estimating models based on data from a set of 54 sewer and water rehabilitation projects in the United States. There are more research studies reported in the literature that had as their main objective to provide ANN-based cost estimation models. Provided below is a brief description of some of these studies, including a summary of results from each of these studies.

Emsley et al., (2002) examined application of the ANN technique to the prediction of total construction costs. Emsley et al., (2002) undertook research that was carried out in two stages. The first stage included an initial pilot study where potentially cost significant variables were identified, the availability of data investigated and strategies for data collection were established. In the second stage, the full data collection program was performed, resulting in the compilation of 288 projects. These projects were obtained from industrial collaborators as well as other quantity surveying and project management practices that were willing to provide information. Data was also generated from the online service of Building Cost Information Service (BCIS). This study resulted in the production of an early
stage building cost modelling software, known as ProCost. This software is a powerful cost estimation package producing single figure estimates of the total cost for any kind of proposed buildings.

Kim et al., (2004) examined the performance of the ANN technique for estimating construction costs. Although the results showed that ANN models were appropriate for estimating construction costs, the study concluded that establishing the best ANN model is quite a cumbersome task and therefore in addition to the accuracy of estimation results, time and accuracy trade-offs should be considered in the cost estimating model. Following this study and as a practical solution for this consideration, Kim et al., (2005) proposed hybrid models of ANNs and genetic algorithms (GAs). Data used in the latter study were the costs of 498 projects of residential buildings built by general contractors from 1997 to 2000 in Seoul, Korea. After a preliminary study using all potential cost influential factors, which were extracted from the historical data, construction costs were found to be affected by several factors including total floor area (m²), number of stories, total units, duration (months), type of roof, type of foundation, type of basement and grades of finishing. These factors were then incorporated in the establishment of a number of models integrating ANNs with GAs. The comparison of these models showed that GAs may help estimators overcome the problem of the lack of adequate rules for determining the parameters of ANNs.

Wilmot and Mei (2005) utilized ANNs in the development of a procedure for estimating the escalation of highway construction costs over time. The data used in this study was extracted from highway construction projects contracted out between 1981 and 1997 by the Louisiana Department of Transportation and Development (DOTD). The data consisted of information on a total of 2,827 highway and bridge contracts. A particular cost index was created in this study, which was used as a measure of overall highway construction cost. The index has been named the Louisiana Highway Construction Index (LHCI) and was defined with the contribution of five main construction categories, called sub-models. These sub-models were excavation and embankment, concrete pavement, asphalt pavement, reinforcing steel and structural concrete. Following the selection of representative pay items in each of the five sub-models, an ANN model was developed for each sub-model. The results from sub-models were then combined in a composite model to provide an overall measure of LHCI. Results demonstrated that this composite model was able to replicate past highway construction cost trends in Louisiana with reasonable accuracy.
Alex et al., (2010) described an ANN methodology to estimate the costs pertaining to the installation of water and sewer services in the City of Edmonton, Canada. The proposed methodology was based upon the analysis of over 800 projects obtained from the City of Edmonton’s drainage division for the period of 1999-2004. Following the extraction of data, various types of ANNs were examined based on a trial and error approach. The best ANN model was then selected based on statistical criteria applied in this study. This model incorporated data related to the average monthly temperature over the various years, the geographical location of the project, the consistency of the number of estimations performed, and features related to the site conditions. The results showed that the best ANN model reduced the variation between estimated and actual costs and reached a prediction accuracy of 80%. This accuracy was reported to be better than the accuracy gained through using the SmartEST estimating software which was commonly used by the City of Edmonton. The software was originally designed by the University of Alberta’s Construction Engineering and Management Group. The authors argued that the best ANN model developed in their study was general enough to be adapted by other estimating software if using the same input factors.

As a concluding remark for this section, it is worth noting that the findings of recent studied (e.g., Farooq, 2007; Fortune and Cox, 2005) revealed that ANN models have only limited application in practice. As with any emerging technique, there is a general feeling of mystery and unease surrounding the implementation of the ANN technique that can be expected to remain until this technique has been proven effective in practice (McKim, 1993b). A number of studies have recently attempted to enhance the practicability of this technique in a professional environment. For instance, Emsley et al., (2002) examined application of the ANN technique to the prediction of total construction costs in the UK. This study resulted in the production of an early stage construction cost modelling software, known as ProCost. This software is a powerful cost estimation package producing single figure estimates of the total cost for any kind of proposed buildings (Soutsos and Lowe, 2005). In another study, Attalla and Hegazy (2003) developed an ANN-based decision support tool to quantify the cost deviation in reconstruction projects. Facilitating the use of ANN models, this tool automates the prediction of the cost deviation for reconstruction projects as a function of the set of control variables specified by the user. Doğan (2005) developed ANN models in the Excel spreadsheet format for the prediction of unit structural cost of residential buildings. Demonstrating the practicality of using spreadsheets in developing ANN cost
prediction models, Doğan (2005) argued that the ANN Excel templates developed can be modified to be used in other facets of the construction industry such as quality, productivity, constructability, value engineering, and scheduling. Farooq (2007) developed an ANN-based cost estimating system for water and sewer repair and replacement projects in the United States. According to Farooq (2007), the developed system offers a multitude of advantages, including (1) reducing the time required for preparing budgetary estimates for this class of projects, (2) improving the accuracy of budget estimates, (3) preventing problems that are usually associated with inaccurate budget estimates such as withholding projects for future fiscal years or constructing projects below optimum standards, and (4) enabling officials to generate cost effective proposals to acquire the funds for future projects.

Specific technical aspects that determine the performance of an ANN model must be carefully researched further, as the process of developing and implementing ANNs for cost estimating applications has a number of problems associated with it (see Chapter 6). These problems make the development of an ANN model for a particular problem a difficult task that is neither simple nor straightforward (Bode, 2000; De la Garza and Rouhana, 1995; Hegazy et al., 1994; Liu, 2006). This difficulty arises in part from the fact that there is a significant lack of guidance available on the development of an optimal ANN model. Consequently, the process of developing such models often requires iterative refining of network parameters, network redesign, and problem reformulation (Attalla and Hegazy, 2003; Hegazy et al., 1994; Kim et al., 2004; Kim et al., 2005; Liu, 2006; Wang et al., 2000). These requirements represent formidable impediments to widespread use and acceptance of ANN-based cost estimation relationships (Smith and Mason, 1997). Also, the common perception of ANN models as “mystical black boxes” (Somers, 1999) may increase concerns regarding the appropriateness and legal defensibility of these innovative models. Researchers are therefore encouraged to advance the current body of published research not only to help inform future applications of ANNs, especially when the prediction is of high concern (Ostberg, 2005), but also to refine the technical aspects of the ANN technique in areas where ANNs can become an effective cost engineering tool.
2.6 Retrofit Cost Estimating (RCE) Models

As described in the previous chapter, one of the major challenges faced by retrofit projects is the ability to successfully and accurately predict the incurred cost of such projects. The cost of retrofitting existing buildings is difficult to estimate (FEMA 228, 1992) due to the large variation in individual building conditions (FEMA 227, 1992). The uniqueness and increasing complexity of retrofit projects also make the cost estimation of such projects more complicated than for other construction projects such as new construction or renovation projects. In the literature, the modelling of retrofit cost has received comparatively little attention by researchers. This paucity may be in part explained by the fact that the study of seismic retrofit of existing buildings began relatively recently. The following sections present the available retrofit cost models found in the literature.

2.6.1 Potangaroa RCE Model

Potangaroa (1985) developed a simple retrofit cost model for unreinforced masonry (URM) buildings. The database used for the development of this model was adapted from a survey in the United States. The cost database was made up of 70 buildings from Veterans Administration (VA) and 16 schools from Los Angeles City Unified School District. Most of the cost estimates for the VA were for proposed strengthening from 1973 to 1975. All estimates included the cost of structural strengthening, restoring finishes, relocating mechanical and electrical fixtures and overhead. No cost for other retrofit works was included in the cost estimates. In the case of LA schools, their costs were over a fifteen year period and, as a result, these costs were adjusted to 1975 US dollar. A further US$2 per square foot was added to the adjusted cost based on an estimate from the LA City Unified School District to allow primarily for replacing the finishes on repaired walls and floors. These costs were then comparable to the VA costs.

Potangaroa (1985) studied the cost database, using the following variables:

- \( C_1 \): The total floor area (ft\(^2\))
- \( C_2 \): The building age (year of original construction)
- \( C_3 \): The design level
- \( C_4 \): The number of floors.
The equation of best fit relating the variables above to the cost of retrofitting ($US/ft^2) was proposed using the regression analysis as follows:

$$C = -7.0265 + (0.000018).C_1 + (0.00767).C_2 + (10.5).C_3 + (0.858).C_4 \quad (2.1)$$

A sensitivity analysis was also undertaken to find those variables that had the most significant impact on the retrofit cost. The results showed that the most significant cost variables were the number of floors and the design level, as represented by peak ground acceleration (PGA). The retrofit costs of two-storey and three-storey buildings were 8% and 16% more expensive than that of single storey buildings, respectively. The building floor area became a significant variable when the number of floors was more than three. The number of floors in the research database was, however, limited to five because URM buildings with more floors were rare. The second significant cost variable was the design level, as a large increase in the PGA from 0.08g to 0.21g increased the relative cost from US$10.5/ft^2 to US$12.0/ft^2, which is an increase of 14%. Building age was found to be the most insignificant cost variable as an increase in age from 1893 to 1934 only resulted in a relative cost rise of 3%.

2.6.2 FEMA Typical RCE Models

The first attempt in the United States at gathering a comprehensive set of retrofit costs was completed in 1988, resulting in the first edition of FEMA 156 and FEMA 157. The cost database consisted of 614 data points (i.e., buildings) and most of the original data points were derived from rather limited studies. The data were mainly related to costs incurred in seismic mitigation of URM structures. The second edition of FEMA 156 (1994) represented a marked improvement over the first edition as it combined an extensive data collection with a rigorous quality control process, resulting in the creation of an extended database of 2,088 data points. This amount of data was collected with the contribution of almost 50 groups or firms and from many different regions of the United States and Canada. The data were also from different building model types and for different levels of expected seismic performance. The data prior to 1970 were eliminated from the database due to their high degree of uncertainty. A crucial aspect of the second edition of FEMA 156 (1994) was to convert all the cost data into the same relative time and location. Cost data were all converted to 1993
US dollar for the State of Missouri. Missouri was picked as the location because it is located at approximately the geographic centre of the United States. Following this conversion, an additional level of quality control was applied. Bounds for this additional quality control for each building group were computed based upon the theory of probability. These bounds were calculated by assuming a lognormal probability density function (PDF) for the cost data. All data that lay outside these bounds (a total of 5% of the data points) were sent back for re-evaluation and possible elimination from the database (Hart and Srinivasan, 2008). The suitability of using a lognormal distribution for the cost data was studied by taking the natural logarithm of the cost data for different building types. The lognormal distribution was found to be a suitable distribution to represent the probabilistic behaviour of the cost data. The lognormal PDF also served as a useful tool to judge whether the costs for different building types behave similarly for various combinations of seismicity and performance objectives. Based on the PDF curves, the original fifteen building types in FEMA were eventually grouped into eight classes.

The next step towards the development of FEMA typical RCE models was to rate the quality of each data remaining in the cost database. To identify the most accurate data, a quality factor was calculated for each cost data point. The values of the quality factor ranged from 1 to 10, with 1 being the least accurate and 10 the most accurate. According to Hart and Srinivasan (2008), the parameters being determined to be the most important in rating the quality of the data were: the date of the cost data, the source of the cost information, and the overall quality and consistency of the data. A point value was assigned to each of these parameters and the total summed for each cost data point to give the overall quality rating of that particular cost data. This overall quality factor was used to develop the “super database” using the Monte Carlo simulation (MCS) technique. In this technique, the number of points simulated for the best quality data point (i.e., the quality factor of 10) was about twice as large as the number of points simulated for the worst quality data point (i.e., the quality factor of 1). The maximum number of cost data points created in the super database for a given building group was 12,000 for each combination of seismicity and performance objective. Considering the total number of building groups, the maximum number of cost data points in the super database was to be 96,000. However, the actual number of data points in the super database was diminished to 83,000 because not all combinations of performance objective and seismicity in a given building group had data points from the cost survey. Each of the
eight building groups, together with their corresponding number of buildings in the original and super database, is shown in Table 2-1.

Table 2-1: Number of data points in each building group (FEMA 156, 1994)

<table>
<thead>
<tr>
<th>Building group number</th>
<th>Building group</th>
<th>Building type</th>
<th>Cost data points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Original database</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>URM – Unreinforced masonry</td>
<td>642</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>W1 – Wood light frame W2 – Wood (commercial or industrial)</td>
<td>164</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>PC1 – Precast concrete tilt up walls RM1 – Reinforced masonry with metal or wood diaphragm</td>
<td>171</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>C1 – Concrete moment frame C3 – Concrete frame with infill walls</td>
<td>372</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>S1 – Steel moment frame</td>
<td>160</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>S2 – Steel braced frame S3 – Steel light frame</td>
<td>97</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>S5 – Steel frame with infill walls</td>
<td>116</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>C2 – Concrete shear wall PC2 – Precast concrete frame with concrete walls RM2 – Reinforced masonry with precast concrete diaphragm S4 – Steel frame with concrete walls</td>
<td>366</td>
</tr>
</tbody>
</table>

FEMA 156 (1994) permits estimates to be made at several levels of sophistication, based on the level of data input into the analysis. FEMA 156 (1994) provides three models for estimating retrofit cost in terms of US$/ft² as follows:

\[
\text{Model one: } C = C_1 \cdot C_2 \cdot C_L \cdot C_T \\
\text{Model two: } C = C_1 \cdot C_2 \cdot C_3 \cdot C_L \cdot C_T \\
\text{Model three} = C_C \cdot (\text{area})^{X_1} \cdot (\text{age})^{X_2} \cdot (\text{no. of stories})^{X_3} \cdot X_4 \cdot X_5 \cdot X_6
\]  

where,

\( C_I \): Building group mean cost


$C_2$: Area adjustment factor

$C_3$: Seismicity/performance objective adjustment factor

$C_L$: Location adjustment factor

$C_T$: Time adjustment factor

$C_C$: Statistically based constant

$X_1$, $X_2$, and $X_3$: Statistically based variables whose values depend on the building group

$X_4$: Statistically based variable whose value depends on the building seismicity/performance objective and the building group

$X_5$: Statistically based variable whose value depends on the building occupancy class and the building group

$X_6$: Statistically based variable whose value depends on the occupancy condition during seismic retrofit and the building group

The above retrofit cost estimating (RCE) models enable the end user to predict “typical retrofit cost” which, by definition, is the “mean structural construction cost”. These RCE models do not include other types of retrofit cost and also should not be used to estimate the retrofit cost of heritage buildings. The taxonomy of retrofit costs is described in the next chapter.

The RCE models, represented by Equations (2.2) to (2.4), increase in complexity as they incorporate more information about a building. The typical retrofit costs obtained from Equation (2.2) are deemed adequate for only very general discussions of potential retrofit costs for large inventories. The typical retrofit costs derived from the use of Equation (2.3) are deemed accurate enough for planning purposes and only when considering multiple buildings. The user obtains the most mathematically rigorous results for typical retrofit costs through the use of Equation (2.4). This equation is also the most accurate model for appraising the retrofit cost of individual buildings. Equation (2.4) is also different from Equations (2.2) and (2.3) in that it uses the results of multi-linear regression (MLR) analysis to match the predicted and actual costs in the super database as closely as possible. According
to FEMA 157 (1995), Equations (2.2) and (2.3) rely on the use of step functions derived from compartmentalization of the data for their ease of use. This compartmentalization means that all the data were sorted into groups of tractable size. For example, for consideration of the building area, the data were classified into four groups (i.e., small (less than 10,000 ft²), medium (10,000 to 49,999 ft²), large (50,000 to 99,999 ft²), and very large (100,000 ft² or greater)). Data at the extreme ends of a given group were assumed to behave similarly to data more centrally located in the group. This assumption may not reflect the true behaviour of the retrofit costs. However, in Equation (2.4) the building area, together with other numerical variables, were assumed to be continuous variables rather than to be step functions, enabling the end user to compute the typical retrofit cost of an individual building with the maximum amount of information.

2.6.3 Other RCE Models

FEMA 227 (1992) provides standard benefit-cost models that could be used throughout the United States to help practitioners and government officials evaluate the economic benefits and costs of seismically retrofitting existing buildings. The retrofit cost of a particular building type for use in the benefit-cost models was derived from both a review of the existing literature and the information on retrofitted buildings obtained during the field data collection phase of this project. Based on the best available information at the time of this project, a range of typical retrofit costs was proposed for various building types. These costs are defined in FEMA 227 (1992) as “hard costs”. According to FEMA 228 (1992), the hard cost is the bid cost for labour and materials for the seismic portion of a retrofit work and includes a component for restoring architectural finishes. FEMA 228 (1992) also suggests multiplying the hard cost value by two to obtain an approximation of the total retrofit project cost. This multiplier covers architecture and engineering fees, permit fees, legal fees, construction financing and other “soft costs” typically associated with the retrofit work. A range of hard costs for different building types is given in Table 2-2 on a per square foot basis. It should be noted that the costs in Table 2-2 were recommended based on a small database of primarily California earthquake-prone buildings. According to Table 2-2, the proposed costs vary highly across the building types. The range of retrofit costs within a given building type is also considerably large.
Table 2-2: Typical hard retrofit costs (FEMA 227, 1992)

<table>
<thead>
<tr>
<th>Building type</th>
<th>Cost (US$/ft²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>URM – Unreinforced masonry bearing wall</td>
<td>10-25</td>
</tr>
<tr>
<td>RM – Reinforced masonry</td>
<td>8-17</td>
</tr>
<tr>
<td>W – Wood</td>
<td>7-17</td>
</tr>
<tr>
<td>S1 – Steel moment frame</td>
<td>8-20</td>
</tr>
<tr>
<td>S2 – Steel braced frame</td>
<td>8-20</td>
</tr>
<tr>
<td>S4 – Steel frames and shear walls</td>
<td>5-12</td>
</tr>
<tr>
<td>S5 – Steel frames and URM infill</td>
<td>5-12</td>
</tr>
<tr>
<td>C1 – Cast in place reinforced concrete frame</td>
<td>10-12</td>
</tr>
<tr>
<td>C2 – Cast in place reinforced concrete shear walls</td>
<td>8-30</td>
</tr>
<tr>
<td>C3 – Cast in place reinforced concrete frame with URM infill</td>
<td>20-25</td>
</tr>
<tr>
<td>PC1 – Precast concrete tilt-up</td>
<td>3-12</td>
</tr>
<tr>
<td>PC2 – Precast concrete frame</td>
<td>8-30</td>
</tr>
</tbody>
</table>

In another retrofit benefit-cost study, Hopkins and Stuart (2003) used the following formula to assess the structural cost of retrofitting on a per square metre basis. The formula includes a significant fixed component to reflect the costs of opening up and making good after the retrofit structural work, regardless of the amount of such work to be done.

\[
C = (0.08 + \sqrt{DR_{ex} - DR_{ret}}/3)) \times NCC
\]  

(2.5)

where,

DR_{ex}: Damage ratio of an existing building

DR_{ret}: Damage ratio of the retrofitted building

NCC: New construction cost (per square meter)

Damage ratio is a measure of the expected percentage of building damage as a function of the intensity of earthquakes (FEMA 227, 1992). This intensity is normally expressed in terms of the Modified Mercalli Intensity (MMI) scale. The values of the damage ratios in Equation (2.5) were derived from a confidential report which modified US data to suit New Zealand building types. One major difficulty with this model is that it depends on the parameters which are unknown before a retrofit project commences. These parameters
can be determined only after the full completion of the seismic assessment and retrofit design of the project. As a result, this model may not be used in the early retrofit planning phase when the detailed information of the retrofit projects is not available. This is a serious limitation to the practical application of this model, particularly if it is to be used by officials having limited engineering background or experience.

Yüzügüllü et al., (2004) studied the impact of the Kocaeli earthquake (1999) on school buildings and described the subsequent retrofit and reconstruction activities. In this earthquake, 820 schools were affected in Istanbul, 22 of which were consequently demolished. This study also assessed the seismic vulnerability of existing school buildings in Istanbul. Following this assessment, an estimated budget was provided for seismically retrofitting those schools that pre-dated the most recent 1998 Turkish building code. At the time when this study was conducted, any repair or retrofit work should follow the requirements of this code, which is normally used for new buildings. During the initial phase of this study, the existing data on the school building stock in Istanbul was improved by collecting additional information on the year of construction, number of floors, total construction area, availability of design projects, and the geologic and soil condition of the region. The formula used to calculate seismic retrofit design cost and seismic retrofit construction cost was simply based on multiplying the total building area by a fixed amount of $US4/m² and $US80/m², respectively. These amounts were based on 2003 US dollar value. The results were then summed up to obtain the total retrofit project cost. In this study, the retrofit objective was to achieve an “Immediate Occupancy” performance level under a 25-year return period earthquake. Schools at this performance level were expected to experience minimal post-earthquake structural damage, with some non-structural damage that could be repaired while the schools are in operation.
2.7 Limitations of the Reviewed RCE Models

The reviewed retrofit cost estimating models are limited in a number of ways as described below:

First, most of the models were proposed on the basis of the data that were originally collected in the US (i.e., FEMA 156, 1994; FEMA 227, 1992; Potangaroa, 1985). The utilization of these models is therefore only warranted for US seismic retrofit projects. The accuracy of these models for being generalized to other countries is uncertain, due to different retrofit practices in different countries and presumably different material and labour costs.

Second, the retrofit cost models proposed in the previously reported studies are error-prone, considering the way in which the models were derived. The derivation of the models was commonly based on two methods: subjective judgement (i.e., FEMA 227, 1992; Hopkins and Stuart, 2003; Yüzyü güllü et al., 2004) and regression analysis (i.e., FEMA 156, 1994; Potangaroa, 1985). As described earlier, the reliability and accuracy of the proposed models is essentially limited either by the uncertainties and human errors implicit in the nature of the subjective judgement or by certain assumptions implicit in the regression analysis.

Third, the sample size in some studies (e.g., FEMA 227, 1992; Potangaroa, 1985) was insufficient to produce a significant model for predicting retrofit cost. In addition, FEMA 156 (1994) generated its cost super database on the basis of a limited number of real historical data. The severity of this limitation is more pronounced for certain building types. For instance, the number of historical data collected in the case of building type S3 (i.e., steel frame with infill walls) was 116. This number for the group consisting of building types S2 (i.e., steel braced frame) and S3 (i.e., steel light frame) was even smaller.

Fourth, previous studies have devoted less effort to properly exploring the extent of the influence of retrofit cost factors on the retrofit cost. In addition, the statistical characteristics of the available retrofit cost models were not reported in sufficient detail. For instance, there is no evidence of the degree of accuracy of the models to explain and predict outcomes. This lack of evidence also applies to the statistical significance of the retrofit cost factors. Similarly, no discussion was provided by the authors of these previous studies as to whether or not the standard regression assumptions are violated in the proposed models by problems such as multicollinearity, outliers, and heteroscedasticity.
Fifth, the lack of a comprehensive set of retrofit cost factors in the available models is also serious. Although these models provided a simplified cost estimating method for retrofit projects, the complicated nature of this class of projects may necessitate the use of additional factors to increase confidence in the results. In addition, the effectiveness of the available models for a varying number of retrofit cost factors was not examined.

Sixth, the prediction ability of the available retrofit cost models was not validated on datasets other than those from which the models were originally derived. This criticism raises a serious question about the reliability and stability of the models.

Seventh, in the studies reviewed no mention was made of the application of ANNs. In general, no attempts have been made so far in the literature to investigate the use of advanced non-parametric methods, including the ANN technique, for modelling retrofit cost.

In an effort to address the limitations of these earlier studies, the current research aims to effectively resolve these limitations by developing reliable retrofit cost estimating (RCE) models based on a variety of combinations of significant factors whose values are known or can be readily determined before a retrofit project commences. These models can be used by decision makers, assisting them to establish reasonable retrofit cost estimates in the early retrofit planning stages.
2.8 Cross-Validation Technique

One of the most crucial drawbacks of the available retrofit cost models, as explained in the previous section, is that the prediction ability of these models was not validated on datasets other than those from which the models were originally derived. This criticism raises a serious question about the generalisation ability of the models. Many statisticians (e.g., Berk, 1984; Ronchetti et al., 1997; Shao, 1993; Snee, 1977) claim that the procedure in which the calculation of the prediction errors is solely based on all available data in the database does not provide an acceptable indication of the generalisation ability of a given model. The main reason for this claim is that prediction errors, when calculated in this manner, are not independent of the data used to fit the model. According to Snee (1977), there are four procedures available for the validation of regression models, being (1) a comparison of model predictions and coefficients with theory; (2) a comparison of results with theoretical model calculations; (3) collection of new data to check model predictions; and (4) the use of the cross validation (CV) technique or data splitting. Of these procedures, the last two provide independent prediction errors and are, therefore, preferred by most practitioners (Kozak and Kozak, 2003). Also, given that collecting new data for model validation is frequently neither practical nor feasible, the CV technique is regarded as an acceptable alternative (Kozak and Kozak, 2003). This technique is considered an effective method of assessing model stability (Palmer and O'Connell, 2009), selecting the best model (Cheung and Skitmore, 2006; Yang, 2007; Zhang, 1997), and obtaining nearly unbiased estimators of prediction error (Shao, 1993; Snee, 1977). Another advantage of the CV technique is that it can be applied to more complicated models, such as artificial neural network (ANN) models (see Chapter 6).

Several methods can be employed for the CV technique, including the one-fold, k-fold, and leave-one-out method. In the one-fold CV method, available data in the entire database is divided into two datasets. The first dataset becomes the development dataset which is used to estimate coefficients of a model, while the second dataset becomes the hold-out dataset which is used to measure the prediction accuracy of the model. In comparing regression modelling to ANN modelling, the use of a development dataset is comparable to using a training dataset, while a hold-out dataset is comparable to a test dataset. Various recommendations are made in the literature on what the size of the test dataset relative to the development dataset should be (e.g., Alex et al., 2010; Hammerstrom, 1993; Williams et al., 2005). In the literature related to construction cost modelling, test datasets have been populated with various portions of data in the entire database (e.g., 10% (Chen and Huang,
2006); 15% (Wilmot and Mei, 2005); 20% (Shehab et al., 2010); 25% (Abu Hammad et al., 2010); 30% (Cheung et al., 2008); 50% (Adeli and Wu, 1998)). Considering these percentages, it becomes apparent that 10% to 50% of the available data are usually accommodated in the test dataset. The test data are also most frequently set aside randomly (Adeli and Wu, 1998; Kozak and Kozak, 2003; Palmer and O'Connell, 2009), although formal algorithms are available to select the test dataset in such a way that this dataset matches, as closely as possible, the distribution of the development dataset (Snee, 1977). The K-fold CV method is generally regarded as an improved version of the one-fold CV method (Kozak and Kozak, 2003). In the k-fold CV method the total number of data \((n)\) is divided into \(k\) equal subsets. This division leads to a procedure in which \((k - 1)/k\) portion of the data is used to estimate coefficients of a model, while \(1/k\) portion of the data is used to measure the prediction accuracy of the model. This procedure is repeated \(k\) times so that each subset is used for model fitting as well as for model validation. An extreme case occurs when \(k = n\), so that each iteration of the model is derived from \(n - 1\) data, and the test dataset contains only one data. This special case of the k-fold CV method results in what is known as the leave-one-out CV method (Cheung and Skitmore, 2006).

Amongst the methods described above, the one-fold CV method has attracted the greatest attention of researchers in the field of construction cost estimation (e.g., Abu Hammad et al., 2010; Adeli and Wu, 1998; Alex et al., 2010; Attalla and Hegazy, 2003; Boussabaine, 1996; Chen and Huang, 2006; Cheung et al., 2008; Elhag and Boussabaine, 1998; Elhag and Boussabaine, 1999; Hegazy and Ayed, 1998; Kim et al., 2004; Kim et al., 2005; Shehab et al., 2010; Williams, 1994; Williams, 2002; Williams, 2003; Williams et al., 2005; Wilmot and Cheng, 2003; Wilmot and Mei, 2005). The popularity of the one-fold CV method can be explained, at least, by two reasons:

First, in comparison with the one-fold CV method, the implementation of the other two methods (i.e., k-fold and leave-one-out) entails much more effort and calculation time. This is particularly true in the case of ANN modelling because, as will be shown later in Chapters 6 and 7, the process of developing ANN models requires an extensive refining of network parameters (Attalla and Hegazy, 2003; Bode, 1998; Boussabaine, 1996; Creese and Li, 1995; De la Garza and Rouhana, 1995; Hegazy et al., 1994; Jain et al., 1996; Kim et al., 2004; Li, 1995; Liu, 2006; Wang et al., 2000; Yeh, 1998). As noted by Smith and Mason (1997), these requirements represent formidable impediments to widespread use and acceptance of ANN-based cost estimation relationships.
Second, and more importantly, it can be inferred from the literature that the one-fold CV method is expected to produce stable results, as these results are unlikely to be significantly affected by a change in the number of data points in the development (training) or hold-out (test) datasets. For instance, Shtub and Zimerman (1993) evaluated the effect of the training dataset on the overall generalization performance of the ANN models developed for estimating the cost of six assembly systems. Two training datasets with 756 and 2268 cases were used for this purpose. Shtub and Zimerman (1993) found that only in one assembly case, the generalization improved when using the larger training dataset. In all the other 5 cases, the generalization with the larger training dataset was not significantly better than the generalization with the smaller training dataset. Bode (2000) observed that although increasing the training dataset size from 10 to 50 improved the predictive performance of his ANN models, further increase in training data up to 200 had almost a negligible effect on this performance. Based on this observation, Bode (2000) concluded that the marginal contribution of additional training data considerably diminishes with growing training dataset sizes (Bode, 2000). In a more recent study, Stockton and Wang (2004) noted that the estimating accuracy of ANN models is more dominated by the number of variables rather than the number of training samples. They then concluded that the greater the number of variables used, the less will be the effect of increasing the number of training samples on estimating accuracy of ANN models. The studies reported so far provide some evidence to support the inference stated at the outset of the current paragraph for ANN models. This inference is also valid for MLR models. For instance, Kozak and Kozak (2003) found a minute difference between the prediction errors obtained during the cross validation process and errors obtained using all data points. According to this finding, Kozak and Kozak (2003) recommended that the model fit statistics (e.g., sum of squared errors, mean square error, coefficient of determination, average absolute bias) be used in the process of model evaluation, and that using different combinations for development and test datasets may be unnecessary. In addition to Kozak and Kozak’s (2003) study, the results of a later study by Stockton and Wang (2004) demonstrated that the estimating accuracy of the regression-based models improves with an increase in the numbers of variables, but there appears to be no marked improvement when the number of data used to develop models increases.

As a concluding remark, it is worth emphasizing that one of the drawbacks of the cross validation technique is that variances of regression coefficients generated from the development data will be larger than those obtained from a fit to the entire data. However, if
the database size is large, the variances of the regression coefficients will be small even if only half of the entire data points are used to estimate these coefficients (Snee, 1977).
2.9 Conclusion

This chapter was devoted to the presentation of results from an extensive literature review on the subject of cost modelling and its related issues. Cost modelling was defined in the literature as an attempt to predict the likely cost of a future project before the design details of the project become available. Following this definition, cost modelling techniques were classified into two broad categories: parametric and non-parametric. A broad range of applications of the regression analysis technique and the artificial neural network (ANN) technique were presented in this chapter for the development of parametric and non-parametric cost estimating models, respectively. The few available studies found in the literature on the subject of retrofit cost modelling were also reviewed, and their associated limitations were accordingly identified and discussed in detail. This review demonstrated that subjective judgement and regression analysis were the two main approaches utilized in the prior studies to develop retrofit cost estimating (RCE) models. The uncertainties and assumptions inherent in these approaches place a major limit on the predictive capability of these models. No attempts have been found in the literature to make use of the ANN technique to develop more advanced RCE models. Taking these and other limitations discussed in this chapter into account, the next chapter presents the methodology used in this research to effectively resolve these limitations, as well as other related problems associated with seismic retrofit cost modelling.
CHAPTER 3

METHODOLOGY

3.1 Introduction

The basic tasks involved in the process of developing cost models are data identification, data collection and data analysis (Wang et al., 2000). The basic aims of these tasks are to identify and collect cost-related information and to analyse this information in order to quantify the cost relationships that exist. In the cost modelling process, therefore, decisions need to be made on which variables are relevant to be collected, how this data collection process should take place and finally which data analysis techniques need to be employed to establish valid cost estimation models.

This chapter describes the methodology deployed in this study to address the questions above. In this chapter, retrofit cost categories are first described and then the factors that are thought to have an appreciable influence on the value of the retrofit cost are identified through an extensive literature review. The chapter continues with a description of the strategies pursued to develop the research cost database. The data collection process adopted in both New Zealand and Iran is described. The chapter concludes with the introduction of appropriate analytical techniques utilized in this research for establishing retrofit cost estimating (RCE) models.

3.2 Retrofit Cost Components and Definitions

Undertaking retrofit, as for other projects, is far from being a single activity; rather, it is a feat of multitasking accomplishments, each of which is essential in order to achieve successful execution of a retrofit project. Such projects would normally go through a procedure that is a net of activities as shown in Figure 3-1. Each activity in this figure has its own cost that is needed to be identified when appraising total retrofit project cost. This total cost is the expected total amount that may be paid to engineering consultants, project management firms, contractors, city councils, and all the parties who are involved in handling the retrofit procedure from beginning (visual rapid screening of the building) to end (retrofit construction work).
Figure 3-1: Retrofit procedure

Step One:
Visual and rapid screening of the structure
(Identify and prioritize seismic prone buildings)

Step Two:
Detailed qualitative and quantitative seismic assessment

Step Three:
Design retrofit measures to enhance building seismic strength

Design different retrofit solutions (at least two)
Select the best retrofit option by taking economic, social, technical, and other constraints into account
Prepare construction and tender documents

Step Four:
Retrofit construction works

End
The total retrofit project cost can be divided into two sub-categories, being “direct costs” and “indirect costs” (FEMA 156, 1994):

i. **Direct cost**

Direct cost is the cost that is directly attributable to all the various activities that are required to be in place in order to obtain the predefined aims of retrofitting, and includes the cost of all steps necessary to be thought in advance in order to take most advantage of a retrofit program (see Figure 3-1). Direct cost is itself composed of “construction cost” and “non-construction cost”.

a- **Construction cost**

Construction cost is the amount that should be paid to contractor(s) and includes contractor’s overhead, profit, and contingency cost. Construction cost also includes change orders during construction work. Retrofit construction work generally starts with demolition and opening up works and then continues with implementation of the seismic retrofit measures, and finally ends with finishing and clean-up work. The construction cost is further divided into two parts: “seismic cost” and “non-seismic cost”.

- **Seismic cost**: This cost is incurred by those construction activities that directly enhance the seismic performance of the structure against earthquake. Such activities may entail both structural and non-structural retrofit works, leading to “structural cost” and “non-structural cost”, respectively.

  The structural cost, as inferred by its name, arises from retrofitting elements which play a crucial role in enabling a building to perform satisfactorily when subjected to earthquake loading. Such elements include structural lateral resisting elements (e.g., beams, columns, joints, and walls), diaphragms and foundations. The structural cost also includes the costs of repairing any elements that have been previously damaged or deteriorated.

  The non-structural cost addresses the costs mostly incurred by seismic mitigation of non-structural components that interact with the structural
system, and also the cost of mitigating other hazards that may be posed by an earthquake, including exterior falling hazards (e.g., parapets, cornices, canopies, exterior non-bearing infill walls and chimneys) and geological hazards. The non-structural cost also includes the costs of necessary works to obtain access to the structural elements and return the buildings to its original condition. This cost, known also as “demolition and restoration cost” or “cover-up cost”, is mainly associated with removing and replacing various finishes that are affected by construction works such as, amongst others, painting, tiling, roofing, and removal and reinstallation of non-structural elements.

- **Non-seismic cost**: Unlike seismic cost, this cost is associated with the construction works that do not directly improve the seismic performance of the structure. However, when a building undergoes seismic retrofit, these construction works may be further accomplished to comply with the local code requirements. Non-seismic construction works can also be considered by owners to improve such aspects as building functionality and appearance. In addition to including the cost of changing and remodelling non-structural components, the non-seismic cost may incorporate the cost of providing fire and life safety facilities, disabled access facilities and hazardous material removal. These costs can be difficult to quantify because they can vary greatly depending upon the individual building characteristics and the requirements set forth in the in-place building acts or seismic codes. Using engineering judgment is therefore a reasonable way to estimate these costs.

b- **Non-construction cost**

The amount paid to anyone other than the contractor in order to complete the project is called “non-construction cost”. This cost covers all activities necessary to be taken before retrofit construction work, as well as all works that increase the reliability of retrofit projects. The amount paid to structural and civil engineers, architects, geology engineers, testing laboratories, and city councils can be therefore placed in this category. A reasonable estimate of the non-construction cost is proposed by FEMA 156 (1994) to be taken as 30% of
the construction cost, while Kappos and Dimitrakopoulos (2008) reported that this percentage could be reduced to about 20%. The latter percentage was largely supported by the California State Building Seismic Program, which assigned one-sixth of the total project cost to the non-construction cost. This cost may include project management fees, architectural, geotechnical, and structural design fees, permits and approvals fees, in addition to material testing fees. The material testing cost is somewhat unique to retrofit projects because in such projects there is almost a need to hire a testing agency to assess the strength of the existing materials during the retrofit design phase. This testing need is more pronounced in cases where no records of as-built documents can be found to support building necessary information. This information is highly required when assessing the seismic performance of the building and also when designing retrofit measures.

**ii. Indirect cost**

Indirect cost comes about as a result of the retrofit work and affects the owner, the tenants, and the community on the whole. It is often the indirect cost that deters building owners from retrofitting their buildings for improved earthquake performance. A retrofit feasibility study conducted in Istanbul (Sucuoglu et al., 2006) affirmed the great importance of relocation cost in making the final decision on which retrofit solution is more economical to adopt. This study showed that residents clearly do not want to leave their buildings during retrofit construction works, and the cost of alternative accommodation and disruption to their business were commonly cited as other preventive issues. The results of another study in Canada also revealed that the total retrofit project cost could be highly affected by the relocation cost, as this cost may form 20% to 30% of the total project cost and also be in the range of 25% to 45% of the construction cost (Cheung et al., 2000). In a more recent study, Kappos and Dimitrakopoulos (2008) suggested that the retrofit indirect cost is suitably taken as 12% of the building replacement value (BRV). This percentage implies that the indirect cost equates to the amount spent for retrofit construction works. Kappos and Dimitrakopoulos (2008) also provided the breakdown of different elements contributing to retrofit indirect cost. They proposed to approximate the rental loss per month, relocation expenses per month, and business interruption cost per day as 0.5%, 2.0%, and 0.3% of BRV,
respectively. Despite these rough estimations, Comerio (1989) was of the idea that the indirect costs as a result of a retrofit project are difficult to measure. Retrofit indirect costs may also vary considerably from project to project, making it more impractical to involve such costs in retrofit cost estimating (RCE) models.

Using the cost distinctions given above, the taxonomy of costs developed in this research is shown in Figure 3-2.
Figure 3-2: Retrofit cost framework
3.3 Effective Factors Contributing to Retrofit Cost

Retrofit costs can be rather uncertain in nature (Grossi, 1998). The results of a survey conducted by Grossi (1998) showed a wide dispersion on what the costs of mitigation might be. Hopkins and Stuart (2003) emphasized that the cost of retrofit could vary significantly and reported that consultants’ records show a wide range of costs and that very little correlation, if any, could be found between. To diminish the uncertainty associated with retrofit costing, the following question needs to be properly answered:

*Which are the important variables deserving the most attention in forming a retrofit cost model?*

To develop reliable retrofit cost estimating (RCE) models, it is therefore necessary to identify and categorise those factors that might significantly contribute to such models. These factors are determined in this study and placed into four main categories as follows:

i. **Building characteristics based variables**

a- **Year of original design and construction**

In recent earthquakes, casualties and damage associated with older buildings, which were designed and constructed using codes that are known to provide inadequate safety, are far worse than that for newer buildings which have been designed and built in accordance with more stringent code requirements. Prior to the early 1970’s much less was known about seismic design, so that much of the construction in the first two-thirds of the last century was designed without adequate detailing and reinforcement for seismic protection and may be deficient according to the seismic requirements of current design standard. Dowrick and Rhoades (1997) studied the effectiveness of past changes in New Zealand earthquake design practices. The study concluded that the introduction of improved ductility for concrete in c. 1970 and for other materials in c. 1980 significantly enhanced the seismic behaviour of structures and has also resulted in a meaningful reduction of heavy damages, while introduction of the first earthquake design code in 1935 and the code changes in 1965 were so modest that they resulted in no considerable difference to the
vulnerability of most building classes, especially the low-rise ones. Seismic mitigation of newer buildings is therefore likely to be simple and inexpensive, while retrofit of older buildings is usually troublesome and costly.

b- Building characteristics (e.g., vertical and lateral resisting system, foundation and diaphragm type, height, area, etc.)

Building characteristics have a dramatic effect on the retrofit cost (FEMA 276, 1999). Recommendations are generally made to begin the retrofit program just for certain building types posing the most risk. It is well documented that some certain building types, such as unreinforced masonry bearing wall buildings and older improperly detailed reinforced concrete frame buildings, are more vulnerable when subjected to ground motions than other types and have historically presented a high risk, depending on local seismicity and building practice (FEMA 356, 2000). A review of recent worldwide earthquakes showed that most of the damage occurred in unreinforced masonry buildings (URM), reinforced concrete buildings with no structurally significant masonry or concrete infill panels (RC), combinations of URM and RC buildings, and finally, adobe buildings (Charleson and Fyfe, 2001; Quadri, 1997; Smyth and et al., 2004). By reviewing the changing nature of New Zealand construction practice after the 1931 Hawke’s Bay earthquake, Ingham (2008) claimed that unreinforced masonry structures still represent the most earthquake-prone class of buildings in New Zealand, despite their use as public buildings being prohibited in 1965, and from 1931 onwards the majority of construction type had changed. Given the same noticeable effect as the building material, poor design of details, poor supervision and construction, and structural deficiencies (i.e., flawed design concepts), such as plan and vertical irregularity, are also of great significance as each can be the main reason for building damage and can form a large portion of damage cost (Beetham et al., 2001; Charleson and Fyfe, 2001; Downes et al., 2001; Dowrick et al., 1995; Dowrick, 1998; King, 1995; Park, 2001; Shephard et al., 1997). In more recent earthquakes such as the 1994 Northridge earthquake, even buildings constructed under recent codes suffered unexpected damage as a direct result of such deficiencies (Quadri, 1997).
c- Non-structural equipment and components

Non-structural components are typically defined as all parts of the building aside from the structural system, including infill walls, glazing, cladding, partitions, finishes, and mechanical, electrical, and plumbing (MEP) systems. Although structural strengthening is most often considered the only path to reduce seismic vulnerability, unsecured non-structural components can be significant life-safety hazards. These components are also the most common cause of secondary damages including fire, flooding, contamination and the loss of life-sustaining systems which in most cases, like structural damage, may result in building closure, loss of functionality and services to the public and loss of revenue to users and owners (Lewis and Wang, 2004; Nichols and Beavers, 2003; Okada, 1992). Kappos and Dimitrakopoulos (2008) reported that damage to contents could be as high as 30% of the building replacement value. In the 1994 Northridge earthquake the vast majority of damage was to non-structural elements, in that, the damage cost to non-structural components was in excess of 35% of the pre-earthquake value of the buildings (Grossi, 1998). Likewise, in the 1931 Hawke’s Bay earthquake and the 1987 Edgecumbe earthquake, the costs due to non-structural damage constituted a considerable portion of the total damage costs (Dowrick and Rhoades, 1995; Dowrick et al., 1995; Dowrick, 1998). It has also been reported that for a relatively small to moderate level of shaking, the total loss is dominated by non-structural and contents damage and little loss is incurred due to structural damage (Beetham et al., 2001; Bradley et al., 2008; Dowrick and Rhoades, 1997). In addition, for a code-complying structural system, damage due to both moderate and large ground motion shaking is reported to be mostly caused by damage to non-structural components and contents and is not due to structural damage or global collapse (Bradley et al., 2008). The observations above indicate that the improvement of seismic performance of a building requires a focus on improving the performance of the entire structure and not just the structural system. The serious need of non-structural mitigation integration with structural retrofit is also heightened in a number of disseminated technical reports such as those issued in 1997 and 1999 in the British Columbia Province of Canada under the names of “earthquake preparedness” and “earthquake preparedness, performance audit”, respectively
(Lewis and Wang, 2004). Given the importance of this issue, many relatively simple and low cost mitigation measures have been developed so far for non-structural elements. The cost of retrofitting such elements has recently decreased to just a small fraction of the cost of retrofitting structural components (California Seismic Safety Commission, 1999; Lewis and Wang, 2004; Nichols and Beavers, 2003).

### ii. Site Based Variables

#### a- Seismicity

Earthquake potential losses monotonically increase for higher seismic intensity (Kappos and Dimitrakopoulos, 2008). The results of cost-benefit studies usually underline the need to deal with existing buildings in those towns and cities that are subject to moderate to high seismicity rather than those located in regions of low seismic activity (Hopkins and Stuart, 2003). In general, the higher the seismicity, the greater the cost of retrofit, as the building will face larger seismic forces and therefore higher level of strengthening will be needed to achieve the same performance.

#### b- Local Site Condition

Ground condition is a major contributor to project cost (Khosrowshahi and Kaka, 1996). Alex et al., (2010) made it clear that the predictive ability of their cost models addressing the installation of water and sewer services would be further enhanced in the presence of geotechnical data. The damage caused by a number of previous earthquakes appeared to be highly influenced by hazards induced by local conditions of the site. For example, in the 1931 Hawke’s Bay earthquake several examples of failure were reported to be caused by unsuitability of the site condition, and the damage to buildings was worst on sites with significant ground damage (Dowrick et al., 1995). Soft soils are another source of hazard because such soils amplify the ground shaking intensity, which may lead to whole or partial collapse of buildings (Beetham et al., 2001). For example, the recent earthquakes of Northridge (1994), Kobe (1995), Armenia (1999), Colombia (1999) and Turkey (1999) have shown that damage is often larger over soft soils than on firm rock. It is therefore essential
to accurately determine which soil type should be applied to any given site, which is a task usually done by geotechnical engineering consultants.

iii. Social based variables

a- Retrofit objective

One of the fundamental tasks in the retrofit effort, which needs to be agreed at the preliminary stages of work, is selecting the appropriate seismic performance objectives for the buildings under consideration depending on the needs and available resources. Higher seismic performances are generally targeted for buildings when their functionality is vital after earthquake and they are required to remain operational following earthquake attacks. Such buildings may include hospitals, fire stations, and emergency management services. Retrofit objective plays a major role in decision-making analysis as the extra cost incurred by an overestimated retrofit objective may significantly change the economic viability of retrofit, resulting in reducing or eliminating a retrofit plan. For example, the results of a study conducted by Holmes et al., (2008) revealed that in the view of building owners, only 5% to 10% increases in the cost is worth to obtain superior objective. On the contrary, underestimating the retrofit objective may also lead to significant post-earthquake repair cost, closure of building, and more importantly, life losses.

b- Building use (building social classification) and importance level

The general stock of buildings can be classified by occupancy. Occupancy determines the building use and the number of occupants (FEMA 154, 2002) and governs the importance level of a building. For instance, AS/NZS 1170.0 (2002) divides buildings based on their importance level into five categories (from low to exceptional). The importance level is important in the sense that it is amongst the factors that the seismic codes usually consider in determining seismic design load (i.e., base shear). According to the Iranian code of practice for seismic resistant design (standard 2800, 3rd revision (BHRC, 2005)), the seismic design load of a very high importance building is 1.75 times greater than that of a building with low importance. The results of damage costs and damage ratios for different occupancy classes and their contents in the 1931 Hawke’s Bay earthquake, 1942 Wairarapa earthquake, 1968 Inangahua
earthquake, and 1987 Edgecumbe earthquake made it clear that building seismic vulnerabilities could be considered as a function of occupancy and social classifications (Downes et al., 2001; Dowrick et al., 1995; Dowrick and Rhoades, 1997; Dowrick et al., 2001). Wells (1996) asserted that occupancy classification has a far greater influence on placing retrofit policy, retrofit time-frame, and retrofit cost, for interim securing and strengthening than does the structural rating. This claim is also supported by the California state legislature as they stated that buildings representing a structural life safety hazard cannot be used for hospital functions after the year 2008. These buildings must be demolished, used for other purposes, or considered for retrofitting. Williams et al., (2009) also suggested that the retrofit of buildings located in regions of low seismic activity and not designed to resist seismic loads, may not be financially viable unless their indirect value (i.e., higher-importance use, expensive contents, human lives, etc.) is significantly greater than the direct structural value. This suggestion may be the case just for limited facilities such as fire stations and hospitals. Non-structural component damage cost is another issue that is highly sensitive to the building use (Dowrick and Rhoades, 1995).

c- Heritage consideration

Historic buildings are typically placed on a city, or national register of historically significant buildings, and all construction and retrofit work on them is subject to review by a regulatory agency. Such buildings typically have features in them such as historic facades, architectural columns or special ceilings that cannot be removed or altered during the seismic retrofit work. The retrofit cost of such buildings is therefore primarily associated with the architectural work involved with the seismic retrofit of the main structure. FEMA 156 (1994) estimates that retrofit total project cost for a heritage building can be considerably higher than that of an ordinary building. The cost can be in the order of three times as much as ordinary buildings. FEMA 156 (1994) also suggests that costs associated with retrofitting heritage buildings should be defined on a case-by-case basis with special attention being paid to the unique nature of each project. Smith (1979) set a cost criterion for strengthening a heritage building, in that, when a building has an historic
appeal and warrants preservation, a greater cost of up to 10% more than the cost of a new building of the same floor area might be justified.

iv. Retrofit policy and solution based variables

a- Seismic guideline and load level considered for retrofit

The following questions naturally arise in the early stages of retrofit mitigation planning to engineers and policy makers which need to be carefully addressed:

“What standard should be used when dealing with earthquake-prone buildings?”

“What level of lateral loading should be selected to be applied to such buildings to protect the inhabitants from seismic damage within accepted economic bounds, yet be technically viable?”

A retrofitting plan that seeks to comply with full new building standard is the most effective mitigation measure, and retrofit efficiency decreases by adopting a lower value of compliance. However, the high cost of seismic retrofit, if done to comply with the latest in-place code requirements developed for the design of new buildings, may be too high and prohibitive for economic feasibility. Retrofit cost generally increases with a trigger level of legislation regime which regulates existing buildings to comply with a prescribed percentage (e.g., 33%, 67% or 100%) of new building standard. There may be budget constraints that restrict the adoption of a certain legislation regime. Smith (1979) argued that strengthening of earthquake-prone buildings to current requirements is often uneconomic. This argument is also supported by a recent study conducted by Kappos and Dimitrakopoulos (2008). The results of their life-cycle cost analysis showed that the optimal retrofit level is far lower than that corresponding to performance levels adopted by current codes. The effect of different building codes in the U.S. on building construction costs was investigated in a study conducted by Williams et al., (2009). The findings from this study suggest that the adoption of some certain codes would increase the construction cost by 10% or more depending on building type. This suggestion is however not compatible with the results
of a case study performed by Hare (1996) on a heritage building in New Zealand. This study came to the conclusion that the retrofit cost and seismic load level are almost independent of each other as different seismic loads have a minuscule effect on the total retrofit cost. This study showed that the retrofit cost when using a load level of full code loading to NZS 4203 (1992) was only 1.3% greater than that when using a load level of 2/3 of full code loading.

b- Retrofit Mitigation measures

The retrofit cost is highly dependent on the type of strengthening method (Kappos and Dimitrakopoulos, 2008). Changes in construction technologies and innovation in retrofit technologies present an added challenge to engineers in selecting a technically, economically and socially acceptable solution. Conventional upgrading techniques usually include the addition of existing walls and foundations and strengthening of frames. Most of these techniques often lead to costly consequences such as heavy demolition, lengthy construction time, reconstruction, and occupant relocation, with all the associated direct and indirect costs. During the past few years, successful and cost effective applications of relatively modern technologies such as damping devices, base isolation, pre-stressing techniques, steel shear plates and advanced composed materials have been reported. The use of these innovative technologies in building retrofit projects was found to be far less intrusive to building occupants and offered savings in retrofit cost. These solutions have been employed more during recent years and seem to have become the norm, especially in the case of critical and important structures (Jury, 1997). Cheung et al., (2000) studied the advantages of utilizing a number of such innovative technologies in the retrofitting of several federal buildings in the Province of British Columbia (Canada), which is the most seismically active zone in the country. They concluded that potentially large savings can be realized by reduction in construction cost and by avoidance of tenant relocation and associated productivity losses. The results of this study showed that by using the new technologies, reduction in the total project cost (including construction cost and indirect cost) was in the range of 40% to 60%. However, there are some instances that prove that employing innovative retrofit techniques would have cost substantially more than a more conventional
retrofit approach. For example, a possible solution for Christchurch Government building and Auckland Town Hall by base isolation was ruled out due to the incurred cost (Hare, 1996; Robertson, 1996). Another limitation in selecting these solutions arises from their efficiency. For example, studies have demonstrated that base isolation is unlikely to be efficient for inflexible structures with a long natural period of vibration, or in deep flexible sites as its use may detrimentally affect the response of the structure (Park, 2001).

c- Occupancy condition during retrofit construction work

The retrofit solution will indicate whether or not the building needs to be vacated during construction work. If for any reason the building functionality needs to be maintained during construction work, then measures may have to be taken to minimize the disruption and, as a result, the cost of the retrofit may considerately increase. Retrofit projects very often involve several moves and relocation of occupants. These relocations may also include the move of workstations or services such as telephones, computer networks, fire alarms and security systems. These activities may cause some delays in retrofit construction works in addition to increasing the associated retrofit cost.

The summary of factors found in the literature to influence and control the retrofit cost is illustrated in Figure 3-3. These factors are grouped into four main categories, as described above.
Figure 3-3: Potential factors which may have influence on retrofit project cost

- Area
  - Total number of stories
  - Age
  - Weight
  - Configuration (regularity in plan and height)
  - Structural type (vertical and lateral resisting system)
  - Diaphragm type
  - Foundation type
  - Non-structural equipment and component
- Social characteristics
  - Occupancy classification
  - Importance level
  - Retrofit objective
  - Heritage consideration
- Building characteristics
- Retrofit policy characteristics
  - Compliancy with first in-place seismic code
  - Seismic guideline and load level used for retrofitting
  - Retrofit mitigation measure
  - Occupancy condition during construction work
- Site characteristics
  - Seismicity
  - Soil type

Retrofit project cost
- Construction cost
- Non-construction cost
- Indirect cost

Figure 3-3: Potential factors which may have influence on retrofit project cost
3.4 Data Collection

In previous sections, retrofit cost framework was developed and the relevant variables that can potentially influence such cost were identified. The next step towards RCE modelling is the data collection, which is the procedure that is pivotal in every research exercise (Lancaster, 2005). The importance of this issue arises from the fact that the use of any mathematical modelling technique is primarily based on the data collected from the previous retrofit studies. If no data are available, then any cost estimation effort is useless. The data gathered are normally compiled in a database. The accuracy of the data comprising the database is a key factor in developing reliable cost estimation models. The robustness of these models is also highly dependent on the amount of data available in the database (Kim et al., 2004). The database should therefore contain a comprehensive amount of reliable data on each variable shown in Figure 3-3.

In this research, two countries, namely New Zealand and Iran, were initially selected to run the data collection program. The procedure and tool employed to gather the required data and the results obtained from the data collection program in each of these countries are discussed in the following sections.

3.4.1 New Zealand Survey

In New Zealand the construction of houses and other buildings is now controlled by the Building Act 2004, which applies to the construction of new buildings and to the alteration and demolition of existing buildings. Territorial Authorities (TAs) are required by the Act to adopt either “passive” or “active” policies on earthquake-prone buildings, where in the former, assessment and retrofit are only required if a building owner applies for an alteration or change of use, while in the latter, evaluation and upgrade are carried out for all buildings.

The New Zealand Society for Earthquake Engineering (NZSEE) guideline for the assessment and improvement of the seismic performance of buildings in earthquakes (2006) include a methodology for assigning priorities and timescale for retrofit. The methodology is based upon the determination of a structural performance score, which is a numerically established index known as percentage of New Building Standard (%NBS). The %NBS is primarily based on the seismicity of the region where the building is located, site soil
condition, compliance of the building with the code in-place at the time of its original design and also building characteristics as shown in Figure 3-3. This baseline value is then adjusted for the most vulnerable structural deficiencies, weaknesses, and geological hazards. The Building Act 2004 requires that buildings with less than 33% NBS should be strengthened to an acceptable level and any building that has greater than 33% NBS does not technically require seismic retrofit intervention. However, NZSEE (2006) states that the starting point of any discussion between engineers and clients should target 100% of new seismic design requirements. NZSEE (2006) also strongly recommends that every effort be made to achieve improvement to at least 67% NBS as this level of strengthening reduces the relative risk from about 20 times to about 3 times that of a new building. This recommendation is made as the most damaging event to the built environment in New Zealand since the substantive beginning of European settlement c. 1840 was due to earthquake.

It was found that 90% of modern New Zealand houses and 99% of old-fashioned buildings would sustain some level of damage when subjected to strong ground motions (Dowrick et al., 1995). Damage costs from New Zealand’s four most destructive earthquakes (excluding the recent 2011 Canterbury earthquakes) amounts to more than NZ$1000 million in 1998 value (Dowrick, 1998). In addition, societal exposure to disaster risk appears to be greater than previously recognized with exponential increases shown over recent years (Shephard, 1997). For instance, according to a New Zealand Treasury estimation, the 2011 Christchurch earthquake could cost up to $NZ 15 billion. Resulting in a death toll of 181, this earthquake was the second largest natural disaster recorded in New Zealand after the 1931 Hawke’s Bay earthquake. Similar situations are also expected to emerge in other regions of the country as the total cost of a large earthquake in Wellington was estimated to be about US$24 billion (Gregory, 1998). It is therefore evident that earthquake preparedness and mitigation measures are essential for all parts of New Zealand.

In an attempt to address the concerns stated above, a national research program entitled “Seismic Retrofit Solutions” commenced in 2004. The primary goal of this program was to develop a national platform of knowledge and expertise associated with seismic retrofit of New Zealand’s earthquake-prone building stock. A further goal was to assist building owners to make more informed decisions on retrofitting their properties, and to address impediments to the implementation of seismic retrofits. Retrofit cost modelling was an integral component of this research program enabling building owners to have a better appreciation of the retrofit decisions they need to make. As a preliminary requirement for
such modelling, the data on retrofit cost and its predictors (see Figure 3-3) needed to be collected in an effective way.

Questionnaires are among one of the most widely used and valuable means of data collection (Lancaster, 2005). In addition, there can be little doubt that the number of surveys being conducted over the World Wide Web is increasing. The ability to collect large amounts of data without interviewers, stationery or postage, and to process answers without separate data entry, makes doing web surveys very attractive (Witt, 1998). In this research, an explicit website was devised as a gateway to completing a semi-structured questionnaire consisting of both open-ended and closed questions that each pertain to either one of the retrofit cost constituents or cost influential factors shown in Figure 3-3. The website (with the address http://www.retrofitcost.net) was developed considering the technical language and approach common to civil and construction engineers and through an extensive iteration with academic staff as well as postgraduate students from the University of Auckland. The development of the website took three months, with another 3 months for testing, auditing, identifying and implementing modification where necessary. During the second three months, the website went through an extensive amount of alteration with an emphasis on increasing its efficacy, practicability and simplicity to ensure a successful outcome.

Many of the findings from researchers (e.g., Dillman et al., 1998a; Dillman et al., 1998b; Smyth et al., 2004) in the field of web surveys and principles recommended for construction of such surveys were implemented in the website. Some examples include effective use of spacing and grouping, providing instructions, allowing respondents to skip certain questions, giving respondents access to their data to either finish uncompleted information or begin with a new survey.

The following are the other key considerations that were taken into account while designing the website:

- Clearly conveying the questions and data requested
- Balancing the need for comprehensive data collection with reasonable survey response length
- Careful utilization of the capabilities of internet technologies and web survey formats such that convenience, understanding, and response accuracy were maximized.
The homepage of the website provides a fast link to start directly with online questionnaires whose structure is shown in Figure 3-4. Two types of questionnaires placed in the website are “fundamental questionnaire” and “supplemental questionnaire”. The former questionnaire consisted of five steps by which it was intended to acquire the essential information for cost modelling. The latter questionnaire, in contrast, is comprised of three steps by which it was aimed to collect the information regarding the factors whose influence on retrofit cost was thought to be of less importance. Although answering the questions of the supplemental questionnaire is completely optional, it was made clear in the website that providing such complementary data would enable the researcher to make the retrofit cost database as thorough as possible. Appendix A presents the requested information embedded in each step of the questionnaires above in conjunction with an associated picture of the web survey.
Figure 3-4: Online questionnaire structure
To ensure a high response rate, the web site was introduced to consultants having been involved in retrofit projects. Each potential consultant was contacted in person and/or via email correspondence to get their commitment to participate in the research. The completion and availability of the website was also nationally announced by means of a paper presented at the 2009 NZSEE conference (Jafarzadeh and Wilkinson, 2009) which is the leading conference in New Zealand held annually to gather researchers and practitioners from different universities, consultants, contractors, and other institutes in New Zealand and all countries that are actively involved in the multi-disciplinary field of earthquake engineering. Despite most of the consultants who were initially approached having expressed their willingness to participate in the online survey, the attempt to obtain required data in New Zealand was largely unsuccessful. Two reasons mainly accounted for this failure. First, although the consultants contacted were informed that the information they would provide would only be used for research purposes and therefore would remain confidential, the retrofit cost and its related data was considered by most consultants to be commercially sensitive information. This sensitivity is by far the main reason that prevented potential contributors from releasing such data to third parties. Second, some of the consultants were of the idea that the effort involved in going through their retrofit archive to extract the information that this study was looking for was beyond their available time and resources and, as a result, they were not prepared to effectively cooperate with this study. These reasons, coupled with the problem that comparatively few retrofit projects seem to have been completed in recent years (as cited by some contributors), greatly hindered this research effort to generate a comprehensive, yet contemporary, retrofit cost database for the NZ earthquake-prone building stock.

3.4.2 Iran Survey

After the 2003 Bam earthquake, which claimed tens of thousands of lives and left many others injured and homeless, the tempo of improving the seismic safety of buildings in both the private and public sectors has accelerated in Iran. One of the most important undertakings of the Iranian government in reducing the seismic vulnerability of the country against earthquake is the program of studying and performing retrofit solutions for important buildings and lifelines. A number of structural groups were addressed by this program including hospitals and emergency buildings, strategic buildings, educational buildings,
bridges, oil and gas infrastructure, communicational buildings, and important bases and lifelines. Because of the wide varieties of buildings, the retrofit costing study conducted in Iran focuses only on public schools as they have received the most attention by far of the national retrofitting program. For instance, the Iranian Parliament passed the “School Safety Act” in 2006 with a budget of US$4 billion for the purpose of either demolishing and reconstructing seismically dangerous schools or retrofitting the vulnerable ones, where possible. This budget enabled 132,000 classrooms to be demolished and accordingly reconstructed, while allowing another 126,000 classrooms to be retrofitted. Since then, the school retrofitting program has been considerably extended from the regions of higher seismic hazard to other parts of the country and, as a result, more cost data on this subject was deemed to be available to collect. Increasing the availability of this data for use in modelling retrofit cost was the principal motive behind collecting data from schools in Iran.

A seven-month field trip to Iran was scheduled to gather information from as many projects as possible within a designated time frame. The questionnaire developed for this purpose was essentially a reproduction of the online questionnaire described in the previous section. Only some minor, yet relevant, changes were implemented to this questionnaire to make it more compatible with the design provisions and current in place retrofit practice in Iran.

To encourage more participants in the survey, the questionnaire was designed such that to meet a certain set of requisites thought to be of importance in stimulating a positive attitude and response towards such surveys. Like the online questionnaire, this questionnaire was also developed in a semi-structured format, reducing the required time for a given questionnaire to be completed. This format also prevents any misunderstanding of the questions by perspective respondents and therefore ensures the questionnaire quality. An introduction of the purpose and objectives of the study was covered on the first page of the questionnaire form. Confidentiality of survey data and approaches to the management and analysis of the data were also addressed. In addition, instructions were provided to potential participants on how to answer the survey questions properly. Each questionnaire was intended to collect required data related to the retrofit of one school.

Before the formal data collection process began, the researcher met with different retrofit professionals to describe the research in detail. Following these meetings, a pilot survey was conducted at the venue of the Iranian Society of Consulting Engineers (ISCE)
with the selected consultants, having been heavily involved in different aspects of the school retrofitting program. This survey was undertaken to further discuss the research objectives, to test the potential response rate, and also to fine tune the questionnaire. The firms were asked to critically review the overall design of the survey and to make their comments on the suitability and comprehensibility of the questionnaire. As expected, all comments received were positive and, as a result, the questionnaire remained unaltered for the main survey.

Besides the firms that participated in the pilot survey, a list of other potential contributors to this research was identified in this survey. These contributors were either engineering consultancy firms or the firms offering project management services. The latter firms were entrusted with the task of reviewing and approving all the reports presented by the former firms. All these firms were contracted by the State Organization of School Renovation, Development and Mobilization (SOSRDM) as the main authority responsible for the development and implementation of the school retrofit program.

All the firms, listed in the pilot survey, were contacted to collect data. Personal contact was the principal communication tool used to get firms to participate in the study. Those firms who expressed a willingness to cooperate were asked to complete the questionnaire for each school retrofit project with which they were associated. Because of time and resources constraints, many of these firms appeared to be unable to fill out the questionnaire using their personnel. Instead, such firms allowed the researcher to access the archive of those schools whose retrofit studies were completed and enacted by the SOSRDM. The archive of each school encompasses all the documents prepared throughout the whole retrofit procedure for that particular school. Such documents include qualitative and quantitative seismic vulnerability assessment reports, the architectural and structural drawings of the school building, material and geotechnical test reports, modelling and analytical files, a detailed retrofit design report and its associated construction tender prices calculated by the engineering consultant. A careful investigation of these reports led to the development of a cost database.

3.5 Cost Database

As opposed to the data collection effort in New Zealand, where it was highly unsuccessful, such effort in Iran within a seven-month time frame resulted in the creation of a database of 158 cases, each pertaining to a particular earthquake-prone public school with
framed structure (i.e., concrete or steel structure). Each case represents the estimated retrofit construction cost together with data on the predetermined influential factors shown in Figure 3-3. All the data were extracted from the relevant approved retrofit reports and documents.

The cost estimation process can vary depending on the characteristics of the cost estimate limitations. These limitations, such as data availability, may restrict the characteristics of the resulting cost estimate (Stockton and Wang, 2004). It is important to gather a minimum amount of past case data in order to reach a certain acceptance threshold of prediction accuracy (Bode, 2000). Elhag and Boussabaine (1999) recommended that more projects need to be tailored in order to enhance the reliability of cost models. This suggestion is further supported by Günaydın and Doğan (2004) who argued that the performance of cost prediction models strongly depends on the quantity as well as the quality of the examples in the database. Günaydın and Doğan (2004) concluded that the bigger the size of the database, the less the prediction error. The current research is a response to the need for a greater amount of quality data.

The number of projects in the cost database is considered to be a good representative sample for the overall population of schools having a framed structure. This database is one of the comprehensive databases that can be identified in the literature for cost modelling purposes, and hence addresses one of the deficiencies in the cost models developed so far, in which the size of the database was relatively small. A number of studies whose database is smaller than that of this research are summarized in Table 3-1. By using real data from 158 projects, it is anticipated that a more sophisticated and sound cost model can be developed in this research.
### Table 3-1: Cost studies with a database smaller than that of the current research

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Data pool</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potangaroa</td>
<td>1985</td>
<td>86 properties</td>
</tr>
<tr>
<td>De la Garza and Rouhana</td>
<td>1995</td>
<td>16 properties</td>
</tr>
<tr>
<td>Khosrowshahi and Kaka</td>
<td>1996</td>
<td>54 properties</td>
</tr>
<tr>
<td>Elhag and Boussabaine</td>
<td>1998</td>
<td>30 properties</td>
</tr>
<tr>
<td>Hegazy and Ayed</td>
<td>1998</td>
<td>18 properties</td>
</tr>
<tr>
<td>Elhag and Boussabaine</td>
<td>1999</td>
<td>36 properties</td>
</tr>
<tr>
<td>Attalla and Hegazy</td>
<td>2003</td>
<td>50 properties</td>
</tr>
<tr>
<td>Skitmore and Ng</td>
<td>2003</td>
<td>93 properties</td>
</tr>
<tr>
<td>Günaydin and Doğan</td>
<td>2004</td>
<td>30 properties</td>
</tr>
<tr>
<td>Georgy and Barsoum</td>
<td>2005</td>
<td>30 properties</td>
</tr>
<tr>
<td>Chen and Huang</td>
<td>2006</td>
<td>132 properties</td>
</tr>
<tr>
<td>Abu Hammad et al.</td>
<td>2010</td>
<td>113 properties</td>
</tr>
<tr>
<td>Shehab et al.</td>
<td>2010</td>
<td>54 properties</td>
</tr>
</tbody>
</table>

### 3.6 Data Reliability

Besides the data size, another key parameter that has an exceptional role in developing accurate cost prediction models is the reliability of the data. Data reliability is primarily related to data source, and the identification of the position held by the person who completed the questionnaire (Oppenheim, 1992). As described in section 3.4.2, the data source is robust in that the most reliable documents were tailored in this research to elicit the required data. The plausibility of these documents was ensured as all the documents were reviewed by a managerial authority and approved through an extensive refereeing process. The suitability of the questionnaire for collecting the required data was also tested and assured. Although this questionnaire was supplied to the target participants, almost all the questionnaires were filled in by the researcher who had detailed knowledge about the research domain and not by different individuals whose knowledge about the information that they were asked to provide was uncertain. The questionnaire for each school was carefully completed after all the documents developed through the retrofit study of that particular school were thoroughly investigated. As a result, no cases in the database have a missing value. This effort also effectively prevented the projects in the database from being duplicated.
Another reason which adds to the credibility of the database is that the database is comprised of the cost data associated with the planned retrofit program. The costs associated with retrofit endeavours performed as a direct response to observed structural damage after an earthquake were not included in this database. This exclusion arises because the post-earthquake retrofit costs are generally deemed to be inappropriate for the purpose of developing generic RCE models. Such costs, when structural damage exists and when there are pressures to open or re-occupy a damaged building in a limited time after an earthquake, would likely be significantly greater than those of a planned pre-earthquake retrofit (FEMA 156, 1994).

Based on the reasons presented above, it can now be argued that the research database is not only extensive but also objective and reliable. Besides the quantity and quality of the database, another important issue affecting the cost estimating process is the type and level of detail of the cost data in any given database. In general, higher levels of estimating accuracy are normally associated with greater levels of data detail (Stockton and Wang, 2004). The current research database conforms to this particular requirement as it accommodates a rather broad distribution of buildings in terms of structural, site and retrofit policy characteristics, as shown in the next chapter. These characteristics, as illustrated in Figure 3-3, are the potential factors expected to have an influence on retrofit construction cost.

3.7 Model Building

Several of the existing studies that address cost related issues were reviewed in Chapter 2. A close examination of these studies provides insight into the mathematical techniques that are useful in the development of a typical cost model.

As will be shown in the next chapter, substantial variability exists among the retrofit cost of the individual buildings in the cost database. The question then arises of which are the relevant factors that account for this variability. To answer this question, the potential factors shown in Figure 3-3 are examined by employing a regression technique. The backward elimination, forward selection, and stepwise regression techniques are employed to identify major factors that have significant impact on the retrofit cost and consequently to develop multi-linear regression (MLR) models. These models offer simple and easy to use equations for prediction of the retrofit cost. Nevertheless, the regression technique is restrained by some
assumptions which, in turn, impose certain limitations on the MLR models’ predictive performance. In addition, the ordinary least squares (OLS) method is erroneous with respect to the loss function. Studies performed by Skitmore and Cheung (2007) and a year later by Cheung et al., (2008) are the most comprehensive works that support the proposition that the traditional loss functions used in the OLS technique are asymmetric, with the degree of asymmetry increasing according to the level of commercial financial viability at stake. Cost models, especially those mathematical models using mean square error or the like for model training and validation, often treat positive errors (overestimates) and negative errors (underestimates) of equal magnitude the same, with an implicit assumption that the disutility of positive errors is equal to that of negative errors (Cheung et al., (2008)). This assumption may lead to biased OLS predictions because the risks associated with underestimated costs are much higher than the risks associated with overestimated costs, making the traditional loss functions asymmetrical. However, the degree of asymmetry involved is not expected to bias the results significantly. Cherkassky and Ma (2004) presented comparisons between three different methods for linear regression, including the OLS method, support vector machine (SVM), and robust regression methods. Cherkassky and Ma (2004) estimated parameters of regression models in these methods via use of different loss functions, being respectively a standard square loss function, Vapnik’s $\varepsilon$-insensitive loss function, and Huber’s least modulus (LM) loss function. The quality of each model was then evaluated in accordance with its prediction accuracy, as measured in terms of mean squared error (MSE). The results of this comparison showed that the methods’ performance becomes similar when the ratio of the number of samples in the development dataset to the number of independent variables is larger than 2 (i.e., $n/d > 2$). This condition also applies here as the value of the $n/d$ ratio in the present research is more than four times as large as the cut-off value of 2 (i.e., $n/d = 121/14 = 8.64$).

Another approach used in the literature to scrutinise cost estimation problems is the artificial neural network (ANN) technique. This technique is of great potential for solving complex problems, such as those relating to cost estimation, because it is not limited by the constraints that are inherent in the OLS technique. Other advantages of the ANN technique, when compared to the OLS technique, are described in detail elsewhere in the thesis (see Chapter 8). ANN models were developed in this research to estimate the retrofit cost on the basis of the variables found to be statistically significant when using the regression technique. The results obtained from the MLR and ANN-based RCE models are then respectively
compared together using the same datasets used for the development and validation of these models.

The function used to compare the predictive performance of atypical modelling techniques varies with different applications. As can be inferred from the study by Skitmore and Cheung (2007), functions related to error ratio are commonly used to examine how and to what extent the predicted cost values are skewed from the actual cost values. Of these functions, many researchers have accepted the mean absolute percentage error (MAPE) for comparing the estimating accuracy of their MLR and ANN-based cost prediction models (e.g., Emsley et al., 2002; Kim et al., 2004; Shehab et al., 2010; Stockton and Wang, 2004; Wang et al., 2000; Williams, 2002; Williams et al., 2005). The reasons for this acceptance are best defined by Skitmore and Cheung (2007), being (1) analysts are likely to have stronger incentives to minimise their mean absolute forecast errors than their mean squared forecast errors; (2) there is a higher turnover of analysts with poor relative performance measured by high mean absolute errors; (3) low mean absolute error forecasters are more likely to stay or be hired by a top brokerage house; and (4) the Wall Street Journal ranks financial analysts on their average absolute forecast errors. In addition, MAPE is useful for the purpose of reporting the results, because the parameter is expressed in generic percentage terms that will be understandable to a wide range of users (Swanson et al., 2011). Considering these reasons, MAPE was adopted in this research as the skewed cost forecasting function for comparing the predictive performance of the MLR and the ANN modelling techniques.
3.8 Conclusion

To achieve the objectives of this research, a four-stage process was considered within the research methodology. In the first stage, total retrofit project cost was defined and separated into two broad categories incorporating direct and indirect retrofit cost. The activities involved in each of these categories were also identified and the features distinctive to each activity were explained in detail. These activities were then integrated into an innovative retrofit cost framework. As shown later, this framework was employed in this research to estimate the cost of those activities that together constitute the construction cost in general, and seismic cost in particular (see Figure 3-2). The next stage in the research methodology was to single out the factors that may potentially affect the value of the retrofit cost. A wide variety of such factors was identified through a thorough literature review. These factors were further classified into four main groups, being building-based, site-based, social-based and finally retrofit policy-based categories.

As the third stage in the research methodology, a description was given of the procedure and tools employed to gather the required data on retrofit cost and its influential factors. A rigorous data collection effort in Iran resulted in the generation of a large database of 158 data points with an exceptional reliability. Each data point is comprised of the relevant information for an earthquake-prone public school with framed structure. For a given school, this information was carefully elicited by the researcher from the final enacted documents, developed through the retrofit study of that particular school. This database was used as the backbone of this research to model the retrofit cost. It should be noted that an attempt was also made to collect such data in New Zealand by creating an exclusive website (www.retrofitcost.net) where the same questionnaire as was used in the Iran survey was administered online to potential contributors. However, this effort was hampered for the reasons explained in this chapter.

The research methodology was completed by introducing appropriate analytical techniques to be used in this research to establish generic reliable retrofit cost estimating (RCE) models. These techniques (i.e., multi-linear regression (MLR) analysis and artificial neural network (ANN)) and the results obtained from each are discussed, presented and compared in their own chapters. As a preliminary step to perform these techniques, the overall characteristics of the data in the cost database are determined in the next chapter.
CHAPTER 4
DATA CHARACTERISTICS

4.1 Introduction

The data from a total number of 158 schools with framed structures (i.e., steel and concrete structures), whose retrofit study were completed and enacted by the associated management company, were collected in Iran. The data for each school in the database contains the corresponding information on its estimated retrofit construction cost and the variables which are assumed to have a various degrees of impact on this cost. As a preliminary stage to retrofit cost estimation modelling, these variables need to be thoroughly analysed. Distribution modelling was performed in order to gain a better understanding of the characteristics of data collected and also to describe the statistical properties of each variable. The reasons for inclusion of each of the cost determinants are also further elaborated in this chapter.

4.2 Retrofit Net Construction Cost (RNCC)

The retrofit cost values were extracted from the tender documents prepared by the engineering consultants after the completion of the retrofit design phase. Only the costs of those construction works which directly improve the building’s seismic performance were estimated in the tender documents. Neither non-construction cost nor non-seismic cost was addressed in these documents. The construction works, in association with the latter cost, were beyond the scope of the school retrofit program. The cost of activities that arise as a result of the retrofit construction work (i.e., indirect cost) also was not considered when calculating the bid price, as it was thought that these activities are unique to specific projects.

The bidding process is usually held to comprise two stages (Newton, 1991): (1) the basic cost estimate – which is intended to establish the cost to the builder, (2) the mark-up – which is an amount added to the basic cost to cover items such as overheads, profit and variations in market conditions. The basic cost and mark-up together comprise the bid price.
The bid price of retrofitting any school in the research database is the product of multiplication of the retrofit net construction cost (RNCC), as the basic cost, and a number of relevant coefficients including:

- Overhead coefficient
- Geographic location coefficient
- Storey coefficient
- Height coefficient
- Mechanical and electrical renovation coefficient
- Contingency coefficient
- Price adjustment coefficient
- Retrofit work complexity coefficient
- Site preparation coefficient.

In the tender documents investigated, no formal method was found to be followed by engineering consultants in defining the values of the coefficients above. Subjective judgement was often used when determining these values. The product of these coefficients multiplied one by another was varied in a broad range from a minimum of 1.00 to a maximum of 3.12. Because of the added uncertainties and complexities of estimating these coefficients, the bid price amount is prone to errors and could not be directly tailored to develop retrofit cost estimating (RCE) models. Instead, RNCC is a better representation of retrofit cost as it provides a common base for such cost. As a result, RCE models in this research were targeted on the RNCC rather than the bid or tender price (see Equation (4.1)).

\[
RNCC = \frac{\text{bid or tender price}}{\prod_{i=1}^{9}(\text{cost coefficient})_i}
\]

As the RNCC values in the survey database were based on different dates, these values were normalized to the base year 2008, using the Iranian Construction Cost Index (ICCI) which is annually published by the Deputy of Strategic Planning and Control. To express the RNCC values in U.S. dollars, an exchange rate of U.S$1=10,000 Rials was used. The Rial is the currency unit in Iran. The average rebased RNCC values in the database is
U.S$95,527 thousand, and the highest and lowest costs are U.S$293,32 thousand and U.S$11,288 thousand, respectively.

When working with data that varies randomly within some bounds, it is often of significant interest to examine different distributions and consequently select the one which statistically matches better to the available empirical data. This issue deserves to be carefully studied as it represents the probabilistic behaviour of a given numerical variable under consideration. Using the Komolgorov-Smirnov (K-S) goodness-of-fit test, the probabilistic behaviour of RNCC and RNCC per unit area values was found to be better described by lognormal distribution than by other continuous distributions examined in this research including, amongst others, normal, logistic, gamma, beta, weibull, and exponential. This conclusion is drawn by comparing the p-values computed for different distributions with respect to their K-S test statistic. The maximum p-value of this test was obtained for a lognormal distribution with a value of 0.092 and 0.379 for data pertaining to RNCC and RNCC per square meter variables, respectively. As these p-values are above the conventional 5% level of significance (p>0.05), the null hypothesis that the data of these variables follow the lognormal distribution cannot be rejected. The RNCC and RNCC per unit area histograms and their corresponding best fitted distribution are shown in Figures 4-1 and 4-2, respectively.

According to Emsley et al., (2002), using a cost variable in terms of cost per square metre would cause any variation in project cost to be proportional to the size of the building, rather than being proportional to the original cost. For instance, assume that the construction costs of two buildings having areas of 50 m² and 100 m² are U.S$1,000 and U.S$2,000, respectively. Given this assumption, the two buildings incur the same cost per area (i.e., 20 U.S$/m²), while in fact the latter costs twice as much as the former. This similarity arises because the value of cost variable, when measured in terms of cost per m², is influenced by the building size and, therefore, this variable does not reflect the real magnitude of the cost. Such influence can be considered in another way by adding the building size to the list of variables by which the sole values of cost can be predicted. Consequently, RNCC rather than RNCC per area was taken into account as the cost variable in this study. As will be shown later in Chapter 5, in order to address the large cost differences, a common solution is to model the natural logarithm of the cost instead of the raw cost. This type of modelling assumes that the natural logarithm of cost is normally distributed, which is proven to be the case in this study by the fact that the lognormal distribution is accurately representing the
distribution of RNCC values in the database. This finding alternatively suggests that the normal distribution gives a precise fit to the natural logarithm values of RNCC.

Figure 4-1: RNCC histogram and lognormal fitting (N=158)

Figure 4-2: RNCC per unit area histogram and lognormal fitting (N=158)
4.3 Building Characteristics

4.3.1 Building Area

When establishing any cost estimation model, the most likely parameter to be considered is building size. Several citations from the literature reinforce the appropriateness and significant contribution of this variable to the cost models (e.g., Abu Hammad et al., 2010; Chen and Huang, 2006; Elhag and Boussabaine, 1999; Emsley et al., 2002; Kim et al., 2004; Kim et al., 2005; Lowe et al., 2006; Stoy and Schalcher, 2007). The importance of the building area when costing retrofit construction works has also been confirmed in previous studies in which the magnitude of the retrofit cost estimate was mainly affected by this variable (FEMA 156, 1994; Potangaroa, 1985).

By superimposing a number of continuous distributions over the data pertinent to the building area, the K-S test revealed that the best fitting distribution to this data is a gamma distribution. However, the null hypothesis that the data came from a particular distribution cannot be rejected at the 5% level of significance for a set of distributions including gamma, max extreme, lognormal, and weibul distributions, as the p-values of their corresponding K-S test are 0.131, 0.159, 0.068 and 0.150, respectively. The variation of the buildings in the database with respect to their area and the gamma distribution fitted to the building area values are portrayed in Figures 4-3 and 4-4, respectively.

![Figure 4-3: Building distribution as a function of building area (N=158)](image-url)
4.3.2 Building Total Number of Stories

Height of building measured in terms of number of stories is another reflection of the project size and is usually considered as a potential independent predictor of cost (Khosrowshahi and Kaka, 1996). In buildings with a greater number of stories, overturning and shear forces may require a proportionately greater RNCC. Furthermore, there are some instances from previous earthquakes where the amount of damage was commensurate with the number of stories. For example, the damage assessment study conducted by Gülhan and Güney (2000) revealed that during the 1999 Marmara earthquake in Turkey, reinforced concrete frame structures with more stories experienced higher level of damages compared to low-rise buildings. The damage ratio also increased as the number of stories increased.

Figure 4-5 shows the distribution of the buildings in the database as a function of the number of stories.
4.3.3 Building Age

Age can be an important determinant of RNCC because older buildings often require taking more lateral resisting measures into account. In addition, as the building gets older, it is more likely that the number of existing structural elements suffering from deterioration increases and, consequently, more elements are subjected to serious modification and replacement.

By employing a life-cycle cost analysis approach, Arikan et al., (2005) came to the conclusion that the age of building, in conjunction with retrofit level (expressed in terms of percentage of the retrofit cost to the initial construction cost), are dominant parameters in making the crucial decision as to whether to demolish and rebuild earthquake-prone buildings or retrofit them. In their study, it was assumed that the incurred retrofit cost would be accordingly increased as the buildings get older, while the rate of increase is dependent on the retrofit levels. According to this assumption, the retrofit cost for a 40-year-old building was computed to be 9% and 28% more than that of a 10-year-old building for retrofit levels of 20% and 80%, respectively. Critical retrofit levels, above which the retrofit is no longer economically viable, were also found to vary almost inversely linearly from 25% for 40-year-old buildings to 67% for 10-year-old buildings. This finding implies that in comparison with the demolish and rebuild option, the retrofit option gradually loses its credibility as building

![Figure 4-5: Building distribution as a function of number of stories (N=158)](image)
age increases. It can also be concluded that retrofitting of older buildings is of less worth compared to buildings with less age.

As buildings in the database of the present study also have different ages which, in turn, may indicate different amounts of retrofit cost, building age is taken as an influential variable. The gamma distribution was again found to be the most suitable distribution to describe the building age variable. In addition to this distribution, the lognormal distribution is also a good representation for this variable. The building age values in the database can be well described by these distributions as the p-values of their corresponding K-S test are above the 5% level of significance (i.e., 0.245 and 0.053, respectively). The spread of the buildings in the database with respect to their age is illustrated in Figure 4-6 and the best fitting distribution (gamma) is displayed in Figure 4-7.

![Figure 4-6: Building distribution as a function of building age (N=158)](image_url)
4.3.4 Building Weight

Heavy buildings are potentially a seismic hazard because the collapse of some certain type of structures in previous earthquakes was reported to be mainly due to excessive building weight (Maqsood and Schwarz, 2010). The general method for determining the total lateral seismic force, known as base shear, is based on the following simple equation which linearly relates seismic induced force \(V\) to the building weight \(W\) and a multi-aspect coefficient \(C\) accounting for other influential factors ranging from region-dependent seismic characteristics to structural characteristics. According to Equation (4.2), the heavier the building the more the seismic force is exerted. Therefore, a heavy building is subject to more force and possibly more seismic damage than is a lightweight building, given the same acceleration.

\[
V = C \times W
\]  

(4.2)

Extreme care is required when attempting to include the building weight in RCE models. From a theoretical perspective a close relationship between the building area and the building weight is well known. It might therefore be assumed that these two variables are highly correlated and that it is not allowable to include both at the same time in an analysis. To resolve this problem, the building weight was first converted to a building weight
indicator in terms of “weight per unit area”, and then entered into RCE models as an influential variable. As will be shown later in Chapter 5, this conversion prevents the multicollinearity problem from occurring. The distribution of the buildings in the database as a function of this variable is also depicted in Figure 4-8.

![Figure 4-8: Building distribution as a function of building weight indicator (N=158)](image)

4.3.5 Building Configuration

Buildings are classified as regular or irregular based on their configuration. Irregular configuration in existing buildings generally arises because of their functional complexity or sometimes because of the architect’s or owner’s desire to create an original or striking architectural layout. However, seismic design codes usually set a number of limitations on the use of buildings having an irregular configuration because they may have a detrimental influence on the effectiveness of the seismic engineering and consequently on building seismic performance.

There are two types of irregularities, being plan and vertical. These irregularities are responsible for two undesirable conditions in an earthquake. Plan irregularity may cause torsional forces to develop which are believed to be the most frequent cause of structural failure during an earthquake (FEMA 454, 2006). Alternatively, vertical irregularity tends to create abrupt changes in strength or stiffness that may concentrate forces at one or several
points in a structure (such as a particular set of beams, columns, or walls) in an undesirable way. Those members may fail, and by chain reaction bring down the entire building (FEMA 454, 2006). The most serious condition of stress concentration occurs when a building has a soft or weak storey which is significantly weaker or more flexible than others. In extreme cases, the vertical irregularity of the soft storey may lead to collapse unless adequate retrofit measures are provided. Retrofit measures are more desirable to focus on removing, or at least reducing, configuration irregularities to the greatest extent possible. These measures should provide a direct load path which is an important characteristic in order to secure good seismic behaviour. The possible measures may involve providing seismic separations in irregular configurations, or adding new lateral resisting elements (e.g., shear walls, braces, etc.) in locations in order to diminish torsional effects, or removing excessive mass through demolishing upper stories.

The distribution of the buildings in the database with regard to plan and vertical irregularities is illustrated in Figures 4-9 and 4-10, respectively. The following figures show that the percentage of the buildings with a regular plan configuration (77%) is slightly less than those having a regular vertical configuration (80%).

![Figure 4-9: Building distribution as a function of plan irregularity (N=158)](image-url)
4.3.6 Structural Type

Typically the evaluation of existing buildings starts with identification of the building type and damaging characteristics of each type (FEMA 454, 2006). The necessity of this identification mainly stems from the fact that the level of vulnerability of a building depends to a large extent on the building structural type and the level to which the original design of the structural system has addressed the seismic forces. The wide range of building types in a given building stock can accordingly result in a wide range of damage states, ranging from very light (a few cracks) to severe (parts of the structure rendered ineffective and potentially unsafe) or, in the worst case, to complete collapse in an expected seismic event. Different levels of expected damage state would indeed govern the extent of required retrofit works and therefore have a significant effect on the retrofit cost. In a general sense, buildings designed and built with the absence of structural lateral force resisting systems are amongst the most vulnerable buildings, that sustain serious damage when struck by an earthquake, and therefore their retrofit work are the more costly.

The structural type is also of great importance in other associated seismic evaluation methods. For instance, to obtain a broad understanding of the vulnerability and the relative risk among the many buildings, one may use available proposed rapid visual screening methods. All these methods consider structural type as the first, and perhaps the most important, indicator to differentiate buildings in terms of their likely seismic proneness.
Usually the buildings without any in-place resisting system are taking as potential candidates for further detailed seismic and economic feasibility studies, to see whether their retrofit is practically warranted. Sometimes the retrofit cost in the case of these buildings is so prohibitive that it may change the final decision in favour of other options (e.g., status quo, demolish and rebuild) rather than retrofit.

The structural type may also serve as a reliable factor when several issues pertaining to the managerial aspect of earthquake policies are to be set. For instance, insurance companies are more concerned with building type than with other factors, and subject earthquake insurance to a minimum deductible amount depending on the type of construction (FEMA 68, 1985). The deductible rate for the structural types indicating the lower seismic vulnerability is considerably less than for the building types that are well-known for their poor seismic performance.

Figures 4-11 to 4-14 demonstrate the representative structural types in this research database. The figures respectively exemplify concrete moment resisting frames, one-way steel braced frames (braced frames exist in only one principal direction while the frames in the other direction were left unbraced), two-way steel braced frames (braced frames exist in both longitudinal and transversal directions), and finally bare steel frames without any lateral resisting elements but infill unreinforced masonry walls.
Figure 4-11: An example of the buildings with concrete moment resisting frames

Figure 4-12: An example of the buildings with one-way steel braced frames
Figure 4-13: An example of the buildings with two-way steel braced frames

Figure 4-14: An example of the buildings with bare steel frames
To examine the extent to which the structural type may affect RNCC, buildings in the database are classified into two general categories: (1) buildings in which seismic resisting system exists in both directions (i.e., concrete moment resisting frames and two-way steel braced frames) and (2) buildings which lack such a system in at least one direction (i.e., one-way steel braced frames and bare steel frames). The distribution of the buildings in the database with respect to these categories is shown in Figure 4-15. According to this figure, buildings are almost evenly distributed between each category.

![Figure 4-15: Building distribution as a function of structural type (N=158)](image)

### 4.3.7 Diaphragm Type

Diaphragms are amongst the critical elements that provide support and stability for the overall building during an earthquake event. Diaphragms, together with the lateral force resisting system, form a horizontal structural system that connects the vertical elements of the seismic force resisting system and carries their loads down to the foundation. Diaphragms therefore require adequate strength and stiffness to resist the lateral forces during an earthquake event. In order to achieve this objective, careful consideration of the existing diaphragm configuration is needed.

Diaphragms can be classified as either rigid or flexible. Rigid diaphragms can be found more in modern buildings while flexible diaphragms were frequently used in older
buildings. There are two types of diaphragms which are dominant in the case of the buildings comprising this research database: (1) brick archaic flooring (also known as jack arch flooring) and (2) concrete joist and hollow infill block flooring. The latter diaphragm type is amongst the rigid diaphragm category and is therefore less likely to be an issue when compared to the former diaphragm type, which is generally categorized as a flexible diaphragm. Flexible diaphragms are of great concern as their flexibility is an existing threat to the building structural stability in the event of an earthquake. A brick archaic flooring system, when subjected to seismic loads, is expected to detach from its supports and consequently lose its integrity and strength. The measures commonly used to mitigate these deficiencies are adding a new reinforced concrete topping slab to form, in combination with existing steel floor beams, a composite deck or, alternatively, adding diagonal strapping members acting together with existing joists as a horizontal braced frame diaphragm. However, regardless of what type of retrofit measure is to be taken into account, diaphragm retrofit would often necessitate partial removal of existing floors and associated finishes, which can be both costly and disruptive.

In this research, the role of diaphragms in the determination of RNCC is studied. Buildings in the database are classified into two general categories with respect to their existing diaphragm types. As shown in Figure 4-16, almost 30% of the buildings have flexible diaphragm, whereas rigid diaphragm accounts for the remaining 70% of cases.

![Figure 4-16: Building distribution as a function of diaphragm type (N=158)](image)
4.3.8 Foundation Type

A thorough seismic evaluation of an existing building should also include an examination of the existing foundation and an assessment of its available capacity to transmit the loads from the superstructure to the soil on which the building is resting. Foundations also have an important role in mitigating damage imposed by site specific hazards like liquefaction. Therefore, the characteristics of an existing foundation need to be carefully studied. Of these characteristics, foundation type is perhaps the most crucial detail whose likely effect on RNCC is examined in this research.

To scrutinize whether or not foundation type would necessitate a meaningful alteration in the retrofit cost, the buildings in the database were classified into two groups: (1) buildings with reinforced concrete isolated footing and (2) buildings with reinforced concrete strip footing. According to Figure 4-17, each foundation type accounts almost equally for 50% of the buildings in the database.

![Figure 4-17: Building distribution as a function of foundation type (N=158)](image)

4.3.9 Non-structural Equipment and Component (Infill Wall Mitigation)

Elements and components that affect the lateral stiffness or distribution of forces in a structure, or are loaded as a result of lateral deformation of the structure, can be classified as
primary or secondary elements, even if they are not part of the intended lateral force resisting system (FEMA 356, 2000). The structural elements discussed so far are amongst the primary elements of a framed structure that provide the capacity of the structure to resist seismic forces induced by ground motion. Other elements and components can be classified as secondary elements as they do not contribute significantly or reliably to the building’s overall response to earthquake ground motions because of their low lateral stiffness, strength, or deformation capacity. However, there are some cases when the retrofit solution is extended to the case of secondary elements, particularly infill walls. It is therefore of interest in the present research to also examine the likely effect of this extension on RNCC variation. Figure 4-18 shows that infill wall mitigation is not considered for almost 61% of the buildings in the database. This percentage implies a common practice of ignoring the likely contribution of these components to the building’s overall resistance against lateral forces.

![Figure 4-18: Building distribution as a function of infill walls mitigation (N=158)](image)

### 4.4 Site Characteristics

#### 4.4.1 Seismicity

A building’s seismic performance is highly dependent on the seismic characteristics of the region where it is located. The seismicity can be adequately described by frequency content of ground motion, duration, and the level of ground shaking. These characteristics
should be properly incorporated into a seismic assessment as they govern both the displacement and energy related seismic demands of the structural system. Furthermore, the base shear being applied to the structure to approximate the earthquake effects (Equation (4.2)) is dependent on local seismicity in a sense that the greater the seismicity, the more lateral force generated and, hence, the building should resist greater earthquake-induced forces. As a result, seismicity is amongst the most critical parameters that should be carefully studied in RCE modelling. To examine the extent of seismicity effect on RCE models, regions should be differentiated on the basis of their seismicity.

In the present study, the method for defining seismicity is directly acquired from the seismicity map embedded in the most recent revision of the Iranian code of practice for seismic resistant design of buildings (Standard 2800, 3rd revision (BHRC, 2005)). The map was developed using the general principles of seismic risk assessment and it is intended to represent broad regional variations in seismic risk. The map demarcates different seismicity regions from each other. Four seismicity regions are introduced in this code, namely, low, moderate, high, and very high. As shown in Figure 4-19, these regions are differentiated based on peak ground acceleration (PGA) having a 10% probability of exceedance in the expected lifetime of a building, which is 50 years.

![Seismicity map of Iran](image)

Figure 4-19: Seismicity map of Iran (Standard 2800, 3rd revision (BHRC, 2005))
It is very important to note that only one building in the database is located in the moderate seismicity region while no building was found to exist in the low seismic zone. This limitation of available data on low and moderate seismicity regions may import significant uncertainty to retrofit cost estimates. To avoid this problem, buildings in the database were classified into two broad groups: (1) buildings located in regions with “very high” seismicity and (2) buildings located in other regions with less seismic activity. The distribution of buildings with respect to this classification is illustrated in Figure 4-20. As can be inferred from this figure, almost 80% of the buildings are located in the former regions while the less severe seismic zones account for the remaining 20% of the buildings.

![Figure 4-20: Building distribution as a function of seismicity (N=158)](image)

4.4.2 Soil Type

In current seismic design codes, the site effect is required to be carefully considered in any analysis procedure as lateral induced forces are also greatly influenced by the subsurface conditions as well as interaction between the soil and the structure, known as soil-structure interaction. This issue basically arises from the fact that the potential ground motion from a single earthquake may vary considerably depending on the seismic characteristic of the soil type. As seismic vibrations propagate towards the earth’s surface, they may be amplified depending on the intensity of the shaking, the nature of the rock and, above all, by the surface
soil type (FEMA 577, 2007). Earthquake shaking tends to be more severe on soft ground than in stiff soil or rock. Consequently, more building damage is expected to occur in areas of soft soils. Studies conducted after the 1989 Loma Prieta earthquake showed that the shaking in the soft soils was 2.5 to 3.5 times that of in rock (FEMA 577, 2007). The soil amplification effect is therefore incorporated when developing design response spectrum which form the basis for calculating the earthquake induced forces in most building seismic design codes.

The design response spectrum of the current Iranian seismic design code (Standard 2800, 3rd revision (BHRC, 2005)) is characterized by four soil types starting with type I through to type IV. The general shape of this spectrum is illustrated in Figure 4-21. Given the same natural period of vibration, Figure 4-21 indicates that the building base shear continually increases from the first to the last soil type.

![Figure 4-21: Response coefficient of buildings with respect to different soil types (Standard 2800, 3rd revision (BHRC, 2005))](image)

For the same reason as stated in the case of seismicity, the soil types in Figure 4-21 are divided into two broad classes: (1) well-behaved and (2) poorly behaved soils. The first two sets of soil types (i.e., I and II) comprise the former class while the latter class includes the last two sets of soil types (i.e., III and IV). The distribution of buildings in the database with regard to this classification is depicted in Figure 4-22. This figure indicates that the ground motion effect on almost two third of the buildings in the database is heightened as a result of the so-called soil amplification factor.
4.5 Social Characteristics

The social-based variables are held fixed in this study as they are not varied with the attributes they can possibly take on. There is only one particular answer for each of these variables. For instance, all the buildings in the database are public schools with the same importance level. These schools are of no historical value and, as a result, their retrofit objective was set to be “life safety”, which allows for un-repairable damage as long as life is not jeopardized and egress routes are not blocked.

4.6 Retrofit Policy Characteristics

4.6.1 Compliancy with the First Building Seismic Design

Building codes provide the minimum acceptable standards used to regulate the design, construction, and maintenance of buildings for the purpose of protecting the health, safety, and general welfare of the building users. The apparent earthquake risk, particularly within the large geographic areas and population that would be affected, combined with the lack of appropriate building codes to reduce earthquake damage, are of great concern. Devastating earthquakes occurred in seismic-prone countries have focused much needed attention on the more stringent seismic practices. The adoption and enforcement of modern building codes is
a major step toward becoming better prepared for an earthquake. During an earthquake, it is more likely that buildings constructed in accordance with the requirements of recent building codes perform far better than buildings designed on the basis of previous immature in-place codes. It might also be true that buildings required to be constructed in compliance with the minimum provisions of premature building practices would, at least partially, suffer a lower level of damage when compared to buildings built at a time when no seismic codes were available to be followed.

A considerable number of buildings in the existing building stock of Iran present a risk of poor performance in earthquakes because there was no seismic design code available or required when they were constructed, or because the seismic design code used was immature and had flaws. Many building types designed under previous seismic provisions are also now considered deficient. Similarly, almost all the schools being studied in the present study fell – to various degrees - short of complying with the current in-place seismic code requirements. By reviewing the available buildings’ as-built records, it became apparent that not all the buildings were even designed or constructed in accordance with the relevant requirements set out in the first Iranian seismic code (Standard 2800, 1st edition (BHRC, 1988)). This issue also holds true for a number of buildings built after the official adoption and enforcement of this code in 1988. To examine whether or not complying with these code provisions would affect the retrofit cost, buildings in the database were divided into two groups as shown in Figure 4-23. According to this figure, almost half of the buildings in the database were designed or constructed without taking these code provisions into consideration.
Figure 4-23: Building distribution as a function of compliancy with the provisions of the first in-place seismic design code (Standard 2800, 1st edition (BHRC, 1988)) (N=158)

4.6.2 Seismic Guideline Used for Retrofitting

Like the social-based variables, this variable is also held fixed in this research because the seismic performance of all the schools in the database was evaluated in accordance with the regulations of a specific technical document, known as “instruction for seismic rehabilitation of existing buildings”. It should be noted that this document was approved as Code No. 360 in 2007. Since then, its use has been mandated for seismic assessment and retrofitting of all public buildings.

4.6.3 Retrofit Measure

A broad range of practical and effective seismic retrofit techniques can be found in building seismic codes, guidelines, etc. These techniques can be traditionally categorized within three basic classes, namely, adding new elements, enhancing performance of existing elements, and improving connections between components (FEMA 547, 2006). In this study, the global stiffness and/or strength of the schools are insufficient to reasonably resist against lateral induced forces during an expected earthquake. Such deficiencies arise due to either a complete lack of seismic resisting system or an inadequate design to an early in-place code. As a result, the retrofit techniques within the former class (add new elements) have obvious
preference and priority over others. Amongst these techniques, there are two well-practiced measures: adding new reinforced concrete shear walls and adding new steel braced frames. These techniques were most frequently considered by structural engineers to enhance seismic performance of the representing schools in the database. General sketches of these measures, their locations where added to the building’s plan, and their distributions along the building’s height are respectively illustrated in Figures 4-24 and 4-25.

Variation in the amount and type of materials in conjunction with different levels of associated labour work required for a complete implementation of each of the aforementioned retrofit measures may cause the retrofit cost to fluctuate. To determine the extent of this fluctuation, buildings in the database were accordingly grouped into two subgroups as shown in Figure 4-26. According to this figure, the number of buildings that were to be retrofitted by adding steel braced frames is more than 2.5 times as many as buildings to be retrofitted by adding concrete shear walls.
Figure 4-24: An example of the buildings to be retrofitted by adding new concrete shear walls

Figure 4-25: An example of the buildings to be retrofitted by adding steel braced frames
4.6.4 Occupancy Condition during Construction Work

The implementation of the retrofit measures discussed above would necessitate all the buildings to be properly evacuated before their retrofit construction work starts. As a result, this parameter is also held fixed in this research.
4.7 Conclusion

This chapter was mainly devoted to explore the statistical properties of the variables included in the research data pool of 158 earthquake-prone public schools. The descriptive statistical summary of the metrically and not metrically scaled variables are provided in Tables 4-1 and 4-2, respectively. In the next chapter, the relationships between RNCC and other variables are analysed using the regression technique.

Table 4-1: Statistical summary of metrically scaled variables in the database (N=158)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Minimum</th>
<th>Lower quartile (25th percentile)</th>
<th>Median (50th percentile)</th>
<th>Upper quartile (75th percentile)</th>
<th>Maximum</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNCC (U.S$)</td>
<td>11,288</td>
<td>62,437</td>
<td>84,649</td>
<td>118,503</td>
<td>293,320</td>
<td>95,527</td>
</tr>
<tr>
<td>RNCC (U.S$) per m²</td>
<td>20.41</td>
<td>41.58</td>
<td>52.07</td>
<td>62.61</td>
<td>226.24</td>
<td>53.60</td>
</tr>
<tr>
<td>Building area (m²)</td>
<td>187</td>
<td>1,276</td>
<td>1,720</td>
<td>2,500</td>
<td>6,100</td>
<td>1,863</td>
</tr>
<tr>
<td>Building number of stories</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3.23</td>
</tr>
<tr>
<td>Building age (year)</td>
<td>5</td>
<td>13</td>
<td>18</td>
<td>27</td>
<td>57</td>
<td>20</td>
</tr>
<tr>
<td>Building weight indicator (ton/m²)</td>
<td>0.15</td>
<td>0.97</td>
<td>1.12</td>
<td>1.28</td>
<td>2.08</td>
<td>1.13</td>
</tr>
</tbody>
</table>
### Table 4-2: Statistical summary of variables in the database that are not metrically scaled (N=158)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Spread</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plan irregularity</td>
<td>77% Regular</td>
</tr>
<tr>
<td></td>
<td>23% Irregular</td>
</tr>
<tr>
<td>Vertical irregularity</td>
<td>80% Regular</td>
</tr>
<tr>
<td></td>
<td>20% Irregular</td>
</tr>
<tr>
<td>Structural type</td>
<td>47% Lateral seismic resisting system exists in both directions</td>
</tr>
<tr>
<td></td>
<td>53% Lateral seismic resisting system does not exist in at least one direction</td>
</tr>
<tr>
<td>Diaphragm type</td>
<td>70% Rigid diaphragm (concrete joist and hollow infill block flooring)</td>
</tr>
<tr>
<td></td>
<td>30% Flexible diaphragm (brick archaic flooring)</td>
</tr>
<tr>
<td>Foundation type</td>
<td>51% Reinforced concrete isolated footing</td>
</tr>
<tr>
<td></td>
<td>49% Reinforced concrete strip footing</td>
</tr>
<tr>
<td>Infill walls mitigation</td>
<td>61% Not considered</td>
</tr>
<tr>
<td></td>
<td>39% Considered</td>
</tr>
<tr>
<td>Seismicity</td>
<td>20% Regions with less seismic risk</td>
</tr>
<tr>
<td></td>
<td>80% Regions with the highest seismic risk</td>
</tr>
<tr>
<td>Soil type</td>
<td>35% Well-behaved soil types (soil types I and II)</td>
</tr>
<tr>
<td></td>
<td>65% Poor-behaved soil types (soil types III and IV)</td>
</tr>
<tr>
<td>Compliancy with the provisions of the first seismic design code</td>
<td>49% Complying</td>
</tr>
<tr>
<td></td>
<td>51% Non-complying</td>
</tr>
<tr>
<td>Retrofit measure</td>
<td>27% Add new concrete reinforced shear walls</td>
</tr>
<tr>
<td></td>
<td>73% Add new steel braced frames</td>
</tr>
</tbody>
</table>
CHAPTER 5

DEVELOPING REGRESSION MODELS FOR ESTIMATING RETROFIT NET CONSTRUCTION COST (RNCC) IN FRAMED STRUCTURES

5.1 Introduction

A literature search revealed that few resources are currently available to support a parametric retrofit cost estimating (RCE) model. The lack of a comprehensive set of cost predictors within these models is particularly serious. Also, little effort has been devoted to properly exploring the extent of the influence of the cost governing factors as the independent (i.e., predictor) variables on the retrofit cost as the dependent (i.e., response) variable. In addition, the prediction ability and reliability of these models is questioned as these models have seldom been validated on datasets other than those from which they were originally derived. Utilizing a set of independent cost variables, some of which have never been previously reported in the literature, an attempt is made in this chapter to tackle and resolve the deficits above through the use of a mathematical, yet traditional, modelling approach; that is, regression analysis.

RNCC and its fourteen predictors were initially analysed in Chapter 3 on the basis of the data collected from 158 earthquake-prone framed structures. Based on this amount of data, the main objective of this chapter is to describe the development of robust regression models to predict the retrofit net construction cost (RNCC) of framed structures, with an emphasis on identifying the most critical predictors that have significant impact on the explicit determination of RNCC. While such models would be valuable in themselves, developing these models fulfils two other purposes: they assist in the development of artificial neural network (ANN) models by providing different combinations of variables that demonstrate a strong relationship with RNCC; and they also provide a useful benchmark against which the performance of the ANN models can be measured.

The chapter starts with a brief explanation of MLR analysis and an investigation of the key issues pertaining to this analysis, including the functional form of variables to be used in the MLR models. The chapter continues with a description of the technique used to develop MLR models and then the criteria examined for selecting the most appropriate model is presented. Detection of the proposed model violations with respect to standard regression
assumption, known also as model diagnostic, constitutes the next part of the present chapter. Model diagnostic is followed by an explanation of the causal relationships between RNCC and each of its predictors in the proposed model with regard to the model functional form. The chapter is concluded with a comparison of the predictive ability of the models, each model containing with a unique combination of the most statistically significant predictors, and utilising the same set of new data that has never previously been exposed to such models during their development process. Given the results of this comparison, as measured by determination of coefficient \(R^2\), it is revealed for the first time in the literature that the retrofit cost can be accurately estimated by using the most parsimonious model in the form of \(C=K\alpha^B\).

### 5.2 Multi-linear Regression (MLR) Analysis

Regression analysis is the most popular statistical technique (Aljahdali, 2003; Koppula, 1981; Wilmot and Mei, 2005). This technique has been traditionally used to develop numerous successful cost estimation models, some of which were presented in Chapter 2. Regression analysis is attractive to modellers not only because it provides a conceptually simple method for investigating functional relationships amongst variables (Chatterjee and Hadi, 2006), but also because it lends so easily to model interpretation and comparison (Lunneborg, 1994). The models developed by regression analysis are therefore capable of measuring the direction and the strength of the relationship that potentially exists between variables, while numerically describing that relationship.

In regression analysis, the dependent (i.e., response) variable and the set of \(p\)-independent (i.e., predictor or explanatory) variables are commonly denoted by \(Y\) and \(X_1, X_2, \ldots, X_p\), respectively. The true relationship between \(Y\) and \(X_1, X_2, \ldots, X_p\) can be approximated by the following regression model:

\[
Y = f \left( X_1, X_2, \ldots, X_p \right) + \varepsilon
\]

where \(\varepsilon\) is assumed to be a random error representing the discrepancy in the approximation. This random error accounts for the failure of the model to fit the data exactly.
The function \( f(X_1, X_2, \ldots, X_p) \) describes the relationship between \( Y \) and \( X_1, X_2, \ldots, X_p \). An example is the linear regression model as follows:

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon
\]  

(5.2)

where \( \beta_0, \beta_1, \ldots, \beta_p \), called the regression parameters or coefficients, are unknown constants to be estimated from the data.

It is assumed that for any set of fixed values of \( X_1, X_2, \ldots, X_p \) that fall within the range of the data, the linear equation above provides an acceptable approximation of the true relationship between \( Y \) and the \( X \)'s. The most commonly used method to estimate the parameters of the model presented in Equation (5.2) is called the ordinary least squares (OLS) method (Studenmund, 2011). Under certain assumptions (to be discussed in detail in this chapter), the OLS method produces estimators with desirable properties. Much of the appeal of this technique lies with its simplicity and also its easy accessibility from many of the popular statistical packages (Aljahdali, 2003). Another attractive feature of this method is that small or minor violations of the underlying assumptions do not invalidate in a major way the inferences and conclusions drawn from the analysis (Chatterjee and Hadi, 2006).

Regression models may have many uses and functionalities which occasionally overlap, and an equation constructed may meet different aims. The main point to be noted is that the purpose for which the regression equation is developed determines the criteria that are to be optimized in its formulation. According to Chatterjee and Hadi (2006), the general applications of these models can be summarized in three main categories:

1- Description and model building
2- Estimation and prediction
3- Control

In present research, the main objective of employing the MLR methodology has matched well with the combination of the first and second categories because on the basis of the developed regression equations, it is intended to firstly clarify the nature of interaction which does exist among dependent and predictor variables and accordingly identify the most influential predictors; secondly propose an equation which can best predict the cost value of a future retrofit project according to a given set of such predictors.
5.3 Variables to Include

Regression analysis is a statistical technique whose basic concept is that the value of a dependent variable can be found using a mathematical formula, which contains a variety of other independent variables whose values are known or can be found (Soutos and Lowe, 2005). In the case of regression models for cost estimating, the dependent variable would normally be the cost and the independent variables are those factors that affect the cost.

Fourteen questionnaire items were initially chosen as possible independent variables affecting the dependent variable of interest, i.e., RNCC, in the context of earthquake-prone framed structures. Table 5-1 gives the description of all these variables. According to Table 5-1, there are three broad types of variables included in the study. Variables $X_1^*$, $X_2^*$, $X_3^*$, $X_4^*$, $X_5$, $X_8$, $X_9$, $X_{11}$, $X_{12}$, and $X_{13}$ directly address different characteristics of the building, while variables $X_{10}$, and $X_{14}$ are of more retrofit solution nature, and variables $X_6$ and $X_7$ are assigned to take the site characteristics into account. Of these predictors, variable $X_5$ may also serve as a useful benchmark when there is an urgent need for setting a viable retrofit policy. These independent variables define the retrofit cost and to various degrees have an impact on RNCC. However, as will be shown later in this chapter, only a handful of these variables are incorporated into the final model, as the inclusion of the remaining variables contributes little to improve the performance of this model.

In addition to categorising variables as either dependent or independent, in Table 5-1 these variables are divided into two other broad classes: quantitative and qualitative variables.

The quantitative variables are the ones comprised of real numbers, including $Y^*$, $X_1^*$, $X_2^*$, $X_3^*$, $X_4^*$. As the range of these variables differed by more than one order of magnitude, it is more appropriate to use the natural logarithm (Ln) of that value, to ensure that the range of values is more evenly distributed. As these variables were all transformed to their natural logarithm format and then used in regression analysis, an asterisk is used to make a distinction between the original and transformed format. Those variables having an asterisk are in their original format, whereas those variables carrying no asterisk are in their transformed state (see Equation (5.3) and Equation (5.4)). Other reasons for this transformation are thoroughly discussed in the next section.
\[ Y = \ln Y^* \]  
\[ X_i = \ln X_i^* \quad i = 1, 2, 3, 4 \]

Table 5-1: Terms and Definitions of Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
<th>Variable Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y^* )</td>
<td>Total Retrofit Net Construction Cost (RNCC) (thousands U.S. $)</td>
<td>Quantitative</td>
</tr>
<tr>
<td>( X_1^* )</td>
<td>Building Area (m²)</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_2^* )</td>
<td>Total Number of Stories</td>
<td>Quantitative</td>
</tr>
<tr>
<td>( X_3^* )</td>
<td>Building Age (years)</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_4^* )</td>
<td>Weight Indicator (ton/m²)</td>
<td>Quantitative</td>
</tr>
<tr>
<td>( X_5 )</td>
<td>Compliancy with first in-place seismic design code</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_6 )</td>
<td>Seismicity</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_7 )</td>
<td>Soil Type</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_8 )</td>
<td>Plan Irregularity</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_9 )</td>
<td>Vertical Irregularity</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_{10} )</td>
<td>Retrofit Measure</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_{11} )</td>
<td>Structural Type</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_{12} )</td>
<td>Diaphragm Type</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_{13} )</td>
<td>Foundation Type</td>
<td>Qualitative</td>
</tr>
<tr>
<td>( X_{14} )</td>
<td>Infill Walls Mitigation</td>
<td>Qualitative</td>
</tr>
</tbody>
</table>

Besides the variables which are quantitative in nature, qualitative or categorical variables can also be considered as predictor variables in the regression analysis. This type of variable can be represented by dummy or indicator variables taking only two values, usually 0 and 1. It should be noted that these numerical values are not intended to reflect a quantitative ordering of the categories, but only serve to identify class membership.

The number of dummy variables created was one less than the number of categories of the original qualitative variable and each case was scored 0 and 1 on each dummy variable. As can be implied from Chapter 4, scores of 1 were used to indicate the presence of an undesired category, believed to have an increasing effect on RNCC, whereas scores of 0 represented the proper category. For example, the original variable seismicity (i.e., \( X_6 \)) was converted to one dummy variable representing two categories of “very high” seismicity and “high to low” seismicity, respectively. When a case of “very high” seismicity occurred, the score of the dummy variable was 1, otherwise 0. Listed in Table 5-2 are the 10 categorical variables (\( X_5 \) through \( X_{14} \)) included in the present study. As shown in this table, the extent of
the contribution and significance of these variables on the dependent variable was studied by assigning each an appropriate dummy variable in a similar way as described for seismicity.

By convention throughout the chapter, the coefficient of an independent quantitative variable in Equation (5.2) is called a partial slope coefficient, whereas a coefficient pertaining to a qualitative variable is called a partial dummy coefficient.

**Table 5-2: Inclusion of the qualitative variables in the regression analysis**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Defined the dummy variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_5$</td>
<td>( \begin{cases} 1, \text{Buildings not complying with the provisions of first in place seismic code} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_6$</td>
<td>( \begin{cases} 1, \text{Buildings located in very high seismicity region} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_7$</td>
<td>( \begin{cases} 1, \text{Buildings sitting on poor behaved soil type} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_8$</td>
<td>( \begin{cases} 1, \text{Buildings with irregular plan configuration} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_9$</td>
<td>( \begin{cases} 1, \text{Buildings with irregular vertical configuration} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_{10}$</td>
<td>( \begin{cases} 1, \text{Buildings retrofitted through adding new steel braced frames} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_{11}$</td>
<td>( \begin{cases} 1, \text{Buildings with the absence of existing seismic resisting elements} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_{12}$</td>
<td>( \begin{cases} 1, \text{Buildings with brick archaic flooring} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_{13}$</td>
<td>( \begin{cases} 1, \text{Buildings with reinforced concrete isolated footing} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
<tr>
<td>$X_{14}$</td>
<td>( \begin{cases} 1, \text{Infill walls mitigation is considered} \ 0, \text{Otherwise} \end{cases} )</td>
</tr>
</tbody>
</table>

**5.4 Functional Form**

Linear regression analysis has been previously performed by using raw cost as the dependent variable (Emsley et al., 2002). According to Lowe et al., (2006), there are a number of assumptions implicit in this choice of variable. First, it is assumed that the standard deviation of error remains constant. That is to say, the cost of a small project can vary by the same monetary amount as a large project. This assumption is highly unlikely to be the case. Further, regression model fitting minimizes the squares of the errors, so models
developed using this technique will be inherently biased towards minimizing the errors for very large projects, where the errors are greatest. Therefore, such model is unlikely to predict well the cost of smaller projects. Given that the RNCCs of the projects for which data have been collected vary between 11,288 and 293,320 U.S.$, the influence of errors in the cost of the larger projects is several orders of magnitude greater than that of the smaller projects, so the effect will be pronounced. The second inherent assumption that is questioned is that the effects of any variables are best expressed as a fixed cost change. However, the cost of a small project would not be expected to rise by the same amount as the cost of a very large project. It is more likely that the cost would rise as a proportion of the project scope, and particularly the project size.

The criticisms discussed above indicate a serious question regarding the meaningfulness of the regression models expressed in a linear functional form when using raw data for cost and its predictors. One solution to this problem is to convert raw cost data into per-square-meter costs. However, as described in Chapter 4, cost per m² may seem to be an unrealistic cost indicator. As an alternative solution, other possible functional forms need to be examined within the models. There are several functional forms that may be used to create a regression model including, to name some, log-log (sometimes called double-log), semi-log, reciprocal, and polynomial forms. According to Studenmund (2011), the log-log form is the most common and popular functional form, that some researchers use as their default functional form instead of the linear form. In the double-log functional form, all the quantitative variables, either dependent or independent, are transformed using natural logarithm function.

Stoy et al., (2007) reported that previously double-log models have produced the best results. One possible reason behind this superiority is that the natural logarithm transformation allows for the nonlinear relationship between the dependent and independent variables to be transformed into an intrinsically linear model that can be expressed using standard linear regression (Freund et al., 2006). Consequently, this transformation is more capable of improving the predictive performance of the regression models by linearizing the data (Williams et al., 1999; Williams 2002). For instance, Williams (2003) has shown that the relationship between the dependent final project cost variable and the independent low bid variable is linear when both variables are transformed using the natural logarithm. Furthermore, this transformation was also found to be beneficial in converting a highly
skewed distribution of a given variable into a near-normal distribution (Khosrowshahi and Kaka, 1996).

Another attractive feature of the double-log model that accounts for its popularity in empirical studies is that each partial slope coefficient (i.e., $\beta_1$, $\beta_2$, ..., $\beta_p$ in Equation (5.2)) measures the partial elasticity of the dependent variable with respect to the independent quantitative variable in question, holding all other variables constant (Gujarati and Porter, 2010). Consequently, if a double-log regression is performed, the meaning of a partial slope coefficient is the percentage change in the dependent variable caused by a one percentage point increase in the independent variable while keeping the influence of other independent variables in the equation constant.

Considering all the justifications given above, the present study, in addition to the linear functional form, utilised the greatest benefits of the features of the double-log functional form when developing MLR models. This approach makes it possible to generate multiple regression models such that the dependent variable is in natural logarithm form and some of the X variables ($X_1^*$ through $X_4^*$), i.e., the numerical variables, are also transformed to appear in the same form, whereas other variables ($X_5$ through $X_{14}$), i.e., the categorical variables, enter the equations in their linear form.

5.5 Development and Testing Datasets

The data from a total number of 158 earthquake-prone schools with a framed structure (i.e., steel and concrete structures) were used to develop and test MLR equations. For each school, the respective data point contains the information of both dependent and independent variables considered in this research. As a normal practice, the available database was randomly partitioned into two datasets: (1) the development dataset and (2) the hold-out or test dataset. The former was mainly used to develop the regression models, whereas the latter was employed to compare the predictive performance of the models on new data never used in their development.

Generally, the development dataset needs to be populated with enough data to ensure the appropriate development of solid models. Likewise, an adequate size for the test dataset is also required to run a meaningful test on the predictive performance of the models. Hence, almost 75% of all the cases (121 schools) were randomly separated out to build the
development dataset, while the remaining 25% (37 schools) were held to form the independent test dataset. These percentages were backed by some researchers to be reasonable (Abu Hammad et al., 2010; Elhag and Boussebaine, 1999; Williams, 1994; Williams et al., 2005).

5.6 Development of MLR Models

In the multiple regression analysis, the identification of the best combination of the independent variables is of paramount importance as different combinations produce different results. In many applications of regression analysis the set of variables to be included in the regression model is not predetermined, and often the first part of the analysis involves the selection of these variables. There are some occasions when theoretical or other considerations determine the variables to be included in the equation. In those situations the issue of variable selection does not arise. But in situations where there is no clear-cut theory, the problem of selecting variables for a regression model becomes an important one. However, the possibilities for this selection are numerous and the coverage of all cases is impractical. Hence, the question to be answered while formulating a regression model is which variables should be included in a regression model. The form of variables is another issue of concern which is already discussed in Section 5.4.

To address the question above, particularly for cases when there are a large number of potential predictor variables, there are a number of techniques which have been proposed as the next best alternative to the best solution (Khosrowshahi and Kaka, 1996). These procedures do not involve computation of all possible equations and have the feature that the variables are introduced or deleted from the equation one at a time. With p variables, these procedures involve evaluation of at most \((p+1)\) equations, as contrasted with the evaluation of \(2^p\) equations necessary for examining all possible equations (Chatterjee and Hadi, 2006). These procedures can be classified into two broad categories: the forward selection procedure (FS) and the backward elimination procedure (BE) (Attalla and Hegazy, 2003; Chatterjee and Hadi, 2006; Emsley et al., 2002; Lowe et al., 2006). There is also a popular modification of the FS procedure called the stepwise procedure (Chatterjee and Hadi, 2006).

The FS procedure begins with no variables in the equation, with the most significant predictor entered at the first step and the independent variables continuing to be added until
none can significantly improve the fit. As reported by Lowe et al., (2006), one problem with this technique is that a variable that correlates well with a number of cost significant variables may be included ahead of other significant variables, because it appears to encapsulate these variables. If this encapsulating variable has a higher significance than the individual variables themselves then the variable will be included first. When other variables are considered for addition to the model, some of the information contained in these variables will already be present in the model, which will make these variables appear less significant than they really are. One possible way of avoiding this problem is to perform a BE procedure. In contrast to the FS procedure, the BE procedure starts with the full equation and successively drops one variable at a time. In this procedure, a variable is deleted according to its contribution to definition of the dependent variable. At each step, a variable that does not pass the tolerance test and therefore has the minimal t-value is deleted first. Then, a tolerance test is again given to all the remaining variables. A variable with the least t-value which does not pass the tolerance is selected as the second deleted variable. This procedure continues until either all variables included in the model have a t-value greater than the threshold value or all variables are deleted. Although FS and BE procedures will normally terminate with identical models (Chatterjee and Hadi, 2006), it is reported in the literature that the latter is preferable over the former because the BE procedure has the advantage of looking at all the available variables in the early stages of the model development process (Attalla and Hegazy, 2003). On the whole, the BE procedure yields models with more variables than the FS procedure, which means that by using the backward elimination procedure it is possible to extract more significant variables than when using the forward selection procedure (Lowe et al., 2006). This possibility was also supported by the results of this study.

Upon running the BE procedure on the fourteen variables, seven variables were automatically eliminated due to their insignificant contribution to the model (p>0.05). The resulting MLR model consisting of seven independent variables is given by:

\[ Y = -1.238 + 0.674X_1 + 0.273X_2 + 0.195X_4 + 0.218X_6 + 0.117X_7 + 0.118X_8 + 0.132X_{11} \]  \hspace{0.5cm} (5.5)

\[ t-value = (-2.840) \hspace{0.5cm} (10.286) \hspace{0.5cm} (2.675) \hspace{0.5cm} (2.031) \hspace{0.5cm} (2.966) \hspace{0.5cm} (2.182) \hspace{0.5cm} (1.986) \hspace{0.5cm} (2.615) \]

\[ p-value = (0.005) \hspace{0.5cm} (0.000) \hspace{0.5cm} (0.009) \hspace{0.5cm} (0.045) \hspace{0.5cm} (0.004) \hspace{0.5cm} (0.031) \hspace{0.5cm} (0.049) \hspace{0.5cm} (0.010) \]

\[ R^2 = 0.710 \hspace{0.5cm} Adjusted R^2 = 0.692 \hspace{0.5cm} F = 39.512 \]
The model above incorporates as many independent variables as are found to have an appreciable influence on the value of the dependent variable. According to Equation (5.5), the $R^2$ value of the resulting MLR model is 0.71, indicating that 71% of the variation in the dependent variable is explained by the variation in the model parameters. Also, the F value of the resulting model (i.e., 39.512) and low probabilities of the independent variables (i.e., $p<0.05$) indicate that this model and its contributing variables are all statistically significant. This finding is due to the fact that when $p$-values of the partial slope and dummy coefficients are less than the common 5% significance level, the null hypotheses that these coefficients are individually equal to zero (i.e., $\beta_s=0$) is rejected and the alternative hypothesis holds. In addition, the computed F value is also significantly large to simply reject the null hypothesis that these coefficients are jointly equal to zero (i.e., $\beta_1=\beta_2=\ldots=\beta_p=0$) even at the lower significance level ($p<0.01$).

In comparison with the BE procedure, both the FS and stepwise procedures resulted in a model having a lower number of significant variables ($p<0.05$). This model is a function of five variables, as shown in the following equation:

$$Y = -1.085 + 0.643X_1 + 0.368X_2 + 0.235X_6 + 0.118X_7 + 0.131X_{11}$$

$$t\text{-value} = (-2.485) \quad (9.945) \quad (3.886) \quad (3.137) \quad (2.158) \quad (2.542)$$

$$p\text{-value} = (0.014) \quad (0.000) \quad (0.000) \quad (0.002) \quad (0.033) \quad (0.012)$$

$$R^2 = 0.692 \quad \text{Adjusted } R^2 = 0.678 \quad F = 51.563$$

The variable selection procedures discussed above should be used with caution, and should not be used mechanically to determine the best variables. Only a total and comprehensive analysis of all the models that can be developed by employing either of the BE, FS and stepwise procedures can lead to a reliable selection of variables. This issue is scrutinised at length in Section 5.6.1. To enable the BE procedure to run through the whole set of variables (i.e., starting with the equation containing the full set of fourteen independent variables and ending up with the equation containing only the most significant independent variable), the cut-off value for the $p$-value was set extremely low ($p<0.0001$). The results of applying the BE procedure are presented in Table 5-3. To distinguish models in Table 5-3 having a different number of independent variables, the name of each model is accompanied
by the number of independent variables present in that particular equation (see Table 5-4). So, MLR models 1 to 14 are the models taking one to fourteen independent variables into account, respectively. The main statistics of the developed regression models are also summarized in Table 5-5. Noteworthy, both the FS and stepwise methods resulted in exactly the same models as did the BE method. As a final remark, it should be pointed out that all of these methods are problematic as they are somewhat arbitrary and restrictive. One approach to mitigation of this problem is to utilise the “best subset regression” (BSR) method. In comparison to the BE, FS, and stepwise methods, the main advantage of the BSR method is that it allows for the assessment of all possible regression models. As a result, the end outcome is a number of best-fitting models that accommodate one predictor, two predictors, and so on. Although being advantageous, the BSR method was not performed in the present study due to software availability constraints. The computer program utilised here to perform regression analysis was the SPSS statistical package (PASW Statistics 18, Release 18.0.0, SPSS Inc., 2009) that does not support the BSR method.
Table 5-3: Regression models developed using the BE procedure

<table>
<thead>
<tr>
<th>Variables in model</th>
<th>Excluded variable</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X14</td>
<td></td>
<td>$Y = -1.300 + 0.702X_1 + 0.257X_2 + 0.214X_4 + 0.190X_9 + 0.113X_7 + 0.139X_{11} + 0.054X_{13} + 0.001X_{15} + 0.071X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X14</td>
<td>X13</td>
<td>$Y = -1.301 + 0.702X_1 + 0.257X_2 + 0.214X_4 + 0.190X_9 + 0.113X_7 + 0.139X_{11} + 0.054X_{13} + 0.071X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X14</td>
<td>X10</td>
<td>$Y = -1.286 + 0.703X_1 + 0.255X_2 + 0.214X_4 + 0.041X_5 + 0.185X_7 + 0.142X_{9} + 0.140X_{11} + 0.057X_{13} + 0.072X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X11, X12, X14</td>
<td>X9</td>
<td>$Y = -1.299 + 0.705X_1 + 0.251X_2 + 0.215X_4 + 0.038X_5 + 0.185X_7 + 0.134X_{9} + 0.057X_{13} + 0.070X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X11, X12, X14</td>
<td>X8</td>
<td>$Y = -1.397 + 0.709X_1 + 0.249X_2 + 0.214X_4 + 0.188X_7 + 0.129X_{9} + 0.053X_{13} + 0.067X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X11, X12, X14</td>
<td>X9</td>
<td>$Y = -1.460 + 0.706X_1 + 0.251X_2 + 0.210X_4 + 0.187X_7 + 0.123X_{9} + 0.047X_{13} + 0.069X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X11, X12, X14</td>
<td>X8</td>
<td>$Y = -1.339 + 0.689X_1 + 0.256X_2 + 0.200X_4 + 0.191X_7 + 0.125X_{9} + 0.042X_{13} + 0.067X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X11, X12, X14</td>
<td>X9</td>
<td>$Y = -1.238 + 0.674X_1 + 0.273X_2 + 0.218X_7 + 0.178X_{9} + 0.118X_{13} + 0.132X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X6, X7, X11</td>
<td>X8</td>
<td>$Y = -1.242 + 0.674X_1 + 0.293X_2 + 0.172X_7 + 0.233X_{9} + 0.129X_{13} + 0.118X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X6, X7, X11</td>
<td>X4</td>
<td>$Y = -1.085 + 0.643X_1 + 0.368X_7 + 0.235X_{9} + 0.118X_{13} + 0.131X_{14}$</td>
</tr>
<tr>
<td>X1, X2, X3, X6, X11</td>
<td>X7</td>
<td>$Y = -1.025 + 0.648X_1 + 0.362X_7 + 0.226X_{9} + 0.136X_{13}$</td>
</tr>
<tr>
<td>X1, X2, X6</td>
<td>X11</td>
<td>$Y = -0.763 + 0.617X_1 + 0.378X_7 + 0.246X_{9}$</td>
</tr>
<tr>
<td>X1, X2</td>
<td>X6</td>
<td>$Y = -0.621 + 0.627X_1 + 0.376X_{9}$</td>
</tr>
<tr>
<td>X1</td>
<td>X2</td>
<td>$Y = -1.273 + 0.771X_1$</td>
</tr>
</tbody>
</table>
Table 5-4: Numbering of the regression models developed using the BE procedure

<table>
<thead>
<tr>
<th>Variables in model</th>
<th>No. of Independent variables in each model</th>
<th>Number assigned to each model</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X14</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X14</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X14</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>X1, X2, X3, X4, X5, X6, X7, X8, X11, X12, X14</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>X1, X2, X4, X6, X7, X8, X11, X12, X14</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>X1, X2, X4, X6, X7, X8, X11, X12, X14</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>X1, X2, X4, X6, X7, X8, X11</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>X1, X2, X4, X6, X7, X11</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>X1, X2, X6, X7, X11</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>X1, X2, X6, X11</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>X1, X2, X6</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>X1, X2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>X1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5-5: Summary of the key statistics of the regression models developed using the BE procedure

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Coefficient of determination (R²)</th>
<th>Adjusted coefficient of determination (Rₐ²)</th>
<th>Residual sum of squares (SSE)</th>
<th>Std. Error of Estimate (σ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.588</td>
<td>0.585</td>
<td>11.731</td>
<td>0.314</td>
</tr>
<tr>
<td>2</td>
<td>0.631</td>
<td>0.625</td>
<td>10.519</td>
<td>0.299</td>
</tr>
<tr>
<td>3</td>
<td>0.660</td>
<td>0.652</td>
<td>9.683</td>
<td>0.288</td>
</tr>
<tr>
<td>4</td>
<td>0.679</td>
<td>0.668</td>
<td>9.148</td>
<td>0.281</td>
</tr>
<tr>
<td>5</td>
<td>0.692</td>
<td>0.678</td>
<td>8.792</td>
<td>0.277</td>
</tr>
<tr>
<td>6</td>
<td>0.700</td>
<td>0.684</td>
<td>8.556</td>
<td>0.274</td>
</tr>
<tr>
<td>7</td>
<td>0.710</td>
<td>0.692</td>
<td>8.267</td>
<td>0.270</td>
</tr>
<tr>
<td>8</td>
<td>0.714</td>
<td>0.692</td>
<td>8.156</td>
<td>0.270</td>
</tr>
<tr>
<td>9</td>
<td>0.715</td>
<td>0.692</td>
<td>8.116</td>
<td>0.270</td>
</tr>
<tr>
<td>10</td>
<td>0.716</td>
<td>0.690</td>
<td>8.095</td>
<td>0.271</td>
</tr>
<tr>
<td>11</td>
<td>0.716</td>
<td>0.688</td>
<td>8.081</td>
<td>0.272</td>
</tr>
<tr>
<td>12</td>
<td>0.717</td>
<td>0.685</td>
<td>8.078</td>
<td>0.274</td>
</tr>
<tr>
<td>13</td>
<td>0.717</td>
<td>0.682</td>
<td>8.076</td>
<td>0.275</td>
</tr>
<tr>
<td>14</td>
<td>0.717</td>
<td>0.679</td>
<td>8.076</td>
<td>0.276</td>
</tr>
</tbody>
</table>
5.6.1 Criteria for Selecting the Most Appropriate MLR Model

The models listed in Table 5-3 are different from each other in some aspects, including the number and type of variables incorporating in these models. To reasonably select the most appropriate model, the adequacy of various fitted regression models needs to be judged in accordance with some reference criteria. In the present research, different criteria that were considered to be most useful were examined and results are provided in the following sections.

5.6.1.1 Leave-one-out Cross Validation

The leave-one-out cross validation technique is essentially a non-parametric MLR analysis, which allows the user to simulate the way in which predictions are produced in practice (Cheung and Skitmore, 2006). Although being a time-consuming method (Arlot and Celisse, 2010; Smith and Mason, 1997), the suitability of the leave-one-out cross validation technique for evaluating regression models has been demonstrated by a number of previous studies (Cheung and Skitmore, 2006; Skitmore and Patchell, 1990). This technique is generally regarded as the ultimate version of the k-fold cross validation technique with $k$ being equal to the number of cases in an original dataset. According to Cheung and Skitmore (2006), the accuracy of statistical inference in the leave-one-out cross validation technique is preserved by dividing a dataset that contains $n$ cases of data into $n$ exploratory sub-samples and $n$ omitted (i.e., hold-out) cases. In the present study, $n$ is equal to 121. Each exploratory sub-sample is used to select a statistical model using the least-squares approach and contains $n - 1$ cases (i.e., 120 cases in this study) that are obtained from the original n-case dataset by the omission of one case without repetition. In addition, each omitted case is used to validate the selected model from an exploratory sub-sample that does not contain the omitted case.

The results obtained from implementation of the leave-one-out cross validation technique are displayed in Table 5-6. In this table, the mean square error (MSE) and the coefficient of determination ($R^2$) were calculated from $n$ models (i.e., $n = 121$) examined for each subset of candidates. A comparison of these statistics reveals that model number seven is the best candidate for predicting RNCC, as this model resulted in both the smallest MSE and the largest $R^2$. According to this table, the values of MSE and $R^2$ are respectively decreasing and increasing from model 1 to 7, indicating that the greater the number of
statistically significant variables in a model, the greater the predictive ability of the model. Nevertheless, the values of MSE and $R^2$ in other models show a completely reverse trend. These values are respectively increasing and decreasing from model 7 to 14, indicating that the predictive ability of these models is continually diminishing as the number of statistically insignificant variables increases in their formulation.

Table 5-6: Results of the leave-one-out cross validation technique for all candidate models

<table>
<thead>
<tr>
<th>No. of model</th>
<th>MSE</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1033</td>
<td>0.5614</td>
</tr>
<tr>
<td>2</td>
<td>0.0942</td>
<td>0.5998</td>
</tr>
<tr>
<td>3</td>
<td>0.0885</td>
<td>0.6242</td>
</tr>
<tr>
<td>4</td>
<td>0.0846</td>
<td>0.6408</td>
</tr>
<tr>
<td>5*</td>
<td>0.0837</td>
<td>0.6446</td>
</tr>
<tr>
<td>6</td>
<td>0.0834</td>
<td>0.6457</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td><strong>0.0826</strong></td>
<td><strong>0.6491</strong></td>
</tr>
<tr>
<td>8</td>
<td>0.0829</td>
<td>0.6477</td>
</tr>
<tr>
<td>9</td>
<td>0.0840</td>
<td>0.6430</td>
</tr>
<tr>
<td>10</td>
<td>0.0858</td>
<td>0.6355</td>
</tr>
<tr>
<td>11</td>
<td>0.0867</td>
<td>0.6316</td>
</tr>
<tr>
<td>12</td>
<td>0.0880</td>
<td>0.6263</td>
</tr>
<tr>
<td>13</td>
<td>0.0901</td>
<td>0.6174</td>
</tr>
<tr>
<td>14</td>
<td>0.0914</td>
<td>0.6117</td>
</tr>
</tbody>
</table>

Note: Shaded is the best model, as indicated by the MSE and $R^2$ values.
* Model 5 is the model derived from implementation of the forward selection and the stepwise techniques.
** Model 7 is the model derived from implementation of the backward elimination technique.

The predictors included in model number seven are $X_1$, $X_2$, $X_4$, $X_6$, $X_7$, $X_8$, and $X_{11}$. To illustrate the cross validated results, Table 5-7 presents the regression coefficients for each of these predictors, together with predicted Ys, MSE and $R^2$ determined by the leave-one-out cross validation technique for model number seven.
Table 5-7: Coefficients, predictions, MSE and $R^2$ values determined by implementing the leave-one-out cross validation technique for “model number seven”

<table>
<thead>
<tr>
<th>Case</th>
<th>$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10}$</th>
<th>$Y$</th>
<th>Square Error $(Y_A - Y_P)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.219 0.60 0.28 0.19 0.22 0.12 0.11 0.12 0.126</td>
<td>4.21</td>
<td>4.49 0.0757</td>
</tr>
<tr>
<td>2</td>
<td>1.348 0.696 0.246 0.219 0.102 0.141 0.125</td>
<td>4.42</td>
<td>4.96 0.0875</td>
</tr>
<tr>
<td>3</td>
<td>1.197 0.666 0.218 0.122 0.11 0.134</td>
<td>4.78</td>
<td>4.60 0.0318</td>
</tr>
<tr>
<td>4</td>
<td>1.256 0.678 0.187 0.215 0.114</td>
<td>4.84</td>
<td>4.66 0.0334</td>
</tr>
<tr>
<td>5</td>
<td>1.232 0.673 0.189 0.214 0.113</td>
<td>4.8</td>
<td>4.50 0.0075</td>
</tr>
<tr>
<td>6</td>
<td>1.236 0.674 0.194 0.217 0.116</td>
<td>4.6</td>
<td>4.54 0.0004</td>
</tr>
<tr>
<td>7</td>
<td>1.233 0.672 0.192 0.217 0.12</td>
<td>4.57</td>
<td>4.46 0.0019</td>
</tr>
<tr>
<td>8</td>
<td>1.199 0.67 0.215 0.112 0.104</td>
<td>5.37</td>
<td>4.96 0.1718</td>
</tr>
<tr>
<td>9</td>
<td>1.264 0.678 0.216 0.114 0.11</td>
<td>4.54</td>
<td>4.34 0.0395</td>
</tr>
<tr>
<td>10</td>
<td>1.264 0.678 0.216 0.114 0.11</td>
<td>4.54</td>
<td>4.34 0.0395</td>
</tr>
<tr>
<td>11</td>
<td>1.252 0.676 0.195 0.116 0.14</td>
<td>5.29</td>
<td>5.19 0.0097</td>
</tr>
<tr>
<td>12</td>
<td>1.247 0.675 0.194 0.117 0.131</td>
<td>4.73</td>
<td>4.76 0.0008</td>
</tr>
<tr>
<td>13</td>
<td>1.247 0.675 0.194 0.117 0.131</td>
<td>4.84</td>
<td>4.90 0.0039</td>
</tr>
<tr>
<td>14</td>
<td>1.238 0.675 0.194 0.117 0.132</td>
<td>4.78</td>
<td>4.79 0.0002</td>
</tr>
<tr>
<td>15</td>
<td>1.235 0.672 0.194 0.126 0.122</td>
<td>4.76</td>
<td>4.39 0.1368</td>
</tr>
<tr>
<td>16</td>
<td>1.241 0.675 0.196 0.117 0.132</td>
<td>4.43</td>
<td>4.41 0.0094</td>
</tr>
<tr>
<td>17</td>
<td>1.288 0.682 0.199 0.124 0.109</td>
<td>4.61</td>
<td>4.35 0.0651</td>
</tr>
<tr>
<td>18</td>
<td>1.224 0.676 0.202 0.115 0.129</td>
<td>4.47</td>
<td>4.69 0.0475</td>
</tr>
<tr>
<td>19</td>
<td>1.290 0.685 0.190 0.113 0.125</td>
<td>4.12</td>
<td>4.48 0.1319</td>
</tr>
<tr>
<td>20</td>
<td>1.234 0.674 0.218 0.116 0.132</td>
<td>5.01</td>
<td>4.98 0.0011</td>
</tr>
<tr>
<td>21</td>
<td>1.22 0.671 0.269 0.127 0.137</td>
<td>5.06</td>
<td>4.67 0.1520</td>
</tr>
<tr>
<td>22</td>
<td>1.248 0.677 0.195 0.115 0.134</td>
<td>4.82</td>
<td>4.68 0.0187</td>
</tr>
<tr>
<td>23</td>
<td>1.214 0.666 0.182 0.122 0.134</td>
<td>3.85</td>
<td>4.24 0.1521</td>
</tr>
<tr>
<td>24</td>
<td>1.248 0.676 0.191 0.116 0.13</td>
<td>4.32</td>
<td>4.40 0.0059</td>
</tr>
<tr>
<td>25</td>
<td>1.235 0.674 0.197 0.116 0.132</td>
<td>4.48</td>
<td>4.42 0.0030</td>
</tr>
<tr>
<td>26</td>
<td>1.231 0.674 0.195 0.118 0.131</td>
<td>4.29</td>
<td>4.37 0.0058</td>
</tr>
<tr>
<td>27</td>
<td>1.087 0.659 0.233 0.172 0.124</td>
<td>3.44</td>
<td>4.15 0.5101</td>
</tr>
<tr>
<td>28</td>
<td>1.301 0.678 0.271 0.127 0.141</td>
<td>4.75</td>
<td>4.10 0.4284</td>
</tr>
<tr>
<td>29</td>
<td>1.273 0.685 0.185 0.098 0.122</td>
<td>4.01</td>
<td>4.72 0.5110</td>
</tr>
<tr>
<td>30</td>
<td>1.231 0.674 0.273 0.116 0.119</td>
<td>4.71</td>
<td>4.64 0.0043</td>
</tr>
<tr>
<td>31</td>
<td>1.255 0.678 0.265 0.115 0.121</td>
<td>4.65</td>
<td>4.49 0.0250</td>
</tr>
<tr>
<td>32</td>
<td>1.234 0.672 0.197 0.121 0.112</td>
<td>4.53</td>
<td>4.40 0.0181</td>
</tr>
<tr>
<td>33</td>
<td>1.259 0.679 0.192 0.115 0.128</td>
<td>3.93</td>
<td>4.20 0.0731</td>
</tr>
<tr>
<td>34</td>
<td>1.229 0.673 0.194 0.114 0.12</td>
<td>4.74</td>
<td>4.55 0.0367</td>
</tr>
<tr>
<td>35</td>
<td>1.245 0.672 0.182 0.111 0.122</td>
<td>4.6</td>
<td>4.12 0.2316</td>
</tr>
<tr>
<td>36</td>
<td>1.186 0.666 0.191 0.122 0.125</td>
<td>4.09</td>
<td>4.31 0.0463</td>
</tr>
<tr>
<td>37</td>
<td>1.305 0.682 0.202 0.113 0.12</td>
<td>4.28</td>
<td>3.93 0.1192</td>
</tr>
<tr>
<td>38</td>
<td>1.232 0.674 0.196 0.116 0.132</td>
<td>4.96</td>
<td>4.92 0.0019</td>
</tr>
<tr>
<td>39</td>
<td>1.235 0.675 0.217 0.115 0.141</td>
<td>4.97</td>
<td>4.85 0.0136</td>
</tr>
<tr>
<td>40</td>
<td>1.312 0.685 0.265 0.123 0.109</td>
<td>4.21</td>
<td>3.96 0.0615</td>
</tr>
<tr>
<td>41</td>
<td>1.326 0.691 0.243 0.222 0.121</td>
<td>4.27</td>
<td>4.64 0.1357</td>
</tr>
</tbody>
</table>

Note: $Y_A$ and $Y_P$ are the actual and predicted values, respectively.
Table 5-7: Coefficients, predictions, MSE and R2 values determined by implementing the
leave-one-out cross validation technique for “model number seven” (Continued)
Case
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100

Y = β0 + β1.X1 + β2.X2 + β4.X4 + β6.X6 + β7.X7 + β8.X8 + β11.X11
β0
β1
β2
β4
β6
β7
β8
β11
-1.207
-1.248
-1.237
-1.234
-1.243
-1.243
-1.243
-1.241
-1.234
-1.238
-1.241
-1.247
-1.298
-1.264
-1.238
-1.223
-1.226
-1.236
-1.24
-1.23
-1.242
-1.233
-1.163
-1.243
-1.148
-1.245
-1.233
-1.253
-1.239
-1.216
-1.242
-1.256
-1.239
-1.268
-1.18
-1.24
-1.24
-1.255
-1.164
-1.24
-1.909
-1.235
-1.246
-1.199
-1.251
-1.229
-1.249
-1.103
-1.213
-1.265

0.671
0.676
0.674
0.674
0.676
0.676
0.676
0.676
0.674
0.675
0.675
0.676
0.68
0.682
0.675
0.671
0.671
0.674
0.675
0.675
0.675
0.673
0.664
0.675
0.661
0.676
0.674
0.674
0.675
0.673
0.675
0.678
0.675
0.678
0.664
0.675
0.675
0.677
0.668
0.675
0.769
0.674
0.675
0.669
0.676
0.67
0.676
0.655
0.671
0.68

0.272
0.272
0.274
0.273
0.266
0.266
0.266
0.267
0.274
0.273
0.275
0.289
0.287
0.259
0.279
0.283
0.282
0.273
0.273
0.267
0.273
0.274
0.273
0.275
0.285
0.269
0.276
0.281
0.273
0.266
0.273
0.266
0.273
0.278
0.291
0.272
0.272
0.272
0.255
0.272
0.232
0.276
0.28
0.277
0.278
0.286
0.277
0.277
0.277
0.261

0.189
0.195
0.195
0.196
0.199
0.199
0.199
0.198
0.195
0.196
0.196
0.179
0.178
0.2
0.192
0.191
0.191
0.197
0.196
0.195
0.197
0.2
0.197
0.197
0.194
0.196
0.194
0.193
0.195
0.198
0.196
0.197
0.196
0.193
0.192
0.196
0.196
0.193
0.244
0.194
0.203
0.19
0.195
0.199
0.196
0.202
0.196
0.167
0.195
0.203

0.214
0.217
0.217
0.219
0.22
0.22
0.22
0.22
0.218
0.217
0.219
0.225
0.221
0.203
0.203
0.221
0.22
0.219
0.219
0.217
0.217
0.216
0.22
0.217
0.216
0.219
0.213
0.231
0.218
0.218
0.218
0.216
0.218
0.216
0.222
0.218
0.218
0.218
0.217
0.218
0.202
0.218
0.219
0.216
0.219
0.223
0.219
0.226
0.218
0.219

0.113
0.116
0.116
0.115
0.119
0.119
0.119
0.119
0.117
0.116
0.115
0.101
0.121
0.119
0.119
0.12
0.119
0.117
0.118
0.115
0.121
0.125
0.12
0.116
0.113
0.115
0.114
0.115
0.117
0.114
0.117
0.115
0.117
0.114
0.121
0.118
0.118
0.118
0.112
0.116
0.147
0.119
0.112
0.114
0.113
0.123
0.113
0.125
0.115
0.119

0.122
0.118
0.119
0.117
0.117
0.117
0.117
0.117
0.118
0.119
0.117
0.109
0.112
0.117
0.116
0.115
0.115
0.117
0.117
0.12
0.12
0.121
0.127
0.119
0.122
0.117
0.117
0.118
0.118
0.11
0.118
0.12
0.118
0.12
0.114
0.118
0.118
0.122
0.104
0.115
0.125
0.118
0.116
0.122
0.116
0.112
0.116
0.11
0.113
0.117

0.137
0.133
0.133
0.131
0.129
0.129
0.129
0.129
0.131
0.133
0.13
0.123
0.139
0.137
0.137
0.135
0.135
0.132
0.133
0.129
0.129
0.126
0.134
0.131
0.125
0.133
0.134
0.128
0.132
0.127
0.132
0.129
0.132
0.129
0.135
0.131
0.131
0.134
0.12
0.131
0.126
0.131
0.135
0.126
0.134
0.139
0.134
0.139
0.129
0.135

YA

Y
YP

Square Error
(YA – YP)2

5.16
4.12
4.57
4.21
4.12
4.12
4.12
4.09
4.38
4.54
4.59
4.07
4.88
4.04
4.34
4.37
4.4
4.37
4.45
5.1
4.53
4.72
3.53
4.12
5.29
4.2
4.29
4.07
4.96
5.38
5.28
4.68
4.59
4.15
4.07
4.59
4.59
4.79
5.53
4.84
4.07
4.53
4.71
5.01
4.81
4.17
4.81
3.21
5.06
4.32

4.85
4.07
4.49
4.28
4.24
4.24
4.24
4.23
4.41
4.48
4.67
4.67
5.21
4.27
4.56
4.60
4.61
4.42
4.54
4.95
4.36
4.43
3.77
4.05
4.95
4.27
4.37
3.87
4.97
5.15
5.30
4.51
4.58
3.93
4.39
4.55
4.55
4.90
5.02
4.76
2.90
4.46
4.88
4.72
4.96
4.68
4.95
3.87
4.92
4.49

0.0949
0.0026
0.0062
0.0053
0.0154
0.0154
0.0154
0.0191
0.0008
0.0039
0.0072
0.3646
0.1116
0.0513
0.0501
0.0539
0.0462
0.0025
0.0078
0.0229
0.0280
0.0831
0.0555
0.0047
0.1162
0.0049
0.0070
0.0392
0.0001
0.0550
0.0002
0.0302
0.0001
0.0483
0.1017
0.0018
0.0018
0.0120
0.2640
0.0066
1.3799
0.0045
0.0290
0.0834
0.0212
0.2651
0.0188
0.4301
0.0208
0.0300

132 | P a g e


Table 5-7: Coefficients, predictions, MSE and $R^2$ values determined by implementing the leave-one-out cross validation technique for “model number seven” (Continued)

<table>
<thead>
<tr>
<th>Case</th>
<th>Y = $\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_4X_4 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8 + \beta_{11}X_{11}$</th>
<th>Y</th>
<th>Square Error $(Y_A - Y_P)^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>-1.252 0.677 0.272 0.195 0.219 0.118 0.117 0.134</td>
<td>4.66 4.77</td>
<td>0.0120</td>
</tr>
<tr>
<td>102</td>
<td>-1.258 0.677 0.27 0.196 0.219 0.119 0.126 0.136</td>
<td>4.54 4.78</td>
<td>0.0564</td>
</tr>
<tr>
<td>103</td>
<td>-1.124 0.675 0.273 0.196 0.218 0.117 0.117 0.132</td>
<td>4.53 4.58</td>
<td>0.0021</td>
</tr>
<tr>
<td>104</td>
<td>-1.267 0.674 0.285 0.182 0.235 0.124 0.117 0.127</td>
<td>3.96 3.70</td>
<td>0.0661</td>
</tr>
<tr>
<td>105</td>
<td>-1.257 0.677 0.277 0.197 0.219 0.112 0.116 0.135</td>
<td>4.83 5.00</td>
<td>0.0302</td>
</tr>
<tr>
<td>106</td>
<td>-1.219 0.671 0.277 0.196 0.218 0.12 0.114 0.129</td>
<td>4.93 4.81</td>
<td>0.0133</td>
</tr>
<tr>
<td>107</td>
<td>-1.17 0.664 0.284 0.189 0.216 0.114 0.121 0.126</td>
<td>5.2 4.94</td>
<td>0.0698</td>
</tr>
<tr>
<td>108</td>
<td>-1.27 0.683 0.251 0.209 0.22 0.108 0.116 0.138</td>
<td>3.97 4.31</td>
<td>0.1130</td>
</tr>
<tr>
<td>109</td>
<td>-1.242 0.675 0.275 0.197 0.219 0.118 0.117 0.133</td>
<td>4.71 4.79</td>
<td>0.0072</td>
</tr>
<tr>
<td>110</td>
<td>-1.239 0.675 0.274 0.195 0.218 0.117 0.118 0.132</td>
<td>4.99 5.01</td>
<td>0.0003</td>
</tr>
<tr>
<td>111</td>
<td>-1.227 0.67 0.297 0.197 0.216 0.114 0.12 0.128</td>
<td>3.97 3.74</td>
<td>0.0547</td>
</tr>
<tr>
<td>112</td>
<td>-1.253 0.677 0.273 0.19 0.215 0.113 0.121 0.127</td>
<td>4.7 4.39</td>
<td>0.0954</td>
</tr>
<tr>
<td>113</td>
<td>-1.285 0.679 0.269 0.198 0.234 0.124 0.119 0.127</td>
<td>4.13 3.88</td>
<td>0.0613</td>
</tr>
<tr>
<td>114</td>
<td>-1.267 0.678 0.27 0.197 0.216 0.123 0.12 0.128</td>
<td>4.4 4.17</td>
<td>0.0526</td>
</tr>
<tr>
<td>115</td>
<td>-1.215 0.668 0.288 0.192 0.22 0.122 0.133 0.139</td>
<td>4.31 4.76</td>
<td>0.2006</td>
</tr>
<tr>
<td>116</td>
<td>-1.244 0.675 0.273 0.195 0.218 0.116 0.118 0.131</td>
<td>4.14 4.11</td>
<td>0.0011</td>
</tr>
<tr>
<td>117</td>
<td>-1.124 0.675 0.274 0.194 0.216 0.115 0.12 0.129</td>
<td>4.64 4.46</td>
<td>0.0315</td>
</tr>
<tr>
<td>118</td>
<td>-1.05 0.647 0.282 0.188 0.225 0.123 0.113 0.136</td>
<td>3.07 3.58</td>
<td>0.2564</td>
</tr>
<tr>
<td>119</td>
<td>-0.938 0.636 0.253 0.195 0.224 0.123 0.142 0.134</td>
<td>2.42 2.99</td>
<td>0.3269</td>
</tr>
<tr>
<td>120</td>
<td>-1.225 0.673 0.275 0.197 0.218 0.116 0.116 0.13</td>
<td>5.04 4.98</td>
<td>0.0038</td>
</tr>
<tr>
<td>121</td>
<td>-1.207 0.671 0.274 0.204 0.214 0.112 0.123 0.124</td>
<td>5.03 4.61</td>
<td>0.1762</td>
</tr>
</tbody>
</table>

$\text{MSE} = 0.0826$

$\text{R}^2 = 0.6491$

$Y_A$: Actual $Y$

$Y_P$: Predicted $Y$

5.6.1.2 Minimum Absolute t-value

One of the main issues with the BE procedure is that it searches a large space of possible models. Hence it is exposed to over fitting the data (Chatterjee and Hadi, 2006). This problem can be mitigated with a well-established stringent criterion for deleting a variable in a successive equation and, finally, selecting the most appropriate equation.

In the present research the following stopping rule suggested by Chatterjee and Hadi (2006) is accepted to select the target equation of interest by which the functional relationships between RNCC and independent variables can be better described when compared with other equations developed through the same BE procedure.

\[
\text{Minimum absolute } t - \text{value of an equation } (\min|t|) > t_{0.05}(n - p) \quad (5.7)
\]
where \( n \) is the number of observations in the development dataset (i.e., \( n=121 \)) and \( p \) is the number of terms in an equation (including a constant and \( (p-1) \) independent variables). \( t_{0.05} (n-p) \) represents the \( t \) value for \( (n-p) \) degrees of freedom (D.F.) at the 5% level of significance. The min \( t \)-value and corresponding cut-off value (\( t_{0.05} (n-p) \)) of each equation are presented in Table 5-8.

The variables selected by using the above stopping rule are \( X_1, X_2, X_4, X_6, X_7, X_8, \) and \( X_{11} \) which, in turn, implies that model number seven can be regarded as the most suitable MLR model for the purpose of predicting RNCC.

### Table 5-8: Models minimum \( |t| \) and corresponding \( t_{0.05} (n-p) \) values

| No. of model | No. of observations in the development dataset (n) | No. of terms incorporating in a model (p) | Min \( |t| \) | \( t_{0.05} (n-p) \) | Selected model based on stopping rule |
|--------------|--------------------------------------------------|-----------------------------------------|--------------|-----------------------------|--------------------------------------|
| 14           | 121                                              | 15                                      | 0.014        | 1.983                       |                                       |
| 13           | 121                                              | 14                                      | 0.191        | 1.982                       |                                       |
| 12           | 121                                              | 13                                      | 0.200        | 1.982                       |                                       |
| 11           | 121                                              | 12                                      | 0.431        | 1.982                       |                                       |
| 10           | 121                                              | 11                                      | 0.536        | 1.982                       |                                       |
| 9            | 121                                              | 10                                      | 0.735        | 1.982                       |                                       |
| 8            | 121                                              | 9                                       | 1.237        | 1.981                       |                                       |
| 7            | 121                                              | 8                                       | 1.986        | 1.981                       | ✔                                     |
| 6            | 121                                              | 7                                       | 1.774        | 1.981                       |                                       |
| 5            | 121                                              | 6                                       | 2.158        | 1.981                       |                                       |
| 4            | 121                                              | 5                                       | 2.604        | 1.981                       |                                       |
| 3            | 121                                              | 4                                       | 3.178        | 1.980                       |                                       |
| 2            | 121                                              | 3                                       | 3.688        | 1.980                       |                                       |
| 1            | 121                                              | 2                                       | 13.043       | 1.980                       |                                       |

#### 5.6.1.3 Residual Mean Squares (RMS) of Errors

One measure that is used to judge the adequacy of a fitted equation is the RMS. For a p-term MLR equation, the RMS is defined as:

\[
RMS_p = \frac{SSE_p}{n - p}
\]  

(5.8)

where “\( SSE_p \)” is the residual sum of squares for a p-term equation. Among a number of developed models, the model with the smallest RMS is usually preferred, especially if the
main objective of using regression analysis is “estimation and prediction” (Chatterjee and Hadi, 2006).

The RMS value of each successive equation obtained by performing the BE procedure is visualized and listed in Figure 5-1 and Table 5-9, respectively. Again, model number seven was the most preferred regression equation for RNCC prediction.

5.6.1.4 Akaike Information Criteria (AIC)

Variable selection in the regression context can also be addressed using statistical measures, known as “Information Criteria”. The Akaike (1973) Information Criteria (AIC) in selecting a model seeks to balance the conflicting demands of accuracy and simplicity in terms of entering the smallest number of explanatory variables required. This approach meets the principle of parsimony and, therefore, this criterion is preferable when the key reason for employing regression analysis is “description and model building” (Chatterjee and Hadi, 2006). The AIC value for a p-term equation can be computed from the following equation:

\[
AIC_p = n \ln \left( \frac{SSE_p}{n} \right) + 2p \quad (5.9)
\]

Equations with a smaller AIC value are preferred although models with an AIC value not differing by two should be treated as equally adequate. However, the equation with the lowest AIC value should be considered as best and, therefore, be adopted.

The fluctuation of the AIC statistic over the whole range of generated equations is illustrated in Figure 5-2. One can also find the numeric value of AIC for each equation in Table 5-9. It is important to note that larger differences in AIC numerical values indicate significant differences between the qualities of the models under consideration.

Having the lowest AIC value, model number seven is once more entitled to be treated as the most desirable MLR model by which the RNCC can best be described.
5.6.1.5 Modified Akaike Information Criteria (AICc)

A modification of AIC to avoid especially the over fitting problem, originally suggested by Hurvich and Tsai (1989), is given by the following equation:

\[
AIC_c^p = AIC_p + \frac{2 (p + 2)(p + 3)}{n - p - 3}
\]  

(5.10)

According to Figure 5-2 or, alternatively, Table 5-9, there is little difference between the AICc and AIC criteria as both conclude with the same model, namely, model number seven.

5.6.1.6 Bayes Information Criteria (BIC)

To compare models by AIC a complete dataset with no missing values is needed, and the AIC value for different models should be calculated on the same set of observations. If there are many missing values for some explanatory variables, application of AIC may be inefficient. To remove the problem of missing values, Schwarz (1978) proposed the following equation as the Bayes Information Criteria (BIC). Another advantage of BIC is that it controls the over fitting tendency of AIC that results in a choice of model with unwarrantedly larger independent variables. The over fitting problem brings forth inappropriate results especially when the number of observations in a given dataset is relatively small.

\[
BIC_p = n \ln \left( \frac{SSE_p}{n} \right) + p (\ln n)
\]  

(5.11)

Unlike the previous criteria examined, the BIC statistic is much in favour of another model, that is, model number five. This model excels ahead of the other thirteen regression equations by possessing the lowest BIC value. The BIC value of each equation is depicted in Figure 5-2. Table 5-9 also provides each equation respective BIC value.
Table 5-9: Results of different statistics for evaluating regression models developed using the BE procedure

<table>
<thead>
<tr>
<th>No. of model</th>
<th>RMS</th>
<th>AIC</th>
<th>BIC</th>
<th>AICc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.099</td>
<td>-278.360</td>
<td>-272.769</td>
<td>-278.015</td>
</tr>
<tr>
<td>2</td>
<td>0.089</td>
<td>-289.555</td>
<td>-281.168</td>
<td>-289.034</td>
</tr>
<tr>
<td>3</td>
<td>0.083</td>
<td>-297.576</td>
<td>-286.393</td>
<td>-296.839</td>
</tr>
<tr>
<td>4</td>
<td>0.079</td>
<td>-302.453</td>
<td>-288.474</td>
<td>-301.462</td>
</tr>
<tr>
<td>5</td>
<td>0.076</td>
<td>-305.256</td>
<td>-288.481</td>
<td>-303.970</td>
</tr>
<tr>
<td>6</td>
<td>0.075</td>
<td>-306.548</td>
<td>-286.978</td>
<td>-304.926</td>
</tr>
<tr>
<td>7</td>
<td>0.073</td>
<td>-308.706</td>
<td>-286.339</td>
<td>-306.706</td>
</tr>
<tr>
<td>8</td>
<td>0.073</td>
<td>-308.341</td>
<td>-283.179</td>
<td>-305.919</td>
</tr>
<tr>
<td>9</td>
<td>0.073</td>
<td>-306.936</td>
<td>-278.978</td>
<td>-304.047</td>
</tr>
<tr>
<td>10</td>
<td>0.074</td>
<td>-305.250</td>
<td>-274.496</td>
<td>-301.848</td>
</tr>
<tr>
<td>11</td>
<td>0.074</td>
<td>-303.459</td>
<td>-269.910</td>
<td>-299.497</td>
</tr>
<tr>
<td>12</td>
<td>0.075</td>
<td>-301.504</td>
<td>-265.159</td>
<td>-296.933</td>
</tr>
<tr>
<td>13</td>
<td>0.075</td>
<td>-299.534</td>
<td>-260.393</td>
<td>-294.303</td>
</tr>
<tr>
<td>14</td>
<td>0.076</td>
<td>-297.534</td>
<td>-255.597</td>
<td>-291.592</td>
</tr>
</tbody>
</table>

Figure 5-1: Variation of RMS statistic within models under study
Figure 5-2: Variation of AIC, BIC, AIC<sub>c</sub> statistics within models under study

5.6.2 The Selected MLR Model for Estimating RNCC

A concise summary of all the criteria examined in the present study is presented in Table 5-10. According to this table, the results taken from applying different evaluating statistics are in good agreement with each other. The results of the leave-one-out cross validation technique, RMS, AIC, AIC<sub>c</sub>, and the minimum absolute t-value indicate that model number seven is the most desired model among all the fourteen different models examined in this study. However, the BIC statistic is entitled to be treated as a more stringent criterion as it suggests that model number five is the most proper model. This model has two variables less than model number seven. It is therefore worthwhile to select model number seven by which one can simply predict the mean value of Y given the values of the independent variables X₁, X₂, X₄, X₆, X₇, X₈, and X₁₁. This model has the added advantage of representing the largest set of significant variables, while including no insignificant variable.
Table 5-10: Summary of model selection criteria and the models suggested by each criterion

<table>
<thead>
<tr>
<th>model</th>
<th>Selection Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Leave-one-out technique</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5*</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7**</td>
<td>✓</td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

* Model 5 is the model derived from implementation of the forward selection and the stepwise techniques.
** Model 7 is the model derived from implementation of the backward elimination technique.

On the basis of the results shown in the table above, another important conclusion can be drawn. This conclusion derives from the consistency observed between the results of different selection criteria and the model brought about by the BE procedure per se (see Equation (5.5)). As described in Section 5.6, the cut-off p-value in this procedure for removing insignificant variables was set to 0.05 (p<0.05). Therefore, it can be concluded that the BE procedure, if being used with the 5% cut-off p-value, is a robust analytical technique for developing MLR models. Besides its robustness, this technique has the great advantage of reducing the computational time. This advantage is of exceptional value in those cases where a large number of independent variables are present, preventing examination of all feasible regression models.

5.7 Detection of the Proposed MLR Model Violations

The validity of a statistical method, such as regression analysis, and the conclusions derived from it depend crucially on certain assumptions. As such, the proposed MLR model in Equation (5.5) is not valid if the standard regression assumptions on which this model is based do not hold. Therefore, before deriving statistically meaningful conclusions from the proposed model, the validation of these assumptions must be first checked. Assumptions are usually made about the data and the model. A simple and effective method for inspecting these assumptions is the examination of the residual (i.e., error) plots. The residual of each
observation is composed from the difference between the actual and predicted values of the
dependent variable; that is, Ln RNCC. Residual plots will point to serious violations in one or
more of the standard regression assumptions when they exist. Presenting a collection of these
plots, the following sections aim to check the validity of the regression assumptions within
the proposed model.

5.7.1 Checking Linearity and Normality Assumption

The following plots of the standardized residuals are used to check the linearity and
normality assumptions:

5.7.1.1 Normal Probability Plot of the Standardized Residuals

This is a plot of the ordered standardized residuals versus the so-called normal scores. The
standardized residuals are simply the residuals divided through their standard deviation
and the normal scores are those that are expected to be obtained if a sample of size n is taken
from a standard normal distribution. If the residuals are normally distributed, the ordered
residuals should be approximately the same as the ordered normal scores. When using the
normality assumption, this plot should resemble a (nearly) straight line with an intercept of
zero and a slope of one. These values are the mean and the standard deviation of a standard
normal variable, respectively. As shown in Figure 5-3, the standardized residuals of the
proposed MLR model lie close to their corresponding normal scores, which is an indication
that these residuals are likely to be normally distributed.
Moreover, Figure 5-4 illustrates the normal distribution best fitted to standardized residuals. As shown in this figure, the mean and standard deviation of the residuals are 0.0 and 0.97, respectively, which are almost the same as those values by which a standard normal distribution is described (i.e., 0.0 and 1.0, respectively).

Figure 5-3: Normal probability plot of the standardized residuals

Figure 5-4: Best fitted normal distribution for the standardized residuals of the proposed MLR model
The normality assumption was also further controlled by examining the following hypothesis:

\[ H_0: \text{The standardized residuals follow the normal distribution} \]

\[ H_1: \text{The standardized residuals do not follow the normal distribution} \]

The chi-square test was used to test the hypothesis above. The test statistics, including the chi-square value and its corresponding p-value, were calculated to be 3.07 and 0.97, respectively. Therefore, the null hypothesis holds absolutely true as the p-value is far greater than the conventional significance level of 5\% (P>0.9).

5.7.1.2 Scatter Plot of the Standardized Residuals against the Predictor Variables

Based upon the standard assumptions, the standardized residuals are uncorrelated with each of the predictor variables. If this assumption holds, this plot should be a random scatter of points. Figures 5-5 through 5-11 illustrate the standardized residuals against each of the predictors by which the proposed MLR model is defined. In each figure, the best regression line that goes through the points is also shown by a red line. The \( R^2 \) value of each line was also computed and shown in the upper left corner of each figure. In all figures, the best regression line completely coincides with the horizontal axis originating at point 0 (i.e., the origin). All the \( R^2 \) values are also practically equal to 0, suggesting that there is really no trend in the residuals and consequently the assumption above is in place.
Figure 5-5: Standardized residuals vs. predictor X1

Figure 5-6: Standardized residuals vs. predictor X2

Figure 5-7: Standardized residuals vs. predictor X4

Figure 5-8: Standardized residuals vs. predictor X6

Figure 5-9: Standardized residuals vs. predictor X7

Figure 5-10: Standardized residuals vs. predictor X8

Figure 5-11: Standardized residuals vs. predictor X11
5.7.1.3 Scatter Plot of the Standardized Residuals against the Predicted Values

Based upon the standard assumptions, the standardized residuals are also uncorrelated with the predicted values. Therefore, the scatter plot of the standardized residuals against the predicted values should also be a random scatter of points. Figures 5-12 depicts the standardized residuals against the predicted values of dependent variable obtained by using the proposed MLR model. Similar to the response observed in Figures 5-5 to 5-11, the points in Figure 5-12 are also scattered randomly within a horizontal band around zero. Similarly, the $R^2$ value of the red regression line in Figure 5-12 also practically equates to 0, indicating that the residuals are clearly uncorrelated with the predicted values.

![Figure 5-12: Standardized residuals vs. predicted values of dependent variable Y](image)

5.7.2 Outliers

In fitting a model to a given body of data, it is necessary to ensure that the fit is not overly determined by one or few observations, known as outliers. Outliers may be found in either the dependent or independent variables.
5.7.2.1 Outliers in the Dependent Variables

Observations with large standardized residuals are outliers in the dependent variable because they lie far from the fitted equation in the Y-direction. Because the standardized residuals are approximately normally distributed with mean zero and a standard deviation close to 1, observations with standardized residuals larger than 2 or 3 standard deviations away from the mean (zero) are called outliers (Chatterjee and Hadi, 2006). The scatter plot in Figure 5-13 shows the standardized residuals obtained for the 121 observations in the development dataset.

![Index plot of Standardized Residuals](image)

According to the figure above, there is only one point (i.e., observation number 91) whose standardized residual is beyond the range as defined above. So, after excluding this observation from the development dataset, the regression analysis was repeated over the remaining 120 cases in this dataset. The resulting regression model is given by the following equation:

\[
Y = -1.909 + 0.769X_1 + 0.232X_2 + 0.203X_3 + 0.202X_4 + 0.147X_7 + 0.125X_8 + 0.126X_{11} \tag{5.12}
\]

<table>
<thead>
<tr>
<th>t-value</th>
<th>p-value</th>
<th>R²</th>
<th>Adjusted R²</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-4.387)</td>
<td>(0.005)</td>
<td>0.749</td>
<td>0.734</td>
<td>47.820</td>
</tr>
<tr>
<td>(11.843)</td>
<td>(0.000)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2.421)</td>
<td>(0.017)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2.269)</td>
<td>(0.025)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2.953)</td>
<td>(0.004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2.920)</td>
<td>(0.004)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2.265)</td>
<td>(0.025)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(2.683)</td>
<td>(0.008)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Considering the equation above, the omission of the observation number 91 led to insignificant changes in the characteristics (e.g., the estimated partial coefficients, t-values) of the proposed MLR model in Equation (5.5) developed using all the 121 observations. In Equation (5.12), like Equation (5.5), the entire partial coefficients still remain significant at 5% level of significance. The outlier also caused a minor change in the quality of the regression because the $R^2$ value of the equation above (i.e., 0.749) is slightly greater than that of Equation (5.5) (i.e., $R^2=0.710$). As the deletion of this outlier did not significantly affect the regression weights and its quality, it then remained in the development dataset. Therefore, the proposed MLR model in Equation (5.5) is still valid to be used.

5.7.2.2 Outliers in the Independent Variables

Outliers can also occur in the predictor variables (the X-space). These outliers can also affect the regression results. Leverage values can be used to measure outlyingness in the X-space. In any analysis, points with high leverage should be flagged and then examined to see if they are also influential. An index plot of the leverage values, such as shown in Figure 5-14, reveals points with high leverage if they exist. Besides the index plot, a commonly used cut-off value for distinguishing high leverage points is $2(p+1)/n$, where $p$ is the number of predictors and $n$ is the number of observations in the development dataset (Hoaglin and Welsch, 1978). Therefore, in the proposed MLR model in Equation (5.5), observations with the leverage greater than 0.132 should be generally regarded as points with high leverage. In Figure 5-14, observations 4, 19, 91, and 119 with leverage value of 0.154, 0.449, 0.158, and 0.232 are therefore flagged as high leverage points as they are located far from the bulk of other points.
To examine whether or not high leverage observations are also influential, a widely used influence measure proposed by Cook (1977) was utilized. Cook’s distance measures the difference between the regression coefficients obtained from the full data and the regression coefficients obtained by deleting the \( i^{th} \) observation, or equivalently, the difference between the fitted values obtained from the full data and the fitted values obtained by deleting the \( i^{th} \) observation. If an observation is influential, its deletion causes large changes and the value of \( C_i \) will be large. Therefore, a large value of \( C_i \) indicates that the point is influential. A practical operational rule is to classify points with \( C_i \) values greater than 1 as being influential (Chatterjee and Hadi, 2006). An index plot of Cook’s distances is shown in Figure 5-15. According to this figure, no value of \( C_i \) exceeds its cut-off value of 1 (\( C_i < 1 \)). However, this figure shows clearly that observation 91 should be considered as an influential observation. It should be noted that this observation was the one that previously appeared in Figure 5-13 to be an outlier in the Y-space, when examining the index plot of standardized residuals. However, the deletion of this observation was proven in Section 5.7.2.1 to have only a minor effect on the final results of the proposed MLR model. As a consequence, the observations with high leverage values are neither influential nor cause problems and therefore they should retain in the development dataset.
5.7.3 Multicollinearity

Interpretation of the MLR models depends implicitly on the assumption, known as multicollinearity, that the independent variables are not strongly interrelated. This interpretation may not be valid if there are strong linear relationships among independent variables. In any regression analysis, one therefore needs to detect the presence of multicollinearity in a given MLR model to avoid the pitfalls resulting from this problem. In present research two approaches are used to distinguish whether multicollinearity actually exists within the independent variables of the proposed model.

5.7.3.1 The Variance Inflation Factor (VIF)

A thorough investigation of multicollinearity involves examining the value of the coefficient determination parameter ($R^2$) that results from regressing each of the independent variables against all the others. The relationship between the independent variables can be judged by examining a quantity called the variance inflation factor (VIF), given by:

$$VIF_j = \frac{1}{1 - R_j^2}, \quad j = 1, \ldots, p,$$

(5.13)
where $R_{ij}^2$ represents the coefficient determination parameter that results when the independent variable, $X_j$, is regressed against all the other predictor variables and $p$ is the number of predictor variables. As a rule of thumb, if none of the VIFs in a given MLR equation are greater than 5, collinearity is not a problem (Studenmund, 2011). The VIFs calculated in the proposed MLR model in Equation (5.5) are:

$$VIF_1 = 1.66, \quad VIF_2 = 1.82, \quad VIF_3 = 1.23, \quad VIF_4 = 1.03,$$

$$VIF_5 = 1.01, \quad VIF_6 = 1.04, \quad VIF_7 = 1.05,$$

The range of VIFs (1.01 to 1.82) implies there is no linear interrelationship among the independent variables and, therefore, multicollinearity is not a problem.

### 5.7.3.2 Principal Component Approach

The principal component approach to the detection of multicollinearity is based on the fact that any set of $p$ variables can be transformed to a set of $p$ orthogonal variables (Chatterjee and Hadi, 2006). When there is a complete absence of linear relationship among the independent variables, they are said to be orthogonal. The new orthogonal variables are known as the “principal components” (PCs). The variance-covariance matrix of PCs is of the form:

$$
\begin{bmatrix}
\lambda_1 & 0 & \ldots & 0 \\
0 & \lambda_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & \lambda_p
\end{bmatrix}
$$

(5.14)

The $\lambda$’s are called eigenvalues of the matrix in Equation (5.14). On the basis of the value of $\lambda$’s, three different rules have been set to examine the presence of multicollinearity (Chatterjee and Hadi, 2006):
1: If the “condition number” is larger than 15. The condition number can be calculated by the equation below:

\[
k = \sqrt{\frac{\text{Maximum Eigenvalue } (\lambda_{\text{max}})}{\text{Minimum Eigenvalue } (\lambda_{\text{min}})}} > 15 \quad (5.15)
\]

2: If the sum of the reciprocals of the eigenvalues is greater than five times the number of predictor variables.

\[
m = \sum_{j=1}^{p} \frac{1}{\lambda_j} > 5p \quad (5.16)
\]

3: If any of the individual eigenvalues are less than 0.01.

\[
\lambda_j < 0.01 \quad (5.17)
\]

If the conditions above do not hold, the variables are regarded as noncollinear. The eigenvalues of the correlation matrix of the independent variables contributing to the proposed MLR model are as follows:

\[\lambda_1 = 1.69, \quad \lambda_2 = 1.18, \quad \lambda_3 = 1.12, \quad \lambda_4 = 1.02, \]
\[\lambda_5 = 0.85, \quad \lambda_6 = 0.82, \quad \lambda_7 = 0.32, \]

Criteria established to inspect the presence of multicollinearity among the independent variables were all examined. The results are shown in Table 5-11. Because none of these criteria are met, it can be concluded that the independent variables present in the proposed MLR model are not collinear.
5.7.4 Homoscedasticity

The constancy of the residual variance is one of the standard assumptions of least squares theory. This assumption is often referred to as the assumption of homoscedasticity. When the residual variance is not constant over all the observations, the residual is said to be heteroscedastic (Gujarati and Porter, 2010). If Heteroscedasticity is present, the application of ordinary least squares will result in estimated coefficients which lack precision in a theoretical sense. The estimated standard residuals of the regression coefficients are often understated, giving a false sense of accuracy. Heteroscedasticity is usually detected by suitable graphs of the residuals such as the scatter plot of the standardized residuals against each of the independent variables or against the predicted values. These plots are already shown in Figures 5-5 through 5-11 for the independent variables and also Figure 5-12 for the predicted values of the dependent variable. These figures exhibit no discernible systematic pattern for the distribution of the standardized residuals as these residuals are randomly distributed and lie within a band parallel to the horizontal line through 0. Therefore, there is no marked evidence of heteroscedasticity in the proposed MLR model.

5.7.5 Autocorrelation

The final potential violation within the regression analysis is known as the autocorrelation problem, where the residuals are not independent. The most celebrated test for detecting autocorrelation is that developed by Durbin and Watson, popularly known as
Durbin-Watson d statistic (Gujarati and Porter, 2010). This statistic has a range of 0 to 4. If a computed d value is closer to 0, there is evidence of positive autocorrelation, but if the d value is closer to 4, there is evidence of negative autocorrelation. The closer the d value is to 2, the more the evidence is in favour of no autocorrelation. In the proposed MLR model, the value of d is found to be 1.916, indicating that the autocorrelation is not a problem.

5.8 Causal Relationships in the Proposed MLR Model

As proved in Section 5.7, no evidence was found that the regression assumptions are violated by the proposed model. This finding validates one of the most important products of regression analysis; that is, the explicit determination of the regression equation, known also as the regression inference. This inference describes the causal relationships which exist between RNCC and each of its predictors in the proposed model. For this purpose, each variable coefficient in this model needs to be accordingly interpreted with regard to the model functional form.

When the quantitative variables in the proposed MLR model in Equation (5.5) (i.e., Y, X₁, X₂, and X₄) are substituted by those in Equations (5.3) and (5.4), the following equation is obtained:

\[ \text{Ln}(Y^*) = -1.238 + 0.674 \text{Ln}(X₁^*) + 0.273 \text{Ln}(X₂^*) + 0.195 \text{Ln}(X₄^*) + 0.218X₆ + 0.117X₇ + 0.118X₈ + 0.132X₁₁ \]

(5.18)

Now by substituting the corresponding variable definitions given in Table 5-1 for those in the regression model above, this model can be rewritten as:

\[ \text{Ln}(\text{RNCC}) = -1.238 + 0.674 \text{Ln}(\text{Area}) + 0.273 \text{Ln}(\text{No. of stories}) + 0.195 \text{Ln}(\text{Weight per unit}) + 0.218(\text{Seismicity}) + 0.117(\text{Soil type}) + 0.118(\text{Plan irregularity}) + 0.132(\text{Structural type}) \]

(5.19)

In the model above, the explanatory contents of the independent variables vary. The greatest explanatory content can be identified for the variable X₁ (Ln (Area)) and the smallest content is contributed by X₇ (Soil type). The power of the explanatory content of the other
independent variables is respectively lessening from $X_2$ (Ln (No. of stories)) to $X_6$ (Seismicity), $X_4$ (Ln (Weight per unit)), $X_{11}$ (Structural type), and $X_8$ (Plan irregularity).

The relations between each independent variable and the dependent one in the proposed MLR model are described in the following sections. The sections are arranged in the order of variables sequence in Equation (5.19).

### 5.8.1 The Relation between Building Area and RNCC

As the last independent variable to quit from the BE procedure, $X_1$ (Ln (Area)) is the most important variable amongst all of fourteen independent variables examined in the case of framed structures in this study. Figure 5-16 depicts the relationship between $X_1$ and Ln (RNCC). The trend line shown in Figure 5-16 also clearly suggests that these variables are positively related.

Ln (Area) is highly indicative as it possesses the largest explanatory content in Equation (5.19). This content can be described in accordance with the functional form of the proposed MLR regression model. As described in Section 5.4, the partial slope coefficient of 0.674 measures the elasticity of RNCC with respect to building area. Specifically, this number states that:

"Holding all other variables constant, if the area increases by 1%, on the average, RNCC rises by approximately 0.67%.”

As the partial elasticity coefficient of 0.674 is less than 1, the RNCC increases proportionally less than the increase in area. Consequently, the retrofit cost indicator (RCI), specified on a dollar-per-square-meter basis (U.S. $/m^2), would decrease as the area increases. Using calculus, it can be shown that the decrease percentage of the RCI mean value (denoted by $\bar{RCI}$) with respect to a given percentage increase of area (denoted by $x$) is determined by the following equation:

$$f(x) = \text{the percentage decrease in the } \bar{RCI} = \frac{0.326x}{1 + 0.01x}$$  \hspace{1cm} (5.20)
Figure 5-17 illustrates the decrease rate in the $\tilde{RCL}$ values for a certain range of $x$ values. From this figure it is apparent that the $\tilde{RCL}$ value drops when the area of the building increases. This figure also suggests that the relationship between building area and $\tilde{RCL}$ is in fact nonlinear such that the $\tilde{RCL}$ declining rate decreases as the $x$ increases. Consequently, as the building total area becomes larger, the $\tilde{RCL}$ values tend to decrease proportionally less than the increase in building area. However, Equation (5.20) should be used with caution. It is important to note that this equation is less erroneous if a small value of $x$ applies. So as the $x$ increases, the error of the results obtained from this equation would also proportionately increase. Another conclusion that can be drawn from Equation (5.20) is that the difference between the RCI values of larger buildings would be less significant in comparison with the difference occurring between the RCI values of smaller buildings. This conclusion suggests that as the building size becomes bigger, the importance of building area in determining RCI values or, alternatively, in describing the RCI total variability is proportionally decreasing.

![Figure 5-16: Distribution of Ln (RNCC) in the development dataset as a function of Ln (Area)](image)

Figure 5-16: Distribution of Ln (RNCC) in the development dataset as a function of Ln (Area)
5.8.2 The Relation between Building Number of Stories and RNCC

$X_2$ $(\text{Ln (number of stories)})$ is the penultimate variable to quit from the BE procedure. Therefore, after the building area, the greatest part of the change in the average value of $Y$ $(\text{Ln (RNCC)})$ can be directly attributed to this variable. Figure 5-18 depicts the variation of $Y$ in the development dataset in accordance with $X_2$. The trend line shown in Figure 5-18 also clearly suggests that $Y$ and $X_2$ are positively related.

$X_2$ has the second largest explanatory content in the proposed model. The numeric value obtained for the partial coefficient of this variable in Equation (5.19) (i.e., 0.273) can be interpreted in the same way as that of the building area, suggesting that:

"Holding all other variables constant, if the number of stories increases by 1%, on the average, RNCC rises by approximately 0.27%.”
Similar to the building area, the RNCC tends to be increased proportionally less than the increase in the number of stories. Like the building area, the interpretation above is more valid when the order of increase rate is small. Figure 5-19 shows the change in percentage increase of RNCC for a given percentage increase in the number of stories. The blue curve in Figure 5-19 represents the actual change, whereas the red straight line appears as the estimated change resulted from the interpretation above. In Figure 5-19, the difference between the actual and estimated change is more pronounced with larger increase rates.

It is also important to note that although the blue curve is positively sloping (i.e., the larger the percentage increase of X, the larger the percentage increase of RNCC), the curve is increasing at a decreasing rate. For instance, if the number of stories increases by 50% (e.g., the number of stories rises from two to three), the RNCC rises by approximately 12%. However, when the number of stories increases by 100% (e.g., the number of stories becomes double), the increase rate of RNCC does not double and is about 21%.

Another conclusion that can be inferred from the discussion above is that the difference between the RNCC values of high-rise buildings would be less significant in comparison with the difference occurring between the RNCC values of low-rise buildings. This finding suggests that, like building area, the retrofit cost of buildings with a larger number of stories will be subject to less alteration than those of low-rise buildings.

Figure 5-18: Distribution of Ln (RNCC) in the development dataset as a function of Ln (Number of Stories)
5.8.3 The Relation between Building Weight Indicator and RNCC

Of the four quantitative variables examined in the present research, $X_4$ ($\ln$ (Weight per unit)) is the third and the last variable found to be statistically significant in appraising RNCC. Similar to building area and building number of stories, building weight exhibits a positive impact on RNCC. As shown in Figure 5-20, the value of $Y$ ($\ln$ (RNCC)) increases as the building weight indicator increases. This indicator is expressed in terms of ton per square metre. The partial slope coefficient of $X_4$ in the proposed regression Equation (5.19) is 0.195, implying that:

“Holding all other variables constant, if the building weight indicator increases by 1%, on the average, RNCC rises by 0.195%.”
5.8.4 The Relation between Seismicity and RNCC

Of the ten categorical variables considered in the present study, seismicity is the most important variable in defining RNCC. According to Equation (5.19), the partial dummy coefficient of this variable, denoted by $X_6$, is estimated to be +0.218. While being statistically significant, this value indicates that there is indeed a difference between RNCC of buildings located in regions with the highest level of seismic activity and RNCC of buildings constructed in the less intensive seismicity regions. Figure 5-21 illustrates the scatter plot of $Y$ (Ln (RNCC)) versus $X_6$. In this figure, the observations in the development dataset are concentrated at two extremes, namely, 0 and 1. The latter value represents the buildings located in “very high” seismicity regions, while the former addresses the buildings located in regions with less seismic activity. The trend line shown in Figure 5-21 signifies that $Y$ and $X_6$ are positively related, suggesting that an increase in earthquake hazard would result in a more costly retrofit endeavour.

To further study the effect of seismicity on the retrofit cost, let $Y_{se1}$ and $Y_{se0}$ respectively denote the RNCC of the latter and former groups mentioned above. Hence, according to the partial dummy coefficient of 0.218 in the proposed regression model, the following equation gives the difference of the retrofit cost between the two groups:
\[ Ln \left( Y_{se1} \right) - Ln \left( Y_{se0} \right) = 0.218 \Rightarrow Ln \left( \frac{Y_{se1}}{Y_{se0}} \right) = 0.218 \] (5.21)

Using logarithmic transformation, the equation above can be expressed in an alternative, but equivalent, form, as follows:

\[ Y_{se1} = e^{0.218} \times Y_{se0} \Rightarrow Y_{se1} = 1.244 \times Y_{se0} \] (5.22)

A very important conclusion can be derived from the equation above; that is:

“Holding all other variables constant, the mean value of RNCC in the buildings located in very high seismic zones is almost 24% more than that of those located in less seismic zones.”

Figure 5-21: Distribution of Ln (RNCC) in the development dataset as a function of seismicity

5.8.5 The Relation between Soil Type and RNCC

Incorporated as the other site-related variable into the MLR analysis, the soil type variable \((X_7)\) is found to be a statistically significant predictor of RNCC. Figure 5-22 displays
the dispersion of Ln (RNCC) according to this qualitative variable. In this figure, the observations are grouped into two values, namely, 0 and 1. The latter value represents a soil type with poor seismic characteristics, while the former value represents a more dense soil type which is also more resistant against induced seismic geological hazards. Like seismicity, the trend line slope shown in Figure 5-22 is positive, indicating that the cost of retrofitting is of higher order in sites where the soil condition is poor.

Being significant at the 5% significance level, the dummy coefficient of 0.117 obtained for $X_7$ in Equation (5.19) points out the effect that different soil types may impose on retrofit cost. Let $Y_{so1}$ and $Y_{so0}$ respectively denote the RNCC of the latter and former groups mentioned above. Following the same procedure described for seismicity, the equation below gives the difference of the retrofit cost between the two groups;

$$Y_{so1} = e^{0.117} \times Y_{so0} \Rightarrow Y_{so1} = 1.124 \times Y_{so0}$$

(5.23)

What can be inferred from the equation above is:

“*Holding all other variables constant, the presence of soil types with poor seismic characteristics would possibly increase the mean value of RNCC by almost 12%.*

Figure 5-22: Distribution of Ln (RNCC) in the development dataset as a function of soil type
5.8.6 The Relation between Plan Irregularity and RNCC

As an architectural-originated characteristic of a building, plan irregularity ($X_8$) turns out to be a significant predictor of RNCC. There is in fact a difference between the retrofit cost of the buildings that are regular in plan and the buildings that are not. Depicted in Figure 5-23 is the distribution of dependent variable $Y$ (Ln (RNCC)) while plan irregularity is taken into account. Like other qualitative variables discussed so far, in this figure the observations are clustered at two points, 0 (standing for buildings with regular plan configuration) and 1 (representing buildings with irregular plan configuration). Let $Y_{P10}$ and $Y_{P11}$ denote the RNCC of buildings with respectively regular and irregular plan configuration. Therefore, it can be shown that:

$$Y_{P11} = e^{0.118} \times Y_{P10} \Rightarrow Y_{P11} = 1.125 \times Y_{P10}$$  \hfill (5.24)

The equation above implies that:

"Holding the influence of other variables constant, the mean value of RNCC in the buildings with irregular plan is nearly 12% more than those in which there is no evidence of plan irregularity."

Figure 5-23: Distribution of Ln (RNCC) in the development dataset as a function of plan irregularity
5.8.7 The Relation between Structural Type and RNCC

In the viewpoint of earthquake-resistant characteristics of buildings, two distinct categories of structure were examined in this study; buildings with the in-place existing seismic resisting elements and buildings with the absence of such elements. Because the qualitative variable of structural type has two classes, an indicator variable \((X_{11})\) was used to introduce these classes into the MLR analysis. Subsequently, the indicator variable takes on the value of 0 for the less earthquake-prone class and 1 for buildings assumed to be seriously damaged in the event of a severe earthquake. Figure 5-24 is the visual display of the distribution of the dependent variable \(Y (\ln (RNCC))\) across structural type classes.

![Figure 5-24: Distribution of Ln (RNCC) in the development dataset as a function of structural type](image)

According to Equation (5.19), the dummy coefficient of variable \(X_{11} (+0.132)\) is statistically significant at the 5% level of significance. In addition, the estimated coefficient has the expected positive sign; indicating that the mean value of RNCC in the buildings suffering from the absence of an existing seismic resisting system is, to be sure, more than those benefitting from the existing lateral force resistant elements. By letting \(Y_{ST1}\) and \(Y_{ST0}\) denote respectively the RNCC of the former and latter structural type, it can be shown that:
\[ Y_{ST1} = e^{0.132} \times Y_{ST0} \Rightarrow Y_{ST1} = 1.141 \times Y_{ST0} \] (5.25)

A salient conclusion that can be drawn from the equation above is:

“\textit{Holding the influence of other variables constant, the absence of earthquake-resistant system in the existing building would possibly increase the RNCC by almost only 14%}.”

5.9 Performance on the Test Dataset

It was intended so far to develop MLR models such that to map the correct RNCC for each of observations in the development dataset. However, the performance of the MLR models over data to which they have not been previously exposed is of high importance because such data are representative of situations in which the MLR models are to be used. Therefore, the final objective of this chapter is to compare the predictive ability of the MLR models when being subjected to unforeseen data in the test dataset. This dataset is comprised of thirty-seven observations which were randomly separated out from the total of 158 observations in the whole database for the purpose stated above.

The last seven MLR models derived from implementing the backward elimination procedure (i.e. models number one to seven in Table 5-4) were utilized to predict the dependent variable in the test dataset. These models were selected because they were all developed using the most statistically significant independent variables. The goodness-of-fit index, \(R^2\), was used as a summary measure to compare the predictive ability of MLR models when being applied to the test dataset. This index is also known as the coefficient of determination because it gives an indication of how the independent variables determine the dependent variable. Therefore, the larger the \(R^2\) over the test dataset, the better the capability of the MLR model to predict unforeseen data which are never used in developing the model.

Care should be taken when predicting the value of RNCC corresponding to a set of values for independent variables in the test dataset that do not lie close to the observed data in the development dataset. There are two dangers in such predictions (Chatterjee and Hadi, 2006). First, there is substantial uncertainty due to the large standard error. More importantly, the linear relationship that has been estimated may not hold outside the range of observations. Therefore, one needs to be cautious in employing regression models for prediction far outside the range of observations in the development dataset. In this study, the MLR models can be
employed to predict RNCC in the test dataset as the values of RNCC along with its predictors in this dataset lie largely inside the range of their corresponding values in the development dataset.

Figure 5-25 shows the performance of the first seven MLR models when being applied to the test dataset. The overall fit of the models to the test dataset, as measured by $R^2$, varies from 0.56 for the third model to 0.69 for the first model. Offering the largest goodness-of-fit, the first MLR model, being made of only one independent variable, provides better predictions than those from other MLR models using a greater number of independent variables. As a result, this model can serve as a more reliable analytical tool for the prediction of RNCC. Besides its reliability, the first MLR model is also more efficient than other MLR models developed. This model reduces the number of independent variables to a manageable number without loss of model accuracy. This reduction also complies with the principal of parsimony which is emphasized in the literature (Gujarati and Porter, 2010; Hwang, 2009). The first regression model enables the end user to predict the RNCC of a project with the least information about the project; that is, building area. Therefore, project managers, cost planners, estimators, and the like can use this model to predict RNCC before a project commences so that they can better plan their projects and so that clients can better arrange their funding requirements. The importance of this issue is more pronounced in cases where no detailed information about the building characteristics, site conditions, and also retrofit solution plans is available at the time of prediction.

As reported in Table 5-3, the first MLR model is defined by the following equation:

$$\ln(RNCC) = -1.273 + 0.771 \ln(Area)$$  \hspace{1cm} (5.26)

\[ t-value= (-2.883) \quad (13.043) \]
\[ p-value= (0.005) \quad (0.000) \]
\[ R^2= 0.588 \hspace{1cm} Adjusted \, R^2= 0.585 \hspace{1cm} F= 170.132 \]

The equation above suggests that a natural logarithm transformation of the dependent and independent variables linearizes the relationship between these variables. This transformation allows a simple linear regression model with “area” as the independent variable and “RNCC” as the dependent variable to be developed. Equation (5.26) can be,
However, transformed by taking the exponential of each side of the equation to show that there is in fact a nonlinear relationship between RNCC and area:

\[
RNCC = 0.280 \cdot (\text{Area})^{0.771}
\]  

(5.27)

On the basis of the equation above, it is now justifiable to conclude that the retrofit cost of framed structures can be soundly predicted by a simple cost-area model expressed in the form of \( C = K A^B \), where \( C \) is the retrofit net construction cost in thousands U.S$, \( A \) is the area in square meter, \( K \) is a constant describing the general level of cost performance for a $1,000 retrofit project, and \( B \) is a constant describing how the cost performance is affected by project size as measured by building area.

The general form of the model proposed above for predicting RNCC is also supported by some researchers in other fields. The time–cost model \( (T = KC^B) \) developed by Bromilow (1969) and the time–area model \( (T = KA^B) \) developed by Chan and Kumaraswamy (1995) are some examples in this regard. Furthermore, a more recent study by Chen and Huang (2006) showed that floor area presents a strong relationship with cost of school reconstruction projects in central Taiwan, which received the most serious damage from the Chi-Chi Earthquake. The results of this study showed that the relationship between reconstruction cost and floor area can be best described by a model similar to the one proposed above. This consistency indicates that the suggested cost-area model can be applied not only to retrofit projects, but also to reconstruction projects.
Figure 5-25: Comparison of $R^2$ values of the first seven MLR models over the test dataset
5.10 Interaction Term

An interaction term is an independent variable in a regression analysis that is the multiple of two or more other main variables (Studenmund, 2011). Main variables refer to each of the elements that constitute the interaction term. Thus, $X_1$ to $X_{14}$ are the main variables in this research. Interaction terms allow researchers to explore the effects of combined predictors on a given dependent variable. However, the introduction of interaction terms into construction cost models is not a common practice, as evidenced by numerous studies where it was not used (e.g., Abu Hammad et al., 2010; Attalla and Hegazy, 2003; Dela Garza and Rouhana, 1995; Elhag and Boussebaine, 1998; Elhag and Boussebaine, 1999; Emsley et al., 2002; FEMA 156, 1994; Georgy and Barsoum, 2005; Günaydın and Doğan, 2004; Hwang, 2009; Khosrowshahi and Kaka, 1996; Kim et al., 2004; Love et al., 2005; Lowe et al., 2006; McKim, 1993b; Potangaroa, 1985; Shehab et al., 2010; Skitmore and Ng, 2003; Soutos and Lowe, 2005; Stoy and Schalcher, 2007; Williams, 2002; Williams et al., 2005; Wilmot and Cheng, 2003; Wilmot and Mei, 2005). In the literature, the inclusion of interaction terms into a cost model was limited to exceptional cases where such inclusion was necessitated by robust theoretical reasons (e.g., Cheung and Skitmore, 2006). The lack of interest in using interaction terms in previous cost estimation studies can be largely explained by three reasons.

The first reason is that the inclusion of interaction terms can make it difficult to describe the effect of main variables on the dependent variable. As can be inferred from the literature (e.g., Ai and Norton, 2003; Brambor et al., 2006; Norton et al., 2004; Balli and Sørensen, 2012), this difficulty arises because coefficients of variables that are involved in interaction terms do not have a straightforward interpretation. In addition, a number of researchers argued that the interaction term undermines the interpretation of the regression coefficients associated with the main variables (e.g., Smith and Sasaki, 1979; Braumoeller, 2004). As noted by Balli and Sørensen (2012), the point is simply that researchers sometimes do not notice the change in the interpretation of the coefficient estimate for the main variables when the interaction term is added.

The second reason is that adding interaction terms to main variables causes the degree of freedom (D.F.) to reduce. This reduction becomes severe in cases where the number of main variables is large, as in such cases hundreds of interaction terms can be constructed and examined in a regression analysis. For instance, as determined by Equation (5.28), the total
number of just bilinear interaction terms (e.g., \(X_1.X_2, X_1.X_3, X_2.X_3\), etc.) that can be included in the present study is 91. According to Equations (5.29) and (5.30), the sole inclusion of these terms would dramatically decrease the D.F. of regression analysis from 107 to just 16. Such a large decrease could, in turn, pose a serious question on the reliability and stability of the regression analysis performed. Noteworthy, this analysis will completely lose its credibility by adding, for example, the set of all possible tri-linear interaction terms (e.g., \(X_1.X_2.X_3, X_2.X_3.X_4\), etc.), simply because this addition reduces the D.F. to less than zero. That is why in the literature (e.g., Gujarati and Porter, 2009; Pearce and Moran, 1994; Shehab et al., 2010) it is asserted that a viable strategy for improving the predictive performance of a regression model is to reduce the number of independent variables. According to Shehab et al., (2010), this reduction in variables not only will improve the degree of freedom of the model but will also eliminate the negative impact of the multicollinearity. As described in the next paragraph, multicollinearity is another problem brought about by interaction terms. In the following equations, \(n\) and \(p\) denote the number of samples in the development dataset and the number of main independent variables, respectively.

\[
\text{Bilinear interaction terms} = \binom{p}{2} = \frac{p!}{2!(p - 2)!} = \frac{14!}{2!(14 - 2)!} = 91 \quad (5.28)
\]

\[
\text{D.F. without "interaction terms"} = (n - p) = (121 - 14) = 107 \quad (5.29)
\]

\[
\text{D.F. with "bilinear interaction terms"} = (n - p) = (121 - 14 - 91) = 16 \quad (5.30)
\]

Finally, the third, and more important, reason is associated with the multicollinearity problem. It has been well established that no matter what form the interaction variable takes (e.g., bilinear, tri-linear, etc.), all main variables should be included in a regression model (Balli and Sørensen, 2012; Brambor et al., 2006). Otherwise, the interaction effect may be significant due to left-out variable bias (Balli and Sørensen, 2012). Thus, for example, \(X_1\) and \(X_2\) should be included when the interaction term is \(X_1.X_2\). As shown in the literature (e.g., Althauser, 1971), this interaction term (i.e., \(X_1.X_2\)) is likely to be highly correlated with the main variables (i.e., \(X_1\) and \(X_2\)). Resultantly, the inclusion of interaction terms may violate one of the most critical assumptions of the OLS method, leading to the multicollinearity problem (Khandker et al., 2010; Osborn, 2012; Smith and Sasaki, 1979). According to Chatterjee and Hadi (2006), this problem is associated with unstable estimated regression coefficients in a sense that the presence or absence of certain variables has a large effect on
the other coefficients. This instability can, therefore, seriously limit the use of regression analysis for inference and forecasting. As explained earlier, there are too many interaction terms to have them exhaustively included in a regression analysis, so that in the literature this inclusion has instead been invoked in accordance with theory or intuition (e.g., Balli and Sørensen, 2012; Brambor et al., 2006; Cheung and Skitmore, 2006; Osborn, 2012). To examine whether the inclusion of interaction terms would have beneficial effects on the prediction of the RNCC, a number of intuitively appealing interaction terms were added to the initial set of 14 independent variables listed in Table 5.1. These terms represent interactions between two numerical variables (X1.X2), between one numerical and one categorical variable (X2.X11 and X4.X7), and between two categorical variables (X6.X11, X7.X8, and X8.X11). Another reason for considering these six interaction terms was that they were all constituted by those variables which were previously shown to be statistically significant predictors of the RNCC. The implementation of the backward elimination (BE) procedure on the new set of independent variables (i.e., 20 variables) resulted in exactly the same MLR model as was obtained before adding the interaction terms (see Equation (5.5)). This similarity indicates that the interaction terms were sequentially removed by the BE procedure from the final model due to their insignificant impact on the RNCC prediction (i.e., \( p > 0.05 \)). However, for the sake of investigating the multicollinearity problem, Equation (5.5) was recalculated when the interaction term X1.X2 was entered into this equation. The result of this recalculation is represented by Equation (5.31). According to this equation, the coefficient of X1.X2 is negative, which is contrary to prior expectation as it is conceivable that a simultaneous increase in building area and number of stories would cause an increase in the RNCC. Multicollinearity is a possibility here and in fact is the case. As illustrated in Equation (5.31), the variance inflation factors for the variables X1, X2, and X1.X2 surpass the cut-off score of 5 (see Section 5.7.3.1), indicating that the multicollinearity problem is brought about by a strong linear relationship amongst these variables. The magnitude of this excess is extremely large for the last two variables, with VIF values of 128.07 and 154.63, respectively.

\[
Y = -2.058 + 0.786X_1 + 1.230X_2 - 1.290 (X_1.X_2) + 0.173X_4 + 0.226X_6 + 0.118X_7 + 0.129X_8 + 0.140X_{11}
\]

\[\text{VIF: } (5.44) \quad (128.07) \quad (154.63) \quad (1.28) \quad (1.04) \quad (1.01) \quad (1.06) \quad (1.07)\]
From the above discussion it is concluded that the inclusion of interaction terms may cause serious problems, particularly when a study is not specifically designed to assess them and when there is no a priori reason to expect an interaction term. In summary, interaction terms are required only when there is a reason to believe that the effect of one independent variable depends on the value of another independent variable. As for the study performed by Cheung and Skitmore (2006), such variables should also be added when there are theoretical reasons for doing so. In other cases, like the study presented here, the inclusion of interaction terms might not only deteriorate the stability of a regression analysis but could also violate the assumptions implicit in this analysis.
5.11 Conclusion

The main objective of this chapter was to develop and validate MLR models for prediction of retrofit net construction cost (RNCC) in response to different independent variables derived from building characteristics, site conditions, and retrofit solution types. Fourteen questionnaire items, some of which have never been studied before, were considered as possible independent variables that may potentially influence RNCC as the dependent variable. The values of RNCC, along with those of its quantitative predictors, were all transformed to their natural logarithm before performing the regression analysis. Because there were also categorical variables involved in this analysis which had no quantitative meaning, these variables were converted into a series of dichotomously scored dummy variables, taking on only two values, namely, 0 and 1. In addition, of the total 158 data, 121 cases were randomly selected for developing the models, while the remaining 37 cases in the test or hold-out dataset were mainly utilized to evaluate how precise the developed regression models are in predicting the RNCC of new data never used in their development.

The development of cost models was achieved through addressing the question of what combination of the independent variables could optimize the prediction capabilities of the MLR models. Using the backward elimination (BE) procedure, fourteen different equations were developed. The model with the best predictive performance was selected based on defined selection criteria including the leave-one-out cross validation technique, minimum absolute t-value, and residual mean squares of errors, together with Akaike and Bayes information criteria. These criteria resolve the question regarding how proficiently the various MLR models developed by implementing the BE procedure predict the RNCC. The selected model can ultimately incorporate seven independent variables which are relevant to building and site characteristics. These variables (including the building area, building number of stories, building weight indicator, seismicity, soil type, building configuration in terms of plan irregularity and building structural type) were all found to be statistically significant predictors of RNCC (p<0.05). The remaining seven variables including the variables addressing the characteristics of retrofit solutions were filtered out because their contribution to RNCC determination was found to be statistically insignificant (p>0.05). The results also showed that each of the predictors significantly impacting the cost of retrofit had a certain level of importance to RNCC variance. Building area was the most important variable in this regard. Building number of stories and seismicity were the second and third most important variables to RNCC variance, respectively. Structural type and soil type were
right in the middle and the building weight indicator and plan irregularity were the last two statistically significant predictors entered into the model, respectively. The comparison of the $R^2$ values obtained from different MLR models indicated that the inclusion of these seven variables improves the explanatory power of the regression analysis from 0.59, as suggested by the first model with the contribution of only the building area, to 0.71.

Analyses were also made to check the validity of the regression assumptions in the final model. This chapter presents a collection of graphical displays that were utilized for accomplishing this task. Using these graphs, no evidence was found that the underlying assumptions of regression analysis (i.e., normality, multicollinearity, homoscedasticity, and autocorrelation) are not in order. Also, no influential outlier was found for the model. These model diagnostics further support the robustness of the proposed MLR model and also prove the sufficiency and reliability of the data used to develop this model. Within the range of the observed data, the validity of inferences proposed by this model was therefore ensured. The inferences described the causal relationships which existed between RNCC and each of its predictors in the proposed model. For this purpose, each variable coefficient in this model was interpreted in accordance with the model double-log functional form. In such a model, the coefficients of quantitative variables can be interpreted as elasticities, and the coefficients for dummy variables can be interpreted as percentage shifts from the base condition. It was shown that, holding all other variables constant, if the area, number of stories, and weight indicator (measured as the total building seismic weight in ton/m²) increases by 1%, on the average, RNCC rises by approximately 0.67%, 0.27%, and 0.195%, respectively. It was also concluded that locating the building in very high seismic zones and the absence of an earthquake-resistant system in an existing building have an adverse effect on retrofit cost, increasing the mean value of RNCC by almost 24% and 14%, respectively. In addition, the presence of soil types with poor seismic characteristics appeared to impose a similarly unfavourable effect on retrofit cost as did the presence of plan irregularity. Both factors increase the mean value of RNCC by almost 12%.

Although the conclusions mentioned above are all valuable by their own right, the most paramount conclusion of this chapter was derived when the performance of the last seven MLR models that resulted from the BE procedure (i.e., models number one to seven in Table 5-4) was compared over the test dataset. These models were selected because they were all developed using the most statistically significant independent variables. The measure used for this comparison was the determination of coefficient ($R^2$) with the higher value
being superior. The results revealed that the first MLR model outperformed other MLR models as it offered the largest value of $R^2$. This model is therefore more able to reduce the difference between the predicted and actual values of RNCC in future retrofit projects. Besides its reliability, this model is also in agreement with the principal of parsimony because it enables the end user to predict the RNCC of a project with the least information about the project; that is, building area whose value is almost always known before a project commences. The importance of this issue is more pronounced in cases where no detailed information about the building characteristics, site conditions, and also retrofit solution plans is available at the time of prediction.

Using calculus, it was argued for the first time in the literature that the relationship between retrofit cost and area can be expressed in the form of $C = KA^B$, where $C$ is the retrofit net construction cost in thousands U.S$, $A$ is the area in square meter, $K$ is a constant describing the general level of cost performance for a $1,000 retrofit project, and $B$ is a constant describing how the cost performance is affected by project size as measured by building area.

As shown in this chapter, the statistical regression analysis can model the linear relationship between dependent and independent variables. It should be noted that although double-log regression models, like those utilized in this research, are nonlinear in variables but they are still linear in the coefficients. Therefore, it may be more realistic to model a complex decision problem, like the one relating to cost estimation, using other nonlinear techniques, such as artificial neural network (ANN) technique. The description of this technique forms the core of the next chapter.
CHAPTER 6

ARTIFICIAL NEURAL NETWORK (ANN) METHODOLOGY

6.1 Introduction

Throughout the years, the computational changes have brought growth to data analysing (Ostberg, 2005). This growth led to the adoption of new and more elaborate modelling methodologies that may provide more accurate means of analysing data (Bullock, 1999). Artificial neural networks (ANNs) are one of the fastest growing and most innovative areas of intelligent computing techniques, which have recently begun to attract well-deserved attention. ANNs have recently emerged as the methodology of choice for many applications in a wide range of disciplines such as mathematics, system science, engineering, business, financial analysis, risk assessment, and managerial decision-making (Garson, 1998; Haykin, 1999; Lawrence 1994; Scarborough, 1995). ANNs have acquired a broad credibility in such a vast array of disciplines because they are robust classifiers, good pattern recognition engines, innovative monitoring and control tools, and powerful predictors.

The ANN methodology is described in this chapter. The chapter starts with an overview to the nature of the ANN methodology and a historical background of the development of this methodology is provided. Paradigms upon which ANN models can be established are then presented and described in detail. The chapter is concluded with an explanation of advantages and disadvantages of utilising the ANN methodology.

6.2 Biological Neuron vs. Artificial Neuron

ANNs are basically computer simulations of complex neuronal systems such as those found in human brain tissues (Ostberg, 2005). Through the imitation of human brain, an ANN has the capability of processing, learning, and storing information gained from its environment and applying the resulting knowledge to new problems. In the human brain, the actual information processing is performed by a simple element, which is the neuron. The neuron has two types of out-reaching tree-like branches (see Figure 6-1): the dendrites and the axon (Jain et al., 1996). In a biological neuron, electro-chemical input signals are received from other neurons through a host of fine input structures, called dendrites. The neuron then
processes the incoming signals by means of the net ionic activity, resulting in a release of an electrical pulse which is sent out through its axon. The electrical signal is then converted to a chemical signal and is passed across a synaptic junction to the dendrites of a receiving neuron (Garson, 1998; Lawrence, 1994).

![Figure 6-1: Biological neuron (Versace and Chandler, 2010)](image)

Just as a biological neural system, ANNs are composed of information processing neurons that are linked through weighted interconnections which adapt to external data (Ostberg, 2005). These artificial neurons receive and process vectors of numbers rather than electro-chemical signals. Within the body of the artificial neuron, mathematical summation and activation functions simulate the ionic activity, and the effect of synaptic junctions on incoming signals is represented by connection weights (Ostberg, 2005). However, the resulting network of artificial neurons is relatively simple and is nowhere near as complex as the neural networks in the human mind that they are supposed to represent (Doig, 1999).

Processing within a single artificial neuron can be represented by a model of one of the first artificial neurons, introduced by McCulloch and Pitts in 1943. This model is depicted in Figure 6-2. In this model, inputs \((x_1, x_2, ..., x_n)\) are multiplied by their respective connection weights \((w_1, w_2, ..., w_n)\) and then summed in the first half of the neuron by a summation function. The summed value is then processed in the second half of the neuron by a mathematically discrete activation function to determine the level of output (Ostberg, 2005).
In this model, the neuron computes the output according to whether the weighted input sum enumerated in the first half of the neuron is above or below a certain threshold $T$. The computation in the $i^{th}$ neuron can be expressed by the following equation (Jeon, 2007):

$$y_i(t + 1) = a\left[\sum_{j=1}^{n} w_{ij}x_j(t) - \theta_i\right]$$  \hspace{1cm} (6.1)

where $t$ is the time instants; $n$ is the number of the inputs to the $i^{th}$ neuron; $y_i$ is the output from the $i^{th}$ neuron. The transfer function $a(f)$ is a binary step function as follows:

$$a(f) = \begin{cases} 
1 & \text{if } f \geq 0 \\
0 & \text{otherwise}
\end{cases}$$  \hspace{1cm} (6.2)

![Diagram of McCulloch-Pitts neuron](image)

Figure 6-2: McCulloch-Pitts neuron (Wasserman, 1989)

### 6.3 Historical Background of ANN

In more recent years, the utilization of the ANN methodology has been going beyond the scopes of academic inquiries. Commercial, yet more practical, use of this analytical
technique is now being explored by the industry in the real world. However, despite the current attention of both researchers and practitioners, the ANN methodology has a considerable history and has been developed for several decades.

According to Eberhart and Dobbins (1990), the history of the ANN methodology can be divided into four periods of time: initial period (1890-1969); depression period (1969-1982); recovery period (1982-1986); heightened interest period (1986-present). Liu (1998) reviewed the significance and effects of each period and described the attainments achieved during each era in the progression of the ANN methodology.

As it is mentioned in previous section, McCulloch and Pitts in 1943 suggested that a network of simple binary neurons could perform highly complex computations. These neurons were presented as models of biological neurons and as conceptual components for circuits that could perform computational works. Although the neuron model could perform logical processing, but it was not really understood how information was stored in the model and how intelligent the knowledge was learned. Moreover, the McCulloch and Pitts model contains a number of simplifying assumptions that do not reflect the true behaviour of biological neurons (Jain et al., 1996). These questions remained unanswered for a few years, until Hebb (1949) assumed that information was stored in the connections between neurons and the learning could be attributed to the process of modifying these connections. Currently, large extents of learning paradigms for ANN models are based on the modifications made to the Hebb’s original concepts.

In 1958, Rosenblatt made another major contribution to the ANN methodology with the development of the perceptron which was the first real artificial neural network. The perceptron provided a simple model permitting extensive mathematical analysis of neural networks. ANN development was then continued in 1960 by Widrow and Hoff, who developed Widrow-Hoff algorithm by which the speed of learning and the accuracy of results were improved. Minsky and Pappert in 1969 conducted an in depth mathematical analysis of the perceptron. Using simple examples, they reported that only a few functions are able to be learned by perceptron. Hence, their work had a discouraging effect on ANN applications and dampened the enthusiasm of most researchers.

The resulting lull in neural network research lasted for several years till the early 1980s. In 1982, Hopfield introduced entirely connected networks called Hopfield networks.
Hopfield’s work laid the foundation for significant advances in the ANN methodology by overcoming the objections, presented by Minsky and Pappert.

The key development in the history of the ANN methodology occurred in the middle of 1980s. In 1986, Rumelhart et al., (1986) popularized a learning algorithm, originally derived by Werbos in 1974 in his PhD research in statistics at the University of Harvard (Abou-Elseoud, 1998; Jain et al., 1996). Rumelhart et al., (1986) took advantage of the feed-forward network’s architecture and also the non-linear feature of artificial perceptrons to produce an easy understood, yet, highly effective training paradigm known as back-propagation learning algorithm. The power and generality of feed-forward networks were known for many years, but prior to the development of the back-propagation learning algorithm there was no efficient and theoretically sound method for training the weights of connections embedded in such networks (Wasserman, 1993).

Since 1986 onwards, the feed-forward networks have become the most popular network amongst both researchers and users of neural networks (Jain et al., 1996). According to Gately (1996), the reason behind this popularity can be in fact attributed to the development of the back-propagation learning algorithm.

### 6.4 ANN Principal Elements

ANNs are information processing systems that abstractly simulate the decision making process of the human brain. Resembling a human brain system that contains a collection of neuron units, ANNs consist of a large number of simple mathematical processing elements (PEs), which are called artificial neurons. These neurons are able to adapt to external data (Ostberg, 2005). The artificial neurons form the heart of ANNs. The functions associated with the artificial neurons provide an ANN the ability to model a wide variety of relationships between input and output variables (Stockton and Wang, 2004; Wang et al., 2000). Each neuron is a simple computational device which has a small amount of local memory and some computing power. Being connected to other neurons, a neuron receives numerous input signals from other neurons, processes them, and generates only one output signal. This output signal is then transmitted as an input signal to many other interconnected neurons. Hence, analogous to biological nerve cells, neurons in an ANN model are able to
learn and remember features and relationships within their environment and subsequently apply this knowledge to new data (Scarborough, 1995).

Like a real nervous system in which biological neuron cells transmit their signals to each other via axon links, artificial neurons within an ANN model also communicate with each other thorough particular pathways which are known as “connections”. In an ANN model, a particular weight is assigned to each connection. The strength of each neuron’s output is determined by the weight of connections between that particular neuron and the neurons in the subsequent layer. The significance of these weights is that the effect of irrelevant inputs is minimized by assigning small weights to their connections, whereas the effect of inputs which have a significant impact on the output of the network is maximized by assigning higher weights to their connections (De la Garza and Rouhana, 1995). Therefore, the performance of a network is highly dependent on these weights wherein the knowledge about a given problem is actually stored. To elevate the network performance to an acceptable level, the network is trained by adjusting the values of connection weights.

From the discussion above, it can be concluded that an ANN model consists of two rudimentary elements:

- Neurons or processing elements (PEs)
- Connections

Neurons and their weighted connections are conceptual components which form circuits within an ANN model that could perform computational tasks. These components are the principal elements, which enable an ANN model to be flexible and desirably map the nature of a problem, which is usually difficult to be solved by utilizing conventional analytic methods.

6.5 ANN Specification

The development of an ANN model is highly problem dependent. What is best for one problem may not be best for another (Abou-Elseoud, 1998). Developing an ANN model requires the determination of an ANN paradigm. This paradigm allows an ANN model to mimic the functions of a human brain, and as a result enables the model to exhibit certain features such as learning, recalling, and generalizing of what have been learnt. An ideal
paradigm must optimally describe the nature of a problem, and must perform well according to certain criteria.

An ANN paradigm is mainly defined by three components:

1- Architecture (topology)
2- Transfer or activation function of artificial neurons
3- Learning (training) algorithm

Architecture is the configuration of an ANN model that arranges the neurons and their connections in a number of layers. Transfer or activation function controls the amplitude of the output of each neuron in the model. Further, learning algorithm is the process of computing the most appropriate weights of connections among neurons.

In the following sections, a thorough review of the above components is presented. A special emphasis is given to the multilayer feed-forward networks, sigmoid logistic function, and the back-propagation learning algorithm as the network architecture, activation function and the learning algorithm, respectively. These are the paradigms utilized in this study.

6.5.1 Architecture (Topology)

A neural network is a type of information processing system whose architecture is inspired by the structure of human biological system (Caudill and Butler, 1990). Similar to the architecture of a real biological neural network, an ANN has a parallel distributed information processing architecture (Fausett, 1994; Lin and Lee, 1996) with a large number of neurons and connections which are arranged in an organized configuration to carry out an extensive computing, and to perform a mathematical mapping (Jeon, 2007). The configuration of a network is defined by its architecture. Basically, neurons in a network are grouped in different layers. The architecture of an ANN model is represented by the number of these layers in the model, the size of each layer in terms of the number of neurons in each layer, and the way the neurons within the model are communicated or connected to each other (Zurada, 1992). The architecture of neural networks can take on an infinite number of forms. However, on the basis of interconnections between the neurons in different layers, the architecture of an ANN model can usually be classified into two broad categories, being feed-forward network and feed-back network which is also known as recurrent network (Abou-
Elseoud, 1998; Jain et al., 1996; Liu, 1998). Figure 6-3 illustrates these categories. Typical networks for each category are also shown in this figure.

In a feed-forward network, a neuron’s output is never dependent on the output of subsequent neurons. This independency means that neurons in a particular layer are connected only to those neurons in the subsequent layer. The neurons of one layer are not connected neither to the neurons in the previous layer nor to the neurons in the same layer. As a result, the signals flow only in one direction which, in turn, causes every neuron in a given layer to send its output only to neurons in the next layer and to receive inputs merely from neurons in the previous layer. Examples of feed-forward networks include the multilayer perceptron network (MLP) and radial basis function (RBF) network.

In contrast to a feed-forward network, a feed-back network may harness at least one feedback loop in which the outputs of specific neurons are sent back as inputs to other neurons (Haykin, 1999). So, in feed-back networks any neuron can be connected to any other neurons in the network and signals can therefore flow in both forward and backward directions. Examples of the feed-back networks include the Hopfield network (Hopfield, 1982), the Elman network (Elman, 1990) and the Jordan network (Jordan, 1986).
According to Jain et al., (1996), feed-forward networks are static as they produce only one set of output values rather than a sequence of values from a given input. Feed-forward networks are memory-less in the sense that their response to an input is independent of the previous network state. Alternatively, recurrent or feed-back networks are dynamic systems. They have a dynamic memory such that their inputs at a given instant reflect the current input as well as previous inputs and outputs. Consequently, when a new input pattern is presented, the neuron outputs are first computed and because of the feed-back paths, the inputs to each neuron are then modified, which leads the network to enter a new state.

Feed-forward neural networks can estimate highly non-linear functional forms and are a universal functional approximator (Funahashi, 1989; Hornik et al., 1989). The most popular and widely in use form of the feed-forward architecture is known as multilayer perceptron (MLP) (Hegazy et al., 1994; Jain et al., 1996). One of the most significant advantages of MLP networks is their generalization and classification capability (Caudill, 1992; Hegazy et al., 1994; Liu, 1998; Maren et al., 1990; Rumelhart et al., 1986). Another advantage of MLP networks is their ability to deal with complex systems and to approximate any non-linear function (Cybenko, 1989; Funahashi, 1989; Haykin, 1999; Hornik et al., 1989; Hush and Horne, 1993; Wilmot and Mei, 2005). These advantages entitle the MLP network to be the most preferred architecture.

MLP networks are relatively easy to construct. A MLP network typically has one input layer, one or several hidden layers, and an output layer. These layers are successively connected in a feed-forward fashion with no connections between neurons in the same layer and no feedback connections between layers. Figure 6-4 depicts a sample of a MLP network with three layers and p, m, and n nodes in input, hidden, and output layer, respectively. Each of the nodes in the network is a neuron and each of the links between the nodes represents a connection. As shown in Figure 6-4, the connections go from the neurons of a certain layer toward the neurons of the following layer and there is no feedback between layers. Therefore, the information is always transmitted from the input layer toward the output layer. In addition, the network shown in Figure 6-4 is said to be fully connected in the sense that every node in each layer of the network is connected to every other node in the adjacent forward layer. However, if some of the communication links are missing from the network, it is said that the network is partially connected. Fully connected MLP networks have been reported in the literature to perform better than the other type (Bailey and Thompson, 1990b; Hegazy et al., 1994).
Selecting an appropriate architecture for the network should receive a well attention because this selection is amongst the most important issues that affect the final model accuracy and its generalization ability. However, little evidence is available to suggest that a specific ANN architecture could be used over a range of applications. To determine the best network architecture that best fits the application under consideration, examining different architectures is inevitable. In particular, the issue that deserves a careful attention is to define the network’s layers, in addition to the number of neurons on each layer.
6.5.1.1 Input and Output Layers and Neurons

Neurons are usually organized into groups called layers, with full or partial connections between successive layers. There are typically two layers that possess connections to the outside world, i.e., an input layer where data are presented to the network, and an output layer in which the solution of the problem takes place. Layers distinct from input and output layers are called hidden layers through which the information is processed.

The number of input and output neurons is determined by the nature of the problem under investigation. The input neurons represent model parameters and accept the data that are presented to the network, whereas the output neurons produce the network results. Consequently, the number of input and output neurons is the same as the number of independent and dependent variables, respectively. Furthermore, the numbers of input and output neurons are not restricted, which is an advantage of ANNs (Garza and Rouhana, 1995; Smith and Mason, 1997). For instance, the MLP network shown in Figure 6-4 is intended to predict \( n \) dependent variables (i.e., \( Y_1... Y_n \)) based on the contribution of \( p \) independent variables (i.e., \( X_1, X_2... X_p \)).

6.5.1.2 Hidden Layers

In an ANN model, there is no direct connection between the input and output layers. Rather, between these layers is the middle or hidden layers which perform the most important function of the network; that is, mapping out the proper connection between input and output variables (Swicegood, 1998). Consequently, the job of hidden layers is to act as an intermediate data abstraction between input and output neurons (Attalla, 1999).

There are normally one or more hidden layers between the input and output layers. However, it was reported that one hidden layer is sufficient to model most applications (Ostberg, 2005; Setyawati et al., 2002; Tamura and Tateishi, 1997). It was also suggested that the use of more than one hidden layer in the network may not be necessary or practical, particularly if problem inputs are chosen as the most important inputs, which adequately represent the problem space (Hegazy et al., 1994). This suggestion is made because the features that the hidden layers are normally used to detecting are directly presented as inputs (Caudill, 1990; Dutta and Shekhar, 1988). Moreover, the universal approximation theorem (Bose and Liang, 1996; Haykin, 1999) states that a single hidden layer is sufficient for MLP
networks to approximate any continuous function. Nevertheless, for some problems, a second hidden layer of neurons between the first hidden layer and the output layer can be used (Chester, 1990; Jeon, 2007; Kecman, 2001). In these problems, using a number of hidden neurons does not meet to solve the problem with one hidden layer or a network with one hidden layer would require an infinite number of neurons (Jeon, 2007). Further, a network with continuous-value inputs may require more than one hidden layer to achieve any arbitrary mapping (Chester, 1990). Finding the most appropriate number of hidden layers is determined by following a trial and error approach. In this regard, a simple heuristic rule of thumb is proposed by Bailey and Thompson (1990b). They suggested to start with one hidden layer and adding more as required. However, it is believed that there is no need to have more than two hidden layers, because including more hidden layers, while increasing the number of training examples which are necessary to properly train the network (Hegazy et al., 1994), would not improve the results (Jeon, 2007).

Surkan and Singleton (1990) came to the conclusion that network performance could be enhanced by adding additional hidden layer. This conclusion is supported by Swales and Yoon (1992), who found that ANN models with multiple hidden layers performed better than the networks, which were composed of a single hidden layer. The same finding was also reported by Yoon et al., (1993) as they detected that the increase in the number of hidden layers may result in an improvement in the performance of ANN models. More recently, Attalla and Hegazy (2003) studied various network architectures to arrive at the best ANN model. They concluded that a particular network, known as Ward Net, with 3 hidden layers produced the best predictive results in their study. In contrast to the studies mentioned above, Dutta and Shekhar (1988) argued that the inclusion of more than one hidden layer had little effect on the network performance. The same argument was also made by Shtub and Zimerman (1993). Their research was aimed at selecting the best network architecture in terms of the number of hidden layers. Three types of architecture and a total of ten possible networks were studied including (1) one network with no hidden layer; (2) four networks with one hidden layer; and (3) five networks with two hidden layers. The results showed that an architecture type with one hidden layer performed better than network with no hidden layer and networks with two hidden layers. Emsley et al., (2002) studied a number of different networks in order to assess the best approach to the ANN modelling. These networks included three- and four-layer MLP networks, RBF networks, and generalized regression neural networks (GRNNs). Of these alternatives, three-layer MLP networks (with
one hidden layer) offered the best performance, in terms of the associated values of $R^2$ and mean absolute percentage error (MAPE). More recently, Chen and Huang (2006) examined 0, 1, and 2 hidden layer(s) for each of the six ANN models upon which it was intended to predict both cost and duration of reconstruction projects contracted by three different parties, namely, central agency, local government, and private sectors. The results illustrated that for every single model studied one hidden layer outperformed its counterpart models with 0 and 2 hidden layer(s).

6.5.1.3 Hidden Neurons

The hidden neurons internally represent the relationships in the data (Hegazy and Ayed, 1998). The approximation capability of an ANN model is primarily defined by the number of hidden neurons (Jeon, 2007). The hidden neurons are able to associate a particular input pattern with the appropriate desired output values (Doig, 1999) which, in turn, enables the model to extract higher-order statistics (Churchland and Sejnowski, 1992). This extraction is particularly valuable when the size of input layer (i.e., the number of independent variables) is large (Abou-Elseoud 1998). Choosing the network size in terms of the number of hidden neurons is an extremely critical step in the design of an ANN model and continues to be one of the most researched topics in ANNs (Abou-Elseoud 1998). Incorrect number of hidden neurons is amongst the reasons accounting for the lack of success in ANN applications. Imprecise training, erroneous architecture, and lack of functional relationship between independent and dependent variables constitute other main reasons of achieving abortive networks (Cotter, 1990; Hornik et al., 1989).

Typically, the more neurons in the hidden layer(s), the more powerful the network (Chu, 1998; Liu, 1998). If enough neurons are not being considered in the hidden layer(s), as compared with the number of training examples, then the trained network would generally appear to be enfeebled in generating outputs that reasonably close to the actual ones (Hegazy et al., 1994; Williams, 1994). The generated outputs are erroneous because during training the network is only able to capture too few distinctive characteristics of the problem. This effect is called under-fitting. Alternatively, having an excessive number of hidden neurons causes the network memorizing the patterns in the training dataset too well instead of learning them. This effect is called over-fitting and leads the trained network to loss its generalization ability. Consequently, the network may fail to produce a correct input-output
mapping for new cases even when these new cases used as a common practice for testing the network are slightly different from the samples used for training the network (Haykin, 1999).

Finding the most appropriate number of hidden neurons per hidden layer is more of an art than science (Shtub and Versano, 1999), as a systematic approach to define this number has not yet been proposed in the literature. The number of hidden neurons is usually determined heuristically through the use of a trial and error approach (Fausett, 1994; Garson, 1998; Hegazy and Ayed, 1998; Kosko, 1992; Williams, 1994). As a result, considerable time must be spent to find the appropriate number of hidden neurons which is one of the pitfalls of ANNs (Bode, 1998; Boussabaine, 1996; Creese and Li, 1995; Hegazy et al., 1994; Kim et al., 2004; Li, 1995; Liu, 2006; Wang et al., 2000; Yeh, 1998).

Varying the number of hidden neurons in the hidden layer, Dutta and Shekhar (1988) found that the number of hidden neurons had only a minor effect on overall performance of the networks. Similarly, Surkan and Singleton (1990) observed that the number of hidden neurons in a hidden layer could be varied over a wide range without influencing the network performance. The same observation was reported in a study by Shtub and Zimerman (1993), who showed that networks with one hidden layer performed almost the same regardless of the number of hidden neurons considered for the networks. More recently, Bode (2000) and Jeong (2004) studied the impact of the number of hidden neurons on the prediction accuracy of ANN models and reported that the ANN models were not substantially influenced by this number. In contrast to the findings of the above studies, Yoon et al., (1993) concluded that the increase in the number of hidden neurons, up to a certain limit, could advance the networks accuracy. Shtub and Versano (1999) concluded that the optimum number of hidden neurons, by which the best predictive performance was obtained, varied between the ANN models they studied. This conclusion was further supported in a more recent study by Liu (2006), who investigated the relationship between the number of hidden neurons and an ANN model performance. The results of this study showed that this relationship is uncertain.

As stated earlier, the appropriate number of hidden neurons still remains controversial and varies by the nature of the problem to be solved. No formal quantitative method has been widely accepted as the optimal technique for determining the best number of hidden neurons. However, in the literature there are some certain equations recommended upon which the number of hidden neurons can be initialized. These equations that have been commonly proposed for MLP networks are summarized in Table 6-1.
Table 6-1: Recommendations for setting number of hidden neurons

<table>
<thead>
<tr>
<th>Number of hidden neurons</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0.75 \times m$</td>
<td>Bailey and Thompson (1990b)</td>
</tr>
<tr>
<td></td>
<td>Salchenberger et al., (1992)</td>
</tr>
<tr>
<td></td>
<td>Hegazy et al., (1994)</td>
</tr>
<tr>
<td>$2 \times \sqrt{m+n}$</td>
<td>NeuroShell 1, User Guide (1989)</td>
</tr>
<tr>
<td>$2 \times m + 1$</td>
<td>Hecht-Nielsen (1990)</td>
</tr>
<tr>
<td></td>
<td>Caudill (1991)</td>
</tr>
<tr>
<td></td>
<td>Hegazy et al., (1994)</td>
</tr>
<tr>
<td>In a range from $(2 \times \sqrt{m+n})$ to $(2 \times m + 1)$</td>
<td>Fletcher and Goss (1993)</td>
</tr>
<tr>
<td>$\frac{m+n}{2}$</td>
<td>Hegazy and Ayed (1998)</td>
</tr>
<tr>
<td>$\frac{m+n}{2} + \sqrt{p}$</td>
<td>Al-Rashidi (1999)</td>
</tr>
</tbody>
</table>

$m$: number of input neurons (i.e. independent variables)

$n$: number of output neurons (i.e. dependent variables)

$p$: number of samples in the training dataset

6.5.2 Transfer or Activation Function

The neural computation allows the development of information processing for which the rules and relationships knowledge are not available (Hecht-Nielsen, 1990). Therefore, it is essential to understand the mechanism how a neuron processes information from several inputs and then converts it into an output. Figure 6-5 describes this mechanism for the $k^{th}$ hidden neuron in Figure 6-4. As illustrated in Figure 6-5, there are three basic components of importance enabling the neuron to operate. First, the strength of connections between neurons is modelled by a numerical value, known as weight. If an input signal has low predictive value, the pathways that connect it to other neurons will be assigned low numerical weights and thus it will have a minor role in the determination of the outcome (Doig, 1999). In Figure 6-5, $w_{k1}, w_{k2}... w_{kp}$ are the weights of connections which respectively plug $X_1, X_2... X_p$ inputs into the $k^{th}$ hidden neuron. $w_{k0}$ is also an externally applied threshold, known as bias weight. Each of these connection weights can be either negative or positive. Negative weight values reflect inhibitory connections, while positive values reflect excitatory connections (Haykin, 1999; Jeon, 2007; Swicegood, 1998). If $w_{kj}=0$, then there is no connection between the neurons $j$ and $k$. The next two components model the actual activity within the neuron cell.
An adder sums up all the inputs, which are modified by their respective weights. Finally, the neuron produces its output ($y_k$ in Figure 6-5) by applying an activation function to the sum of the weighted values ($u_k$ in Figure 6-5). This output is then transmitted to the neurons of the next layer as one of their incoming inputs.

As shown in Figure 6-4, each neuron in the hidden and output layer has an activation function that is used to compute the output signal from the input signals in a similar way as described above. The activation function controls the amplitude of the output of the neuron and allows the network to reach its best performance in generalization capability (Liu, 1998). It is also believed that ANNs are best suited to problems which involve complex non-linear
relationships because of their adaptivity owing to activation functions (Flood and Kartam, 1994).

Mathematically, activation functions are used in ANN models to map unconstrained numerical data into a pre-specified bounded space. An acceptable range of the numerical bounded space is usually between 0 and 1, or -1 and 1. There are several types of activation functions which can be represented in a linear or non-linear format. However, there are some activation functions which are more used by researchers.

The most widely used non-linear activation function is the S-shaped logistic activation function (Abou-Elseoud 1998; Doig, 1999; Hansen and Messier, 1991; Haykin, 1999; Jain et al., 1996; Liu, 1998; Ostberg, 2005; Swicegood, 1998; Tam and Kiang, 1990; Trippi and Turban, 1996), which is also known as “binary sigmoid function”. It has been stated that the MLP networks using the logistic activation function can approximate any continuous non-linear function provided that a sufficient number of hidden neurons are also available (Cybenko, 1989; Funahashi, 1989; Hornik et al., 1989; Hush and Horne, 1993). The logistic activation function is determined by Equation (6.3) and it is used to map the input values from the interval (-∞, +∞) into the interval (0, 1). So, the output is always between 0 to 1 (see Figure 6-6).

\[ y = f(x) = \frac{1}{1 + \exp(-x)} \]  

(6.3)

Figure 6-6: Logistic activation function
Another popular activation function which is similar to the logistic activation function is the hyperbolic tangent activation function, known also as “bipolar sigmoid function”. As opposed to the logistic (0, 1) output range, the hyperbolic tangent activation function squeezes the input values from the interval \((-∞, +∞)\) into a range of \((-1, +1)\) (see Figure 6-7). The hyperbolic tangent function is mathematically defined by the following equation:

\[
y = f(x) = \tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)}
\]  

(6.4)

![Hyperbolic tangent activation function](image)

Figure 6-7: Hyperbolic tangent activation function

According to Schalkoff (1997), the reasons behind the importance and popularity of the activation functions mentioned above are as follows:

- They are semi-linear, monotonically increasing and differentiable everywhere
- They are expressible in the closed form
- Their modifications or extensions lead or relate to other activation functions
- The derivative of their functions with respect to the network is very easy to form
- They have a biological basis.

As an alternative to the S-shaped logistic and hyperbolic activation functions, some modellers may prefer to use activation functions that are radially symmetric, often in a classic bell shape curve, such as the Gaussian function (see Figure 6-8).
6.5.3 Learning Algorithm

Among the many interesting properties of neural network, the ability of the network to learn from experience and to improve its performance through learning is of primary significance (Abou-Elseoud 1998). ANNs can learn through a network of massively interconnected yet simple mathematical processing elements (Haykin, 1999) which enables the network to identify input-output mappings. An ANN model has to be thought such that the application of a set of inputs produces the desired set of outputs. The overall goal of training an ANN model is therefore to minimize the difference between values of actual and predicted output variable. This goal is accomplished through a systematic step-by-step training procedure where the predicted outputs are compared with the corresponding actual outputs and as a result the connection weights are adjusted to minimize the difference between them (Somers, 1999). This systematic training procedure, known also as learning algorithm, teaches the network the likely patterns in the data and accordingly lets the network change its connection weights. The iterative procedure of learning algorithm continues until satisfactory performance, as defined by the modeller, is reached or the network fails to converge on a solution. Figure 6-9 represents the abstract process of training a network.
Several learning algorithms have recently been developed to deal with different problems. These learning algorithms vary a great deal with regard to the type of input patterns they accept, the output patterns they produce and the learning characteristics and the complexities and limitations they exhibit (Jain et al., 1996; Jeon, 2007; Kohonen, 1988; MacKay, 1992). Table 6-2 specifies the relationship between the application requirements and the capabilities of selected learning algorithms. Several of these algorithms (e.g., back-propagation and Boltzmann machine) offer flexibility in terms of size (number of layers), connectivity (how neurons are linked together), and training methods. Other learning algorithms (e.g., ART, Kohonen network, and Hopfield) are more tightly contained in the nature of their topology or training method. However, there is no systematic approach available to determine the best learning algorithm for a particular application. Consequently, the selection should be based on the comparison of the application requirements to learning algorithm capabilities.
### Table 6-2: ANN learning algorithms (Boussabaine, 1996)

<table>
<thead>
<tr>
<th>Learning paradigm</th>
<th>Activation function</th>
<th>Number of layers</th>
<th>Type of input accepted</th>
<th>Size of hidden layers</th>
<th>Connectivity</th>
<th>Learning algorithms</th>
<th>Learning parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Back-propagation</td>
<td>Sigmoid</td>
<td>2 or more</td>
<td>Continuous</td>
<td>Small to medium</td>
<td>Fully interconnected</td>
<td>Generalized delta rule</td>
<td>Learning constant and Momentum</td>
</tr>
<tr>
<td>Counter propagation</td>
<td>Kohonen and Sigmoid</td>
<td>2 or 3</td>
<td>Continuous</td>
<td>Same as Kohonen</td>
<td>Fully interconnected</td>
<td>Kohonen and Grossberg</td>
<td>Kohonen and Outstar</td>
</tr>
<tr>
<td>Madline</td>
<td>Signum</td>
<td>2</td>
<td>Bipolar</td>
<td>Small to medium</td>
<td>Fully interconnected</td>
<td>Delta rule and max, min, majority</td>
<td>Learning constant</td>
</tr>
<tr>
<td>Outstar</td>
<td>Sigmoid</td>
<td>1 continuous</td>
<td>Binary</td>
<td>N/A application</td>
<td>Varies by</td>
<td>Outstar function</td>
<td>Decay time attack</td>
</tr>
<tr>
<td>ART2</td>
<td>Sigmoid</td>
<td>1</td>
<td>Grey scale</td>
<td>Increases b data type</td>
<td>Fully interconnected</td>
<td>ART2</td>
<td>Vigilence, gain</td>
</tr>
<tr>
<td>Kohonen network</td>
<td>Competitive learning</td>
<td>1</td>
<td>Continuous data type</td>
<td>Equals number of interconnected</td>
<td>Fully interconnected</td>
<td>Kohonen size, alpha</td>
<td>Neighborhood</td>
</tr>
<tr>
<td>Boltzmann machine</td>
<td>Varies</td>
<td>2 or more</td>
<td>Binary continuous</td>
<td>Small to medium</td>
<td>Fully interconnected, random</td>
<td>Boltzmann Temperature</td>
<td></td>
</tr>
<tr>
<td>Hopfield network</td>
<td>Hard limiting</td>
<td>1</td>
<td>Binary bipolar</td>
<td>N/A</td>
<td>Fully interconnected</td>
<td>Hopfield</td>
<td>None</td>
</tr>
<tr>
<td>Adaline</td>
<td>Signum</td>
<td>1</td>
<td>Bipolar</td>
<td>N/A</td>
<td>Fully interconnected random</td>
<td>Delta rule</td>
<td>Learning constant</td>
</tr>
<tr>
<td>Bam</td>
<td>Clamped linear</td>
<td>1</td>
<td>Bipolar</td>
<td>N/A</td>
<td>Fully interconnected</td>
<td>Bam</td>
<td>Retention, gain</td>
</tr>
<tr>
<td>Perceptron</td>
<td>Perceptron</td>
<td>1</td>
<td>Continuous</td>
<td>N/A</td>
<td>Random</td>
<td>Perceptron convergence</td>
<td>Learning constant</td>
</tr>
</tbody>
</table>

As stated earlier, the main objective of training is to update the connection weights to reduce the errors between the predicted and actual output values to a satisfactory level (Jeon, 2007). This update continues until the trained network becomes able to successfully generalize on new cases (Hegazy et al., 1994). The errors between the predicted and actual
output values determine which neuron in the network should be modified and how. Any learning algorithm aims therefore to find the most appropriate weight values by which the trained network can reach an optimal performance. In this performance, the network outputs are as close as possible to their corresponding actual outputs such that the generalization ability of the network is also secured.

Training a network should be principally accomplished based on certain learning rules. These learning rules intend to obtain the weights which optimize a predefined criterion function which is also commonly referred to as an error function. This function determines the nature of a learning rule (Abou-Elseoud, 1998). Learning rules can be generally categorized into three broad classes as follows (Abou-Elseoud, 1998; Bailey and Thompson, 1990a; Haykin, 1999; Jeon, 2007; Liu, 1998; Ostberg, 2005):

- **Supervised Training** or **Associative Training**: Supervised training is a process that incorporates an external teacher which provides a network with a training dataset which is completely comprised of inputs and corresponding actual outputs. A teacher is capable of applying exact corrections to the network outputs when an error occurs. The network processes the inputs and consequently generates the outputs. The generated outputs are then compared to the actual outputs and their difference is calculated. The difference would then serve as a means to accordingly adjust the connection weights. Usually, the weights are synthesized gradually, and at each step of the learning process they are updated so that the error between the network’s output and a corresponding desired output is reduced. Supervised training has established itself as a powerful learning algorithm for the design of ANNs. Examples of supervised learning algorithms include the delta rule (Widrow and Hoff, 1960) and the generalized delta rule (Rumelhart et al., 1986).

- **Unsupervised Training** or **Self-organization Training**: Unsupervised training is a process that does not incorporate an external teacher and relies upon local information and internal control. In this training method, the network learns as a student by its own by interacting with the data and without the assistance of a teacher. Unlike supervised training, the actual outputs of the training sample are unavailable and, therefore, the learning must to some extent be accomplished based on observations of response to inputs. During training, the network detects regularities and separates properties. Repetitive exposure to sample data causes network connection weights to
converge to a state that best represents the characteristics of the sample data. Hence, in unsupervised training, the network is supposed to discover statistically salient features of the input population and the network must develop its own representation of the input stimuli as there is no feedback from the environment to indicate what the desired outputs of a network should be or whether they are correct. Examples of unsupervised training include the Kohonen (Kohonen, 1989), and Carpenter-Grossberg Adaptive Resonance Theory (ART) (Carpenter and Grossberg, 1988).

- **Reinforcement Training or Hybrid Training**: Reinforcement training is a type of training that may be considered as an intermediate form of the above two types of training. In the reinforcement training, a network is given a scalar evaluation signal instead of being provided with the actual output and evaluation can also be made periodically instead of with every input sample in the training dataset. The training system does some actions on the environment and gets a feedback response from that environment. Subsequently, the training system grades its action good (rewarding) or bad (punishable) based on the environmental response and accordingly adjusts its parameters. The parameter adjustment is continued until an equilibrium state occurs; following which there will be no more changes in the training parameters. An example of a reinforcement training rule is the genetic algorithm (GA) (Goldberg, 1989).

### 6.5.3.1 Back-propagation Learning Algorithm

The most common learning algorithm used in conjunction with MLP networks is called back-propagation learning algorithm developed by Rumelhart (Rumelhart et al., 1986). It is estimated that over 80% of all ANNs in development use the back-propagation learning algorithm (Rao and Rao, 1995). Besides having gained such a broad credibility to become the first choice of artificial network developers (Caudill & Butlet, 1992), the back-propagation learning algorithm has been shown to be more efficient than other learning algorithms in solving problems. For instance, by evaluating different learning algorithms including back-propagation, conjugate gradient descent, Levenberg Marquardt, quick propagation, delta-bar-delta and quasi-Newton training algorithm, Emsley et al., (2002) concluded that the ordinary back-propagation learning algorithm was the most efficient learning algorithm.
The back-propagation learning algorithm, known also as the gradient descent learning algorithm, is a supervised training where an error function is used to teach the network. Initially, a set of mapping data with input and output variables is fed to the network. Then, the back-propagation learning algorithm iteratively modifies the connection weights to ensure all inputs are correctly mapped to their corresponding outputs. The weights that minimize the error function are pursued to be used in the final network. The error function used in the back-propagation learning algorithm is the mean square error of the network. This error function is originally taken from the least mean square error algorithm, also known as the “delta rule” or “Widrow-Hoff rule”, and is common to many other learning algorithms (Hegazy et al., 1994; Liu, 2006). Similar to the residuals found in multi-linear regression (MLR) analysis, the errors in the ANN methodology are the difference between the outputs predicted by the network and the actual outputs.

During the back-propagation learning algorithm, the neural network simply processes input-output patterns in a procedure, which consists of two passes through different layers of the network (Abou-Elseoud 1998; Jeong, 2007; Liu, 2006; Williams, 1994). This procedure is schematically shown in Figure 6-10. The procedure is repeated over the training dataset until the difference between the outputs generated by the network and the actual outputs converges to a predetermined error threshold (Caudill and Butler, 1990).

- **Forward Pass:** The learning process consists of an ordinary forward pass of the data, in which the input pattern is applied to the network and its effect is propagated forward through the network, layer by layer, to reach the output layer. In the forward pass, each neuron gets the sum of weighted inputs from a previous layer and performs transmission based on the given activation function with the addition of the bias for that neuron. After performing the activation function, the output for this neuron is sent to the neurons in the following layer. The same procedure is performed in the neurons of the following layers based on the designated weights and activation functions until the data transmission reaches the last layer, which is the output layer. At this stage, the output pattern is calculated by the network and then compared to the actual output pattern. Upon the completion of this stage, the system has to decide whether further learning is required. This decision is accomplished by comparing the obtained total difference between the actual outputs and the network generated outputs with an acceptable error threshold given by the network developer. If the output pattern generated by the network is correct, then no adjustments are made to the weights and
consequently no further learning occurs. Alternatively, an incorrect output pattern from the network necessitates adjustments in the weights.

- **Backward Pass:** The back-propagation learning algorithm adjusts the connection weights between layers to make the network outputs move closer to the actual outputs. This adjustment is done in accordance with the error correction rule which aims to minimize the mean squared error between the actual and generated outputs. Through the implementation of the backward pass, the error in the output layer is transformed by the derivative of the activation function and is then used in the previous layer to adjust its connection weights. The error is then propagated backward into previous layers, one layer at a time and against the direction of connections to update the weights of the neurons and their connections in the same architecture. The process of back propagating the network error continues until the first layer (i.e., input layer) is reached.

![Back-propagation learning procedure](image)

*Figure 6-10: Back-propagation learning procedure*
In order to better understand the back-propagation learning algorithm, it is of benefit to describe this learning algorithm in mathematical terms. According to Williams (1994), the relevant equations for the back-propagation learning algorithm can be summarized as follows:

1. Apply the input vector to the input neurons:

\[ X_p = (X_{p1}, X_{p2}, \ldots, X_{pN})^T \]  

(6.5)

\( N \) is the number of neurons on the input layer.

2. Calculate the network input values to the hidden layer neurons:

\[ net_{p_j}^h = \sum_{i=1}^{N} w_{ji}^h x_{pi} + \theta_j^h \]  

(6.6)

where \( w_{ji}^h \) = the weight on the connection from the \( i^{th} \) input neuron to the \( j^{th} \) hidden neuron; \( \theta_j^h \) = a bias term that permits a more rapid convergence of the learning process; and the superscript \( h \) refers to quantities on the hidden layer. The bias term works as a horizontal shift for the origin of the activation function to suit the magnitude of signals incoming to a given neuron. A bias neuron is usually added to the input and hidden layers to facilitate network processing (Hegazy et al., 1994) (see Figure 6-10). Being connected to successive layers, these bias neurons are treated as regular neurons having commonly inputs of 1.0 (Hegazy and Ayed, 1998).

3. Calculate the outputs from the hidden layer:

\[ i_{pj}^h = f_j^h (net_{pj}^h) \]  

(6.7)

where \( f_j^h (net_{pj}^h) \) = the activation or transfer function considered in the hidden layer.

For instance, by applying the logistic activation function, the output of a hidden neuron is
calculated as follows. As stated earlier, this activation function compresses the range of outputs between 0 and 1.

\[ i_{pj} = f_j^h(\text{\textit{net}}_{pj}^h) = (1 + e^{-\text{\textit{net}}_{pj}^h})^{-1} \]  
\[ (6.8) \]

4. Calculate the network input values to the output layer neurons:

\[ \text{\textit{net}}_{pk}^o = \sum_{j=1}^{L} w_{kj}^o i_{pj} + \theta_k^o \]  
\[ (6.9) \]

where \( w_{kj}^o \) = the weight on the connection from the \( j^{th} \) hidden neuron to the \( k^{th} \) output neuron; \( \theta_k^o \) = a bias term; \( L \) is the number of neurons in the hidden layer; and the superscript \( o \) refers to quantities on the output layer.

5. Calculate the outputs on the output layer:

\[ o_{pk} = f_k^o(\text{\textit{net}}_{pk}^o) \]  
\[ (6.10) \]

where \( f_k^o(\text{\textit{net}}_{pk}^o) \) = the activation function considered in the output layer.

6. Calculate the error for the output neurons:

\[ \delta_{pk}^o = (y_{pk} - o_{pk}) f_k^\phi(\text{\textit{net}}_{pk}^o) \]  
\[ (6.11) \]

where \( f_k^\phi(\text{\textit{net}}_{pk}^o) \) = the derivative of the output layer activation function; and \( y_{pk} \) = the actual output value.

7. Calculated error terms for the hidden layer:

\[ \delta_{pj}^h = f_j^h(\text{\textit{net}}_{pj}^h) \sum_k \delta_{pk}^o w_{kj}^o \]  
\[ (6.12) \]

where \( f_j^h(\text{\textit{net}}_{pj}^h) \) = the derivative of the hidden layer activation function.
Every weight update on the hidden layer depends on the error term $\delta_{pk}$ on the output layer. The known errors of the output layer are then back-propagated to the hidden layer to determine the weight changes in that layer.

8. Update weights on the output layer:

$$w_{kj}^o(t + 1) = w_{kj}^o(t) + \eta \delta_{pj}^o x_{p_{kj}}$$  \hspace{1cm} (6.13)

where $\eta$ = the learning rate. In this step, weights are modified slightly to produce a smaller error the next time the input pattern is presented. The amount of weight modification is proportional to the amount of error. In addition, the learning rate refers to the minimum amount that each weight may change after each training case. Larger learning rates indicate greater weight changes.

9. Update weights on the hidden layer:

$$w_{ji}^h(t + 1) = w_{ji}^h(t) + \eta \delta_{pj}^h x_{p_i}$$  \hspace{1cm} (6.14)

10. Calculate the error term:

$$E_p = \frac{1}{2} \sum_{k=1}^{M} \delta_{pk}^2$$  \hspace{1cm} (6.15)

$M$ is the number of neurons in the output layer. The error term is a measure of how well the network is learning. When the error is acceptably small, training can be discontinued. As stated earlier, the back-propagation learning algorithm is based on minimizing squared error that in this regard this algorithm is similar to the multi-linear regression (MLR) algorithm.

The primary use of the back-propagation learning algorithm is to perform supervised training on MLP networks which then may be used for either continuous mapping (e.g., prediction, function approximation) or classification purposes (Wasserman, 1993). As such, MLP networks trained by back-propagation learning algorithm are the best known and by far the most utilized supervised ANN models in various applications (AbouRizk et al., 2001;
Aljahdali, 2003; Attalla and Hegazy, 2003; Hegazy et al., 1994; Jeon, 2007; Jeong, 2004; Moselhi, 1998; Wasserman, 1993; Wilmot and Mei, 2005; Zhao et al., 1998). The popularity of the back-propagation learning algorithm arises from some certain reasons. One of these reasons is its simplicity (Hegazy et al., 1994; Jeon, 2007; Rumelhart et al., 1986). This algorithm updates the network’s connection weights through a simple computational procedure consisting of only two phases. Moreover, in complicated applications where the input and output relationships are non-linear, the back-propagation learning algorithm enables MLP networks producing reasonably accurate results (Abou-Elseoud, 1998). Such networks have the ability to simulate many statistical functions surprisingly well (Doig, 1999). Another reason accounting for its unprecedented interest amongst statisticians and other primary researchers is that MLP networks trained by the back-propagation learning algorithm have the ability to reveal an acceptable level of generalization (Caudill, 1992; Hegazy et al., 1994; Maren et al., 1990; Rumelhart et al., 1986). Consequently, such networks tend to give correct outputs when presented with the inputs they have never come across with. With all these advantageous properties, some issues must be carefully considered when training an ANN model by the back-propagation learning algorithm.

When using the back-propagation learning algorithm, there are some parameters whose values should be properly set in order to avoid the learning process being too slow, and also to ensure obtaining the convergence. Normally, the values of these parameters need to be heuristically set as there is little theory to guide their determination. The desirable performance of the back-propagation learning algorithm can be very sensitive towards these settings. An unfortunate choice can cause slow convergence or network paralysis where learning virtually stops (Wasserman, 1993). Therefore, the process of developing back-propagation models often requires iterative refining of the parameters embedded in the back-propagation learning algorithm. This iterative procedure needs to be accomplished through an extensive trial and error approach which, in turn, requires a considerable time and effort to be spent. These parameters are as follows:

- **Learning Rate**: MLP networks trained by the back-propagation learning algorithm are sensitive to the learning rate which modifies adjustments to the connection weights. The learning rate prevents excessive weight changes during the training phase. Through the recursive process of the back-propagation learning algorithm, the weights of a network are allowed to be adjusted according to a learning rate which specifies the magnitude of change for each backward pass. The higher the value of the
learning rate, the faster and more radical changes would occur in the adjustment of connection weights. Thereby, a large learning rate may speed up training through rapid error correction in a network. However, using a larger learning rate increases the risk of overshooting the best solution and may potentially lead the training to non-convergent solutions. Alternatively, using a smaller learning rate may be appropriate in allowing the network to gently converge, but it may cause the network to require more iteration through the training dataset, which, in turn, extends the training time. The value of the learning rate ranges from 0 to 1. A suitable learning rate has to be obtained through a trial and error approach by experimentally testing different learning rates. Kim et al., (2004) studied the effect of different learning rates from 0.5 to 0.9 (in steps of 0.1) on results of the ANN models that they established. Kim et al., (2004) concluded that a specific rule that yields the best value of the learning rate could not be found. However, some researchers (Hegazy et al., 1994; Pao, 1989) suggested that the learning rate can generally be set at 0.7. This suggestion was supported by the results of a recent study conducted by Chen and Huang (2006).

- **Momentum:** Ideally, the network should find the global minimum, but this cannot always be guaranteed (Liu, 1998). The neural networks could be trapped in a local rather than global minimum of the error function and, therefore, arrive at an unacceptable solution (Abou-Elseoud, 1998). Because MLPs may also get stuck in local error minima and may prematurely stop learning, techniques have been developed to overcome this problem. One of these techniques is called momentum. Momentum is not an integral part of the back-propagation learning algorithm, but it can be added as an extension to this algorithm. Momentum serves to stop the learning process from heading in an undesirable or looping direction (Bullock, 1999) which, in turn, minimizes the chance of getting stuck in a local minimum. The momentum value determines how much of the previous corrective term should be remembered and carried on in the current training. The momentum is in fact the factor that helps proportion the change in previous weight to a new weight, which allows the model to converge more promptly to an optimized solution. Normally, the larger the momentum value, the more emphasis is placed on the current correction term and the less on previous terms (Trippi and Turban, 1994). In addition to improving training accuracy, momentum is often thought to exhibit advantages in speeding up training convergence (Doig, 1999; Gately, 1996). Similar to the learning rate, a trial and error
approach needs to be tailored to choose the most efficient value for the momentum. Different values in a range of (0, 1) can be tested in this regard. Pao (1989) and Hegazy et al., (1994) proposed to commonly set the momentum coefficient at 0.9. Also, some researchers have examined the predictive performance of ANN models while altering the value of momentum. For instance, Kim et al., (2004) implemented different ANN models to compare the estimation error rates of these models through changing the value of momentum from 0.1 to 0.9 (in steps of 0.1). Kim et al., (2004) found that the error of ANN models was relatively small in those cases with a momentum of 0.6, 0.7, and 0.8. Almost the same finding was reported in another study by Kim et al., (2005). In this study, however, they simultaneously examined the effects of the number of hidden neurons and the momentum value on the predictive ability of ANN models. The study limited the momentum values from 0.3 to 0.9 at the step of 0.1 and limited the number of neurons in the hidden layer between 12 and 33. The results showed that when the number of neurons in the hidden layer was 12, 16 and 33, the difference between predictive performances of ANN models was low without significant fluctuation for momentum between 0.5 and 0.8. In contrast to the above studies, the results of some studies indicated that it might be advantageous to select a smaller value for the momentum. For instance, Chen and Huang (2006) altered the momentum value from 0.1 to 0.7 at the step of 0.2 to determine the best value for each of the six predictive ANN models developed to estimate cost and duration of reconstruction projects. Chen and Huang (2006) concluded that setting the momentum value to 0.3 enables all the six ANN models to reach their highest predictive performance.

- **Iteration or Epoch:** In ANN models trained by the back-propagation learning algorithm, training cases are presented to the network sequentially. In some cases, it may be necessary to present each training example to the network many thousands of times. Generally, the number of times that the training examples are presented to the network is measured in iterations, known also as epochs. An epoch is therefore the one time presentation of each training sample to the network. For instance, if there were n observations in the training dataset, one epoch would consist of presenting, in random order, each of the n observations to the network (Bullock, 1999). After each iteration, the error between the network outputs and the actual outputs is propagated backwards to adjust the weights in a manner that is mathematically guaranteed to
converge (Rumelhart et al., 1986). The network becomes more knowledgeable after each iteration and the results of the previous iterations are used as the starting point for the next iterations. This processing and learning of weights continues until the network is said to converge or another event determined by stopping criteria occurs. No formula exists to calculate the exact number of iterations that should be used in the ANN training. However, too much iteration, like too many hidden neurons, may cause over-fitting problem which, in turn, causes ANN models to lose their generalization ability. Alternatively, not enough iterations of the training dataset could result in the problem of inadequate learning. This problem is also known as under-fitting problem.

**Stopping Criteria:** One of the most critical questions in training a network is when to terminate training. This question can be defined by implementing different stopping criteria. The stopping criteria for training an ANN model can generally be that the network reaches either training tolerance or its maximum number of iterations. Training tolerance values represent the allowable variation when the network’s outputs are compared to the corresponding actual outputs in either training or selection datasets. The description of these datasets, together with the test dataset, is presented in Section 6.6. For instance, the network training can be stopped when the network performance no longer improves on the training dataset by a specified amount or when the network no longer shows improved performance when tested on the selection dataset. As an alternative to training tolerance criteria, training of the network may also cease when a pre-specified maximum number of iterations have been completed. A training tolerance of 0.0 requires an exact match of the network’s outputs and the actual outputs in the training dataset which, in turn, precludes connection weights being updated. A higher training tolerance, will allow more variation in the network’s outputs before errors are propagated back through the network. A smaller training tolerance will provide more accurate results, but at the cost of more iterations and, hence, learning time. For the purpose of setting a viable, yet practical, training tolerance, it is recommended that networks can be considered to have learned the training examples reasonably well if the total sum of squared error is of the order of 0.01 to 0.001 (Bishop et al., 1991).
- **Connection Weights:** A network has a parallel distributed architecture with a large number of neurons and connections with varying weights. In the back-propagation learning algorithm, a set of weights needs to be first randomized, generally to small non-zero values. The network accuracy performance is sensitive to the initial set of network weights. In addition, initializing proper interconnection weights is also advantageous in terms of boosting training precision and also lessening the training time. In the literature, different values for initializing the connection weights have been proposed. For instance, Hegazy et al., (1994) suggested to begin with values between -0.5 and +0.5, whereas Hegazy and Ayed (1998) found that setting the initial weight values to a range between 0.5 and 1 is appropriate, particularly for inputs scaled to a range (-1 to 1).

- **Scale Function or Normalization:** Neural network training can be made more efficient if certain pre-processing steps are performed on the network variables. One of these steps, which is widely adopted, is normalization of the variables. When variables are loaded into an ANN model, they must be scaled from their numeric range into the numeric range that the network deals with more efficiently. Normalization is often so helpful since it removes any bias from the combination of variables belonging to different numeric ranges (Jeon, 2007). Another advantage of normalization is speeding up the convergence and, as a result, lessening the learning time (Bullock, 1999). There are two main numeric ranges that ANN models can more effectively operate in: (1) zero to one and (2) minus one to one. The latter scale function is more popular as it enables back-propagation networks provide improved performance (Hegazy and Ayed, 1998; Williams, 1994). The following formula illustrates how the data is transformed by this scale function:

\[
V_{norm}^p = 2 \times \left( \frac{V_{p(i)} - V_{min}}{V_{max} - V_{min}} \right) - 1 \tag{6.16}
\]

where \(V_{min}^i\) = minimum in the dataset of the input vector \(V^i\) (ith), \(V_{max}^i\) = maximum in the dataset of the input vector \(V^i\) (ith), and \(V_{p(i)}^i\) = input (p\textsuperscript{th}) value in the vector \(V^i\) (ith).
6.6 Training, Selection and Testing Dataset

In most cases, ANN modellers train and evaluate the performance of their models by implementing a standard method, called “cross-validation”. In this method, the main database is randomly partitioned into three subsets consisting of a training dataset, a selection dataset, and a test dataset (Garson, 1998; Haykin, 1999; Scarborough, 1995). These subsets should be representative of the samples of potential input-output pairs that the network will experience in practice. The subsets should be independent from each other and selected by random, but their samples need to be drawn from the same population.

- **Training Dataset**: Training dataset is the first subset which is used for computing and updating the network’s connection weights. As stated earlier, to develop ANN models trained by a supervised learning rule it is necessary to produce a training dataset with target output data. Therefore, the training dataset consists of a number of samples or observations, whose attributes are known. The performance of ANN models strongly depends on the quality of the training dataset (Bode, 2000; Günaydın and Doğan, 2004). The input and output pairs in the training dataset are fundamental in ANN methodology, because they convey the necessary information for discovering the best relationship amongst variables. It is highly important to have reliable input data so that the output produced by the network can compute a meaningful correlation (Al-Rashidi, 1999). To improve the generalization performance of ANNs, the quality of the training dataset needs to be carefully examined by taking the following set of rules into account (Freeman and Skapura, 1991; Moody, 1992; Williams, 1994). First, the back-propagation learning algorithm does not extrapolate well. Hence, data in the training dataset must span over the entire range of expected data space. Second, the distribution of training inputs must be sufficiently dense over the whole range of expected inputs to produce an accurate interpolation. If one does not have training dataset that covers a significant portion of problem conditions, then ANN methodology is probably not the right solution (Ostberg, 2005).

Besides quality, the performance of an ANN model hinges heavily on the size of the training dataset (Attalla and Hegazy, 2003; Bode, 2000; Günaydın and Doğan, 2004; Jeon, 2007; Liu, 1998; Wang et al., 2000). Generally, the larger the training dataset, the richer that dataset in describing the problem domain, the less the prediction error, and the more accurate the network is likely to be (Stockton and Wang, 2004; Wang et
al., 2000). In addition, the best way to avoid over-fitting problem, which is mainly caused by inappropriately large number of hidden neurons and leads the network to lose its crucial ability to generalize, is to use lots of samples in the training dataset. It is believed that the inclusion of a greater number of data in the training dataset would significantly improve the accuracy and the generalization abilities of ANN models (Boussabaine, 1996; Elhag and Boussabaine, 1999; Emsley et al., 2002; LeCun, 1989; Smith and Mason, 1997; Shehab et al., 2010). In the literature, there is no rigorous theory to help in selecting the training dataset size. Rather, the number of data required for training the model depends on the nature of the problem and the number of independent (i.e., input) variables. For instance, the results of Bode (2000) study suggested the suitability of ANNs for cost estimation tasks if the number of cost predictors is not significantly above five or six, and if at least 50 to 100 past cases are available to form the training dataset. Bode (2000) also concluded that the marginal contribution of additional training data decreases with growing the size of the training dataset. This conclusion was in a good agreement with the results of a former study conducted by Shtub and Zimerman (1993). They evaluated the effect of the training dataset on the overall generalization performance of the ANN models developed for estimating the cost of six assembly systems. Two training datasets with 756 and 2268 cases were used for this purpose. Shtub and Zimerman (1993) found that only in one assembly case, the generalization considerably improved when using the larger training dataset. In all the other 5 cases, the generalization with the larger training dataset is not significantly better than the generalization with the smaller training dataset. In another study, Liu (1998) studied the effect of three groups of training datasets with sizes of 60%, 75%, and 90% of the total usable collected projects on the predictive performance of ANN models. The results showed that the 90% sized training dataset had the smallest standard deviation and the largest coefficient of determination ($R^2$). Liu (1998) then came to the conclusion that, while there is no limit of the training dataset size, the more samples in the training dataset, the better the performance of ANNs. In this regard, some researchers have suggested that using training samples as much as ten times of the number of input variables would be typically a reasonable choice (Bailey and Thomson, 1990a; Eberhart and Dobbins, 1990; Jeon, 2007; Smith, 1993). As another empirical, yet simple, proposition for establishing the training dataset, it is recommended that at least 50% of the total
samples in the entire database should be accommodated in the training dataset (Alex et al., 2010; Williams et al., 2005).

- **Selection Dataset:** ANN models trained by the back-propagation learning algorithm tend to over-train when they were only trained by applying the training dataset in the training stage (Adeli and Wu, 1998; Tzafestas et al., 1996). As a network trains to a point of suitable stability on the training dataset, the network performance is evaluated using a second set of input-output pairs in the selection dataset to which the network has not yet been exposed. While being a portion of the training dataset that are not used to train the ANN model, the selection dataset acts as an independent check on how well the network is trained. The selection dataset is used to monitor model performance during the training phase. The error on the selection dataset is monitored during the training process. The selection dataset error normally decreases during the initial phase of training, as does the training dataset error. However, as the training proceeds, the selection dataset error may increase while, by contrast, the training dataset error retains its decreasing trend. In such a case, it is so likely that overtraining or over-fitting problem has occurred. The overtraining problem causes the neurons to memorize the input and output mappings in the training dataset, without any capability to generalize. To eliminate this problem, the training should be stopped and the connection weights attributed to the minimum selection dataset error are returned. The optimal ANN model is therefore the one that has the lowest selection dataset error. Consequently, the training phase terminates when the network performed the best on the selection dataset rather than the training dataset. This termination would keep an ANN model from over-fitting on the training dataset (Williams, 1994), while would allow the development of the model and its validation to occur at the same time (Attalla, 1999). It has been suggested in the literature to randomly separate one-sixth to one-third of all cases for the selection dataset (Alex et al., 2010; Shehab et al., 2010; Williams et al., 2005). It has been also suggested that a reasonable way to create the selection dataset is to remove a random extraction of about 10% to 40% of the samples in the training dataset before starting the training phase (Ward Systems Group Inc, NeuroShell 2, User Guide, 1998). This suggestion is more meaningful for cases where the samples of the selection dataset are to be drawn out from the training dataset.
- **Test Dataset**: Once a network is trained using the training and selection datasets, it is not generally known whether the trained network will yield reasonable outputs for the inputs which have not been used in the training phase. Theoretical work by Hecht-Nielsen (1989) has shown that neural networks can learn input-output relationships to the point of making perfect forecasts with the data on which the network is trained. However, perfect forecasts with the training dataset do not guarantee optimal forecasts with another new dataset due to differences in the two datasets (Swicegood, 1998). To remove this concern and examine the confidence level of the network, it is essential to fix network parameters after the training phase, and then measure the performance of the trained network on the observations in the third subset which is the test dataset. The further use of the trained network is justified only when the network shows a satisfactory generalization performance on the test dataset. Another important issue that needs to be carefully considered is the range of data values in the test dataset. If either independent or dependent variables in samples of the test dataset contain values which are not in the range of corresponding values in the training and/or selection datasets, the prediction constitutes an extrapolation outside the latter datasets. Similar to the regression models, because of the unreliability of extrapolation with neural networks, such samples should be avoided and the measures of prediction error should be recalculated again after exclusion of these samples from the test dataset (Smith and Mason, 1997). As for the training dataset and the selection dataset, there is no robust rule for determining the appropriate number of samples in the test dataset. However, it has been suggested that the number of samples in the test dataset need to be at least a third of samples in the training dataset (Hammerstrom, 1993).

### 6.7 Generalization vs. Overtraining

The ANN methodology is an attempt to improve the odds that the right decision will be made by creating a forecast for tomorrow based on the results from yesterday (Bullock, 1999). To accomplish this attempt, ANNs simulate the structure and learning behaviour of biological neuronal systems and, therefore, are able to exhibit complex capabilities including learning, recalling, and, most importantly, generalizing the learned relationships. Generalization is an absolutely indispensable requirement for improving the network sustainability and achieving a proficient network.
As stated earlier, in the back-propagation learning algorithm, one starts with a training dataset and uses this algorithm to compute the connection weights between layers of a network. By means of this training and learning, ANNs create their own internal representation of the problem, but it is not generally known whether they can successfully map unforeseen data not used in the training dataset. It is hoped that the trained network will be capable of generalizing previous cases to new cases. A network is said to generalize well when it is found to adequately respond to situations they have not been previously exposed to. So, the generalization performance of a network is normally tested over the test dataset to examine whether the network is able to correctly model the relationship of input-output pairs in the test dataset. In general, generalization serves as a measure of (1) the sufficiency of the training dataset to cover much of the problem's solution space and (2) the suitability of the network configuration to the problem being modelled, irrespective of how well the network performed during training (Hegazy et al., 1994).

Overtraining or over-fitting is one of the problems that may occur during network training. The network learns irrelative details about the individual samples rather than the basic structure of the data presented to it (Boussabaine and Cheetham, 1995). In this situation, the error on the training dataset is driven to a very small value, but when the network is subjected to a new dataset, the error is quite large. It happens when the network learns too many specific input-output relations in the training dataset. Consequently, the network memorizes the training pattern rather than learning it, causing the network loses its generalizing ability.

The generalization and over-fitting features of a neural network model are mainly determined by two factors: (1) number of input (i.e. independent) variables and (2) number of hidden neurons. If the number of input variables in the network is much smaller than the total number of samples in the training dataset, then there is little or no chance of over-fitting (Liu, 2006). In addition, a large number of hidden neurons leads to a small error on the training dataset, but not necessarily leads to a small error on the test dataset. As illustrated in Figure 6-11, the error rate of the training dataset would always be reduced by increasing the hidden neurons. However, adding the neurons in hidden layer(s) would first cause a reduction in the error rate of the test dataset but then increase this rate. To prevent the overtraining problem from occurring, it is very important to set the number of hidden neurons to a proper value. This prevention also enables the network to reach its highest possible generalization
capability. Assuming the generalization performance of the ANN model is acceptable over the test dataset, the ANN model can then be widely deployed to solve the problem at hand.

![Error rate of training and test datasets as a function of number of hidden neurons](Jeon, 2007)

Figure 6-11: Error rate of training and test datasets as a function of number of hidden neurons (Jeon, 2007)

### 6.8 Pros and Cons of the ANN Methodology

The ANN methodology has recently been the focus of much research since its inception in the middle of the twentieth century. ANN models have been widely utilized in climate control systems, real-time process control, numerous consumer electronic devices and appliances, robotic vision and control systems, voice recognition, bomb identification in luggage, missile guidance systems, flight control systems, stock market forecasting, investment analysis, bankruptcy forecasting, economic policy making, as well as many other applications (Haykin, 1999; Ostberg, 2005). ANN methodology has gained a broad acceptance in such a vast array of disciplines because of salient advantageous, some of which are summarized below:

- The major property that makes ANNs to be superior to other conventional techniques is their ability to be trained on historical information as well as real-time data (Hegazy and Ayed, 1998).
- ANNs are true models of the problem in that they are not reducible to a simple formula or procedure (Lawrence, 1994).

- ANN models can identify and place importance on predictors that classical statistical techniques do not (Doig, 1999).

- In comparison to conventional techniques, ANN methodology requires less expert opinion and judgement. This feature arises from one of the unique characteristics of the ANN methodology that makes it to be free of many constraints and limitations inherent in mathematical concrete modelling (Liu, 1998; Swicegood, 1998).

- ANNs’ operations are flexible (Hornik et al., 1989). ANNs continue learning from any point to improve their performance. This improvement stems from the built-in capability of ANNs, which enables them to adapt their weights in response to changes in the data. Therefore, as new inputs enter a network or the accuracy of output diminishes, the neural network can modify its behaviour and redefine processing.

- ANNs can work with various types of variables (e.g., continuous, categorical, ranked, and dichotomous).

- A neuron is basically a non-linear mathematical device (Abou-Elseoud, 1998). Consequently, an ANN model, made up by a heavily interconnection of neurons, is non-linear by its own. The ANN methodology is therefore considered as the best problem solving approach when data exhibit non-linear characteristics (De Groot and Würtz, 1991).

- ANNs are particularly valuable in solving problems which are difficult to be analysed with conventional modelling approaches, particularly when a priori knowledge of the problem is lacking and when the relationships amongst the variables are unknown, complex, or, as stated earlier, non-linear (Alex et al., 2010; Attalla, 1999; Berends, 1998; Boussabaine, 1996; Doig, 1999; Garson, 1998; Jeong, 2004; Liu, 1998; Rafiq et al., 2001; Rumelhart et al., 1994; Shtub and Zimerman, 1993; Somers, 1999). Unlike conventional techniques, ANNs are able to detect any patterns found in the data and can provide a larger opportunity to estimate the dependent variable for various
conditions without a precise knowledge of all contributing independent variables and their relationships (Al-Rashidi, 1999; Attalla and Hegazy, 2003; Bode, 2000; Jeong, 2004).

- ANNs provide a great degree of robustness or fault tolerance (Bode, 2000). When an element of a neural network fails, the network can continue without any problem because of its massively parallel nature within which input variables, processing neurons, and output variables are immensely connected to one another.

- ANNs are robust even in the face of flawed or missing data, and tend to degrade gracefully as the presence of quality data declines (Garson, 1998; Lawrence, 1994; Liu, 1998).

- ANNs’ ability to automatically learn from representative examples is one of the major advantageous of ANNs which makes them attractive and exciting (Jain et al., 1996). ANNs can learn from these examples without being in need of programming (Bullock, 1999). An ANN designer normally decides over the network basic parameters which are necessary to design an ANN model (i.e., architecture, activation function, and learning algorithm). Then, the model automatically adjusts these parameters.

- The numbers of input and output neurons in ANNs are not restricted, which is an advantage of ANNs (De la Garza and Rouhana, 1995; Smith and Mason, 1997). In particular, ANNs will be more desirable for those cases in which more input (i.e., independent) variables are taken into account (Attalla and Hegazy, 2003; Jeon, 2007; Liu, 1998). In these cases, ANNs become more divers and are able to detect any patterns found in the data and provide larger opportunity to investigate different factors affecting the problem at hand (Attalla and Hegazy, 2003). Again, this advantage of ANNs arises from their massively parallel and interconnected processing system, which enables ANNs to simultaneously consider numerous factors.

- ANNs are able to provide higher levels of success in solving major tasks for which a large amount of data is available (Alex et al., 2010; Al-Rashidi, 1999; Jeon, 2007; Liu, 1998; Peng, 2006). Moreover, researchers may also benefit from ANNs on
occasions where data for design and interpretations are less than adequate (Al-
Rashidi, 1999; Shehab et al., 2010).

- In comparison with conventional modelling techniques, ANNs can better simulate the
  true behaviour of variables that are characteristically sparse and noisy (Abou-Elseoud,
  1998; Garson, 1998; Swicegood, 1998). The importance of this issue is more
  pronounced in the field of cost estimation because such estimation is often performed
  for new products and processes, for which good quality historical data does not exist.
  Thus, the cost model must make the most of sparse, noisy and approximate
  information (Smith and Mason, 1997).

- In terms of the range of application areas, there is potentially no limit to the type and
  variety of tasks that can be estimated using the ANN methodology as this
  methodology is not problem type dependent (Wang et al., 2000).

Despite the advantages mentioned above and despite the good performance of ANNs
in previous research efforts, the process of developing and implementing ANNs has a number
of problems associated with it. These problems give, at least in part, an explanation why over
the last 20 years one would find only a small handful of published research in which ANNs
or other forms of computational intelligence were utilized (Collins and Clark, 1993; Somers,
1999); rather, traditional techniques such as MLR equations abound (Ostberg, 2005).
Following are the factors which may explain the sluggish acceptance of neural network
modelling with respect to conventional approaches:

- User-friendly neural network modelling software was not widely available until the
  early 1990’s (Ostberg, 2005). Neural modules have been slow to become available
  within statistical packages such as SPSS and SAS which are more popular amongst
  researchers (Ostberg, 2005).

- The ANN approach does not mitigate any of the difficulties associated with
  preliminary activities when using statistical parametric methods (Smith and Mason,
  1997). The analyst is still left with a choice of cost predictors and frequently must
  make a one-time commitment to collecting specific cost data before analysis begins.
- Although ANNs excel at various types of prediction, casual analysis may be problematic because ANNs neither automatically explain their reasoning nor provide an audit trail that fully explains how the system reaches its conclusion (Touretzky and Pomerleau, 1989). All the knowledge that the network acquires during learning is implicitly encoded in its numeric weights. In large MLP networks with large number of neurons and connections, a huge number of numeric values for the weights will result, making it extremely difficult to interpret and draw any meaningful explanation for the solution process (Hegazy et al., 1994). Resultantly, the user would not be able to understand and explain the estimation results from ANN models. Thus, the explication of problems and relationships may be difficult (Collins and Clark, 1993; Hegazy and Ayed, 1998; Somers, 1999). This difficulty has contributed to the “black box” image of ANNs (Boussabaine, 1996; Garson, 1991; Hegazy et al., 1994; Hegazy and Ayed, 1998; Kim et al., 2004; Shtub and Zimerman, 1993). According to Smith and Mason (1997), explaining the black-box nature of ANNs in arriving at their answers could be much like explaining how one plays tennis by doing a dissection of the tennis player’s brain tissue. Hence, in decision-making process, other models like case-base reasoning and MLR are quite better and more successful than ANN models in explaining the casual relationships among the variables of the problem at hand (Kim et al., 2004).

- ANN modelling techniques are numerous and often quite complex. Training a neural network is an algorithmic procedure and the results can most certainly be replicated as long as one uses the identical computer software, the same initial weights, the same training data, and the same deterministic method of presenting the data during training. However, if even one of these parameters is altered, the resulting neural network would almost certainly be different from the original one. This difference is apt to be extremely minor; however it is not inconceivable that major differences could occur (Smith and Mason, 1997). That is why it has been stated that neural network modelling is still more art than science (Swicegood, 1998) and the one who employs this type of modelling is an artist whose work is never finished, or at least, is an artisan who is never sure the analysis he or she presents might not be sub-optimal (Garson, 1998).
ANNs are limited in a sense that one cannot simply look at the coefficient estimates that the network develops and determine the significance of a given input variable (Swicegood, 1998). ANNs are also unable to estimate the overall effect of individual input parameters (Doig, 1999).

The fields for which researchers employ the ANN methodology to develop their models are quite different. Therefore, simply learning the fundamentals of ANN modelling can prove to be difficult. This difficulty is even more pronounced with the unfamiliar vocabulary being employed within the context of ANN modelling. As a result, the learning curve in designing and interpreting ANN models is more imposing than that of conventional statistical models, where decisions are fewer and guidance is readily available from texts and software (Smith and Mason, 1997).

ANNs can generalize well on new cases if over-fitting is avoided (Attalla, 1999). One of the major drawbacks that ANN users may overlook is over-fitting (Attalla, 1999), which forces networks to develop an exact implementation of the model on the given training dataset. Over-fitting impairs the network objective as defined to minimize the generalization error or error on the new dataset which is not used in the training dataset (Pados and Papantoni-Kazakos, 1994).

ANN models need training to operate and their knowledge acquisition process can be highly time-consuming (Boussabaine, 1996; Hegazy et al., 1994; Hegazy and Ayed, 1998; Jain et al., 1996). In particular, larger and more complex networks, despite their greater representational and modelling power, need much longer training time (Openshaw, 1993).

ANN models require multiple training as there is no explicit set of rules to determine whether a given learning algorithm is suitable for a particular application. The learning algorithms, such as back-propagation, may be suitable for some applications but not for others (Hegazy and Ayed, 1998).

In the ANN methodology, numerous values for parameters, embedded in a given training algorithm (such as learning rate, momentum, and initial weights), need to be
examined by trial and error to improve the performance of ANN models (De la Garza and Rouhana, 1995; Jain et al., 1996; Kim et al., 2004).

- Designing the network architecture and setting its parameters (such as number of hidden layers, and number of hidden neurons in each hidden layer) is also a cumbersome task as there is no appropriate rule to determine these parameters. Therefore, to determine the best network architecture that best fits the application under consideration, an extensive trial and error processes is inevitable (Attalla and Hegazy, 2003; Bode, 1998; Boussabaine, 1996; Creese and Li, 1995; Hegazy et al., 1994; Jain et al., 1996; Kim et al., 2004; Li, 1995; Liu, 2006; Wang et al., 2000; Yeh, 1998).

- Conventional methodologies such as the regression analysis are well practiced for prediction purposes, whereas the ANN methodology is not. The common perception of ANN models as “mystical black boxes” (Somers, 1999) may increase concerns regarding the appropriateness and legal defensibility of these models.

6.9 ANN Computational Software Packages

Several software packages have been recently available for the application of ANNs. One of these software packages is called “NeuroShell 2” (Neuroshell 2: Release 4.0, 1998) which is developed by Ward Systems Group® Inc. Literature search revealed that several research studies have employed this software as one the most preferred analytical tool in order to develop predictive ANN models (Alex et al., 2010; Al-Rashidi, 1999; Attalla and Hegazy, 2003; Hegazy and Ayed, 1998; Kim et al., 2004; Shehab et al., 2010; Williams, 1994). The software has broadly been utilized for its ease of use; speed of training, and for its host of different architectures, which take advantage of the back-propagation learning algorithm for training. NeuroShell 2 software includes a simplified set of procedures for designing and executing a complete and powerful ANN application with flexible user-optimization of training parameters. In this software, the user is allowed to specify the learning rate, momentum, activation functions and initial weights in the design module. Also, the software has multiple criteria for stopping training in addition to different methods for handling missing data, pattern selection and viewing weight and neuron values during training (Attalla and Hegazy, 2003). Given the advantages stated above, NeuroShell 2
software was employed in the current study to develop robust non-parametric retrofit cost models, using the ANN methodology.

Other commercial software packages available for ANN computing include Matlab developed by MathWorks Inc., NeuralWorks Professional II developed by NeuralWare Inc., Trajan Neural Network Simulator developed by Trajan Software Ltd, and Braincell developed by Promised Land Technologies Inc. The latter ANN computing software package can be run as an add-in for Microsoft Excel to take advantage of Excel’s rich data handling and graphing capabilities.
6.10 Conclusion

A thorough description of the ANN methodology was provided in this chapter. This methodology is one of the fastest growing and most innovative areas of intelligent computing methodologies, which assume that human learning can be imitated with computer simulations of real biological nervous system. ANN models consist of two basic components, being processing elements (PEs), also known as neurons, and connections. Each neuron within an ANN model communicates with other neurons thorough connections. The operation of ANN models is based on three fundamental factors, being architecture, activation function, and learning algorithm.

How the neurons and their associated connections are arranged within a network determines the network architecture. A literature search revealed that the multilayer perceptron (MLP) architecture is the best known and most utilized type of network architecture. Activation function, as a control unit, dictates when neurons activate, and controls the amplitude of the output of each neuron in an ANN model. The non-linear logistic activation function is the most popular activation function used in the literature. Finally, learning algorithm governs the adjustment of connection weights in a network. This adjustment enables the network to achieve a desirable performance. The back-propagation learning algorithm is the most common learning algorithm, which is used in conjunction with MLP networks.

The performance of a MLP network trained by the back-propagation learning algorithm is controlled by a number of parameters such as the number of hidden layers, the number of hidden neurons in each hidden layer, learning rate, and momentum. A fixed combination of these parameters which can be applied to every problem does not exist. Selecting these parameters is highly dependent on the problem context, the level of knowledge about the problem, and the modeller’s preference. Usually, the best combination of these parameters for a particular application is identified heuristically through the use of a trial and error approach. The main objective in this identification is to obtain an ANN model with the highest generalization ability, where generalization is a measure of the model’s performance on data not used in the training process (i.e., the test dataset).

The advantages and disadvantages of the ANN methodology were also reviewed in this chapter. According to this review, this methodology is particularly beneficial when a priori knowledge about a problem is lacking, and when the relationships amongst variables
are unknown, complex, or non-linear. The black box nature of this methodology and the lack of appropriate guidelines for the successful development of ANN models are amongst the major disadvantages of this methodology.

A number of studies utilizing the ANN methodology as a novel non-parametric tool for cost estimation were presented in Chapter 2. The results of these studies suggest that this methodology has an impressive potential to support the general cost estimation applications. In the next chapter, the efficacy of applying the ANN methodology to predict the RNCC of framed structures is investigated.
CHAPTER 7

DEVELOPING ANN MODELS FOR ESTIMATING RETROFIT NET CONSTRUCTION COST (RNCC) IN FRAMED STRUCTURES

7.1 Introduction

The RNCC models developed in Chapter 5 are commonly based on traditional linear regression method, which is based on simple assumptions and has certain limitations, and imposes stringent conditions that are inherent in the linear regression. Given these limitations, there is a need for an alternative modelling technique that can offer more precise predictions. As explained in the previous chapter, the ANN methodology is a promising alternative modelling technique for the establishment of sound predictive models. However, the applicability of this methodology in the case of retrofit cost modelling has not been previously investigated.

In this chapter, ANN models are developed for the first time in the literature to predict the RNCC. This development was based on those variables which were found to be statistically significant in Chapter 5. An ANN model was developed for each of the last seven MLR models developed by implementing the backward elimination (BE) procedure. The ANN models were all trained and tested on the identical datasets as were used to develop and validate the MLR models, respectively. A novel procedure is presented in this chapter that allows for the proper development of ANN models trained by the supervised back-propagation learning algorithm. Using this procedure, the values of the parameters, which are involved in the creation of a feed-forward back-propagation network, were determined such that the best generalization performance was obtained for each of the ANN models developed. These parameters include the number of hidden layers, the number of hidden neurons, the learning rate, and the momentum. Finally, the sensitivity of the ANN models with respect to these parameters, together with the number of input variables, is studied.

7.2 Variables to Include

A mounting consensus does exist amongst researchers that whatever analytical methodology is utilized, the most important factor in producing robust, and valid outcomes is
a rational selection of input (i.e., independent) variables (Church and Curram, 1996), which is
a process referred to as feature selection within the field of computational intelligence
(Garson, 1998). In addition, another important factor that accounts for the poor performance
of a network is that the variables are not also presented in the most appropriate way.
Therefore, selecting an appropriate functional form for each variable to be included in a
neural network is also crucial to making the network work properly.

In this research, on the basis of the results taken from the MLR modelling technique,
seven independent variables whose influence on RNCC (i.e. the dependent variable) was
found to be statistically significant were tailored to construct ANN models. Three of these
independent variables were numerical variables, while the remaining four were basically
qualitative in nature. All these variables, together with the dependent variable, were first
manipulated to appear in the same fashion as they did in the MLR models, and then used for
the development of ANN models. As such, the numerical variables were transformed to their
natural logarithm (Ln) format, while qualitative variables were translated into the binary
codes (i.e., 0 and 1) that the networks were capable of processing them. For best results, this
translation was performed in a way that maintained the relative order of meaning for possible
situations that a given qualitative variable can take on. Therefore, the latter code was assigned
to those situations expected to have an increasing effect on RNCC. Table 7-1 provides details
related to dependent and independent variables considered for ANN modelling. This table
also illustrates how these variables were to be entered into ANN models.

Table 7-1: Details of variables considered for ANN modelling

<table>
<thead>
<tr>
<th>Variable</th>
<th>Type of variable</th>
<th>Functional form of variable (as entered into the ANN models)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>Numerical</td>
<td>Ln (Y)</td>
</tr>
<tr>
<td>X1</td>
<td>Categorical</td>
<td>Ln (X1)</td>
</tr>
<tr>
<td>X2</td>
<td>Categorical</td>
<td>Ln (X2)</td>
</tr>
<tr>
<td>X6</td>
<td>Categorical</td>
<td>1, For buildings located in very high seismicity region; 0, Otherwise</td>
</tr>
<tr>
<td>X7</td>
<td>Categorical</td>
<td>1, For buildings sitting on poor – behaved soil type; 0, Otherwise</td>
</tr>
<tr>
<td>X11</td>
<td>Categorical</td>
<td>1, For buildings with irregular plan configuration; 0, Otherwise</td>
</tr>
</tbody>
</table>

Y*: Total retrofit net construction cost (RNCC) (thousands U.S. $)
X1*: Building area (m²)
X2: Total number of stories
X4: Weight per unit (ton/m²)
X6: Seismicity
X7: Soil type
X8: Plan irregularity
X11: Structural type
7.3 Number of ANN Models to Be Developed

Corresponding to each of the first seven MLR models presented in Chapter 5, an ANN model was to be developed for the prediction of the RNCC. The ANN models were therefore numbered in accordance with the number of independent variables involved in these models. The combinations of these independent variables in ANN models were also in complete agreement with those used when generating MLR equations. These ANN models, together with their corresponding independent variables, are listed in Table 7-2.

Table 7-2: ANN models to be developed in this research

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>No. of independent variables in ANN models</th>
<th>Independent variables in ANN models</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>(X_1)</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>(X_1, X_2)</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>(X_1, X_2, X_6)</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>(X_1, X_2, X_6, X_{10}, X_{11})</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>(X_1, X_2, X_6, X_7, X_{11})</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>(X_1, X_2, X_6, X_{10}, X_7, X_{11})</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>(X_1, X_2, X_6, X_{10}, X_7, X_{11})</td>
</tr>
</tbody>
</table>

7.4 Training, Selection, and Testing Datasets

In this research, the data from 158 retrofit projects were used to train and test the ANN models. According to the suggestions provided in Chapter 6, seventy-five percent of these projects (i.e., 121 samples) were randomly selected and used during the training of ANNs, whereas the remaining twenty-five percent (i.e., 37 samples) were used to provide an independent test dataset. Ten percent of the training samples (i.e., 12 samples) were also extracted at random to constitute the selection dataset.

Notably, the training and test samples were exactly the same as those which were previously used in Chapter 5 to create and validate the MLR models, respectively. As a result, the same set of samples is used in the next chapter to compare the accuracy and
generalization capabilities of ANN models with those of their corresponding MLR counterparts.

7.5 Development of ANN Models

As described in Chapter 6, an ANN designer needs to first select an ANN paradigm which is mainly defined by “architecture”, “transfer or activation function”, and “learning algorithm” and then set the most appropriate values for various network parameters, particularly those embedded in a given architecture as well as learning algorithm. In essence, choosing a network’s paradigm is highly dependent upon the problem context, the designer’s preference and also the level of knowledge of the system to be modelled. However, determining the best values for network parameters is quite a burdensome task as there is no appropriate rule to determine these values. Therefore, one needs to iteratively examine numerous values for each of the parameters under study and consequently select the values that best fit the problem at hand. The following sections describe in detail the ANN modelling paradigm utilized in this research.

7.5.1 Architecture

The first step in creating an ANN model is selecting an appropriate architecture for the model. In this research, the multilayer perceptron (MLP) architecture was selected and utilized to develop predictive RNCC models. As described in the previous chapter, the MLP network is the most popular and widely used form of feed-forward architecture, which exhibits a number of significant features. Its generalization and classification capability (Caudill, 1992; Hegazy et al., 1994; Liu, 1998; Maren et al., 1990; Rumelhart et al., 1986), together with its ability to deal with complex systems and, as a result, to estimate highly non-linear functional forms (Cybenko, 1989; Funahashi, 1989; Haykin, 1999; Hornik et al., 1989; Hush and Horne, 1993; Wilmot and Mei, 2005) are amongst the most prominent reasons accounting for why it is the most preferred architecture.

Three exclusive sets of fully connected MLP networks were examined for each of the seven ANN models listed in Table 7-2 to find their corresponding best architectures. These MLP networks are generally recognized by the number of hidden layers that they are
composed of. Figures 7-1 to 7-3 are graphical representation of these networks which are henceforth called MLP networks with one, two, and three hidden layers, respectively. The network with three hidden layers is also known as Ward Net. Networks of more layers are not considered in this research as it is believed that using more than five layers is of no benefit (Ward Systems Group Inc, NeuroShell 2, User Guide, 1998).

7.5.1.1 Input and Output Neurons

In the ANN modelling, the input layer accommodates the independent variables of interest, while the output layer represents the dependent variables. As such, the number of input neurons corresponds to the number of independent variables and similarly the number of output neurons is akin to the number of dependent variables to be predicted by the ANN models.

In MLP networks as shown in Figures 7-1 through 7-3, the first slab represents the input layer in which the input neurons should be placed, while the last slab (i.e., slab 3, slab 4, and slab 5 in MLP networks with one, two, and three hidden layers, respectively) represents the output layer which is, for all the ANN models studied, composed of a single
neuron representing the dependent variable of interest. Table 7-3 summarizes the number of input and output neurons for each of the ANN models in Table 7-2.

Table 7-3: Number of input and output neurons in ANN models

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>No. of input neurons (m)</th>
<th>No. of output neurons (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>1</td>
</tr>
</tbody>
</table>

7.5.1.2 Hidden Neurons

Choosing the network size in terms of the number of hidden neurons is an extremely critical step in the design of an ANN model (Abou-Elseoud 1998). This issue merits attention because the approximation capability of a neural network is primarily defined by the number of hidden neurons (Jeon, 2007).

A number of equations that are widely recommended in the literature to define the most likely appropriate number of hidden neurons were presented in Chapter 6. To set the initial number of hidden neurons in each of the ANN models studied in this research, these equations were all examined and, from the values calculated for a given ANN model, the value that was quite close to the largest one was initially selected. This selection was done because increasing the number of hidden neurons, up to a certain limit, would typically enhance the performance of an ANN model, and, as a consequence, it may be better to err on the side of using more rather than fewer hidden neurons (Chu, 1998; Liu, 1998; Yoon et al., 1993).

For each of the ANN models studied, Table 7-4 presents the equations being used, and the number of hidden neurons suggested by each equation, together with the value taken for initializing this number. According to Table 7-4, the last equation in this table (i.e., Equation 5) consistently outperforms other equations, giving the largest values for the
number of hidden neurons in all the ANN models. This equation is also the only equation which is formulated in a manner that permits it to integrate “the number of training dataset” in its calculation scheme. It is worthwhile to note that the selected values in the last column are in complete agreement with the ones proposed by the Neuroshell 2 software for networks with two hidden layers.

Table 7-4: Initial number of hidden neurons in ANN models

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>Equation 1 $\sqrt{m \times n}$</th>
<th>Equation 2 $m$</th>
<th>Equation 3 $2 \times \sqrt{m} + n$</th>
<th>Equation 4 $2 \times m + 1$</th>
<th>Equation 5 $\frac{m + n}{2} + \sqrt{p}$</th>
<th>The selected value for initializing the number of hidden neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>2</td>
<td>1.41</td>
<td>2</td>
<td>3.83</td>
<td>5</td>
<td>12.5</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>1.73</td>
<td>3</td>
<td>4.46</td>
<td>7</td>
<td>13</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>9</td>
<td>13.5</td>
<td>12</td>
</tr>
<tr>
<td>5</td>
<td>2.24</td>
<td>5</td>
<td>5.47</td>
<td>11</td>
<td>14</td>
<td>12</td>
</tr>
<tr>
<td>6</td>
<td>2.45</td>
<td>6</td>
<td>5.90</td>
<td>13</td>
<td>14.5</td>
<td>14</td>
</tr>
<tr>
<td>7</td>
<td>2.65</td>
<td>7</td>
<td>6.29</td>
<td>15</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

$m$: number of input neurons (See Table 7-3)  
$n$: number of output neurons (See Table 7-3)  
$p$: number of samples in the training dataset ($p$ equals to 121 in this research)

7.5.2 Transfer or Activation Function

The transfer or activation function for the hidden layer(s) is ordinarily one of the sigmoid functions such as the logistic, the hyperbolic tangent, and the Gaussian functions. These functions are capable of inducing non-linearity into the network, which is one of the powerful characteristics of neural networks. For the output layer, when presented with continuous-valued targets with a bounded range, it has been suggested to use either logistic or hyperbolic tangent function (Bailey and Thompson, 1990b; Hegazy et al., 1994; Liu 2006).

In this research, neurons in both hidden and output layers of the ANN models were all activated by utilizing the S-shaped logistic transfer function. This function is the most widely used non-linear transfer function, which enables the ANN models to approximate any
continuous non-linear function. As described in the previous chapter, the logistic activation function maps the coming input signals with values from the interval \((-\infty, +\infty)\) into the output signal whose value lies within the interval \((0, 1)\). Figure 7-4 shows the shape of the logistic activation function, together with its mathematical expression.

![Logistic activation function](image)

**Figure 7-4**: Logistic activation function and its mathematical expression

### 7.5.3 Learning Algorithm

The back-propagation learning algorithm was described in Chapter 6 as the best known and most utilized type of supervised training method. Having been successfully used in a wide variety of applications, the back-propagation learning algorithm was also utilized in this research to train the ANN models.

As explained in the previous chapter, the performance of the networks trained by the back-propagation learning algorithm is controlled by a number of learning parameters embedded in this algorithm. The values of these parameters should be properly determined to ensure that the best performance is achieved for the ANN models. A practical approach to the determination of these parameters is presented in the following sections.
7.5.3.1 Initial Connection Weight

Before starting the training phase in the back-propagation learning algorithm, it is necessary to initialize the value of connection weights to a small non-zero value. Considering the general suggestions presented in the previous chapter, this value for all the connection weights was initially set to 0.3.

7.5.3.2 Scale Function

In order to make the training of an ANN model more efficient, the values of input variables are generally normalized to certain ranges of values. This normalization is accomplished by using a proper scale function. According to Chapter 6, the most appropriate scale function linearly scales the values of input variables into the range from minus one to one. This scale function was utilized in this research.

7.5.3.3 Learning Rate and Momentum

The learning rate is a parameter that affects how connection weights within a network are updated. These updates also include a portion of the last weight change, as defined by the momentum parameter, to simultaneously accelerate the training convergence and improve the training precision. The values of these parameters can be set to any amount from 0 to 1. As described in Chapter 6, a specific rule that yields the best set of values for these parameters is not available in the literature. Some problems tend to produce better results when using low values for these parameters, while others do better with higher values. Therefore, a suitable set of values for these parameters has to be determined through following a systematic trial and error approach.

In this research, a set of three MLP architectures was distinctively considered for each of the seven ANN models in Table 7-2. For each of the resulting twenty-one ANN models, the values of the learning rate and momentum parameters were iteratively changed from 0.1 to 0.9 in step of 0.1. In the case of each ANN model studied, this iteration resulted in the examination of 81 different sets of values for these parameters. As described later in this chapter, the set which yielded the best generalization performance was selected for each ANN model.
7.5.3.4 Stopping Criteria

As can be inferred from the discussion in Sections 6.6 and 6.7 of the previous chapter, the training process can constantly continue to make a model that works better and better on the training dataset. So, it is of crucial concern to know when a model has trained enough to yield reasonably accurate results for the dataset which was not used during training.

According to Chapter 6, most experts agree that a major reason that causes an ANN model to fail is overtraining or overlearning. If overlearning occurs, the model will then just memorize the training dataset. This type of models can master the training data, but fail when presented with new data. To overcome the overlearning problem, a technique called cross-validation has been widely used in previous studies (Alex et al., 2010; Al-Rashidi, 1999; Attalla and Hegazy, 2003; Hegazy and Ayed, 1998; Kim et al., 2004; Shehab et al., 2010; Williams, 1994; Williams et al., 2005). The cross-validation technique employing an entirely separate dataset (i.e., the selection dataset) to evaluate how well the network is predicting or classifying. What usually happens is that the error for the training dataset continues to get smaller forever, or at least gets to the point where it remains fairly unchanged. Alternatively, the error for the selection dataset continues to get smaller to the optimal point and then it slowly begins to get larger. Using the cross-validation technique would allow the network to be saved at this optimal point. At this point, the network is expected to generalize well on new data. The procedure described above is graphically illustrated in Figure 7-5.

![Figure 7-5: The general behaviour of training and selection datasets during the training procedure](image-url)
The cross-validation technique was implemented in the Neuroshell 2 software by means of calibration module. This module limits overlearning by preventing memorization to happen. When using the calibration, a network is first trained with a set of training samples and then, while learning is temporarily disabled, the current state of the trained network is applied to the selection dataset, and the error factor corresponding to this dataset is determined. This error factor is calculated as the mean squared error (MSE). The training process is repeated for several sets of training samples, each set yields a particular error factor for the selection dataset. These sets are composed of different training samples, but have the same size, which is defined by “calibration test interval”. In addition, each time that the training process is successfully accomplished for a set of training samples is determined as an “event”. The whole training process entails hundred thousands of these events. This process may eventually terminate when the number of the events, which result in no further improvement in minimizing the error factor of the selection dataset, exceeds a certain predefined amount. This amount is called “training stopping event” and should be defined in advance before the training process starts. In this research, the “calibration test interval” and “training stopping event” are set to 10 and 100,000, respectively.

7.6 Network Selection Criterion

The overall performance of a network is generally assessed in terms of two abilities, being the accuracy ability, and more importantly, the generalization ability. The former ability is evaluated over the training dataset, and shows how successful the network is in capturing the patterns in this dataset. The latter ability is, however, evaluated over the test dataset, and shows how successful the network is in capturing the patterns of the samples for which the network has never been trained. Although it is reasonable that a well-trained network has a similar performance over the training and test datasets (Jeon, 2007), the bottom line in deciding upon the selection criterion lies more within the generalization ability, rather than the accuracy ability. The main reason for this decision is due to the fact that the networks selected based on the generalization ability are likely to better predict the actual values of new cases that the network has never encountered before. The selection criterion in terms of generalization also ensures the potential problem of overlearning to be properly controlled and removed, which otherwise would normally cause the network to be of no further use. Considering the reasons stated above, the generalization ability was preferably
utilized in this research as the selection criterion to evaluate different networks within each model, and accordingly select the most suitable network.

7.7 Selection Indicators:

The coefficient of determination ($R^2$), mean squared error (MSE), and mean absolute error (MAE) are statistical indicators, which were employed to examine the generalization ability of the networks. Of these indicators, the former was mainly used in this research to compare the suitability of all the networks developed for a given ANN model, and accordingly to select the most appropriate network for that particular model. This indicator compares the precision of a given network to that of a benchmark network wherein the prediction is just the mean of all samples. A perfect fit would result in an $R^2$ value of 1, a very good fit near 1, and a very poor fit less than 0. If predictions made by a network are worse than what could be predicted by just using the mean of samples, then the $R^2$ value is less than 0. The formulas used for computing these indicators are as follows:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (7.1)$$

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n} \quad (7.2)$$

$$MAE = \frac{\sum|y_i - \hat{y}_i|}{n} \quad (7.3)$$

where $y_i$ is the actual value of the $i^{th}$ variable, $\hat{y}_i$ is the network predicted value of $y_i$, $\bar{y}$ is the mean of the $y_i$ values, and $n$ is the number of samples in either the training dataset (i.e., 121) or test dataset (i.e., 37).

7.8 Results

One of the main goals of this research is to develop the most suitable ANN model for each set of independent variables that were found in Chapter 5 to have a significant effect on the determination of the RNCC. This development consists of two main phases: (1) the preliminary and (2) the final phase. Each phase of this development has its specific objectives, which are discussed in the following sections.
7.8.1 Preliminary Development Phase

Two objectives were set out in the preliminary phase for the development of each of the seven ANN models in Table 7-2. These objectives are: (1) to select the most appropriate architecture of the models in terms of the number of hidden layers, and (2) to determine the most appropriate value for the learning parameters of the models in terms of the learning rate (LR) and momentum (M). In this phase, the number of hidden neurons in each model was initialized to its predetermined value, as presented in Table 7-4. For a given ANN model, this phase resulted in the development of a network which is henceforth called “the best primary network or model” for that particular model.

As described in Section 7-5-1, three sets of MLP networks with different architectures were considered for each ANN model. Also, different combinations of learning parameters, as described in Section 7-5-3-3, were iteratively examined during the training procedure of each set. For the last ANN model (i.e., model number seven), Tables 7-5, 7-6, and 7-7 present the results of this iteration for the MLP networks with one, two, and three hidden layers, respectively.

The most commonly used summary measures were calculated for the selection indicators, introduced in Section 7-7. The results of this calculation are presented at the end of each of the tables stated above for both the training and test datasets. These measures include minimum, maximum, mean, and coefficient of variation (C.O.V). By comparing the training and test datasets in terms of these measures, it was found that overlearning occurred in the case of some certain combinations of learning parameters. For instance, the values of C.O.V for $R^2$ in the test dataset were 4.14, 2.53, and 4.12 times greater than their respective values in the training dataset for the networks with one, two, and three hidden layers, respectively. These results imply that the relative variability of $R^2$ values in the training dataset was much lower than the relative variability of $R^2$ values in the test dataset. This implication is a consequence of the fact that the application of some certain combinations of learning parameters in a given network caused the network to show a completely different performance during the training and testing procedures. At the presence of these combinations, the network performed extremely well in predicting the samples in the training dataset. However, when being applied to the test dataset, the network poorly predicted the samples in this dataset. For instance, in Table 7-5, the greatest value of $R^2$ in the training dataset ($R^2 = 0.808$) was obtained for a network with LR= 0.3 and M= 0.9. The worst
prediction of the test samples was also obtained by the same network, resulting in the minimum value of $R^2$ in the test dataset ($R^2 = 0.079$). Therefore, while this network produced the best results for the training dataset, it produced the most inaccurate results for the test dataset. As a result, the network with $LR = 0.3$ and $M = 0.9$ was the least favoured network amongst all the networks examined in Table 7-5.

The performance of networks within the same architectural category was evaluated with reference to their relative generalization ability. In this evaluation, the learning parameters which resulted in the largest $R^2$ value over the test dataset were taken as the best learning parameters for a given architecture. Figures 7-6, 7-7, and 7-8 illustrate the best values of learning parameters for networks with respectively one, two, and three hidden layers, across the whole range of ANN models studied. Resultantly, for each ANN model, a group of three sets of networks with the best learning parameters were developed. Each of these networks within a given ANN model corresponded to a certain type of architecture. Table 7-8 summarises the characteristics of these networks for each model, and presents their performance over the training and test datasets, as measured in terms of $R^2$, MSE, and MAE.

Considering the number of ANN models developed in this research, the total number of the groups stated above amounted to seven. The accuracy and generalization abilities of the networks in each group are depicted and compared in Figures 7-9 and 7-10, respectively. According to this comparison, the order of networks priority within a given group is subject to change with respect to the type of the dataset being processed by these networks. It can be seen from Figures 7-9 and 7-10 that the network which represented the highest accuracy ability in a given group was not similar to the network which represented the highest generalization ability in that particular group. This observation holds true for all the groups but the one before the last (i.e., group number six), in which the order of the priority of networks remained intact for both the accuracy and generalization abilities. In other groups, however, the networks priority with respect to these abilities differed either slightly or substantially. For instance, in the first group (i.e., model number one), the network with two hidden layers was the most preferred network in terms of the accuracy ability, whereas the network with one hidden layer outperformed other networks in terms of the generalization ability. As another example, in the fourth group (i.e., model number four), the accuracy ability was constantly decreasing from the network with one hidden layer to the networks with two and three hidden layers, respectively. Nevertheless, the reverse is true when comparing the generalization ability of these networks.
Table 7-5: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “one hidden layer”

<table>
<thead>
<tr>
<th>Learning Rate (LR)</th>
<th>Momentum (M)</th>
<th>R² Training dataset (121 samples)</th>
<th>MSE Test dataset (37 samples)</th>
<th>MAE ** Test dataset (37 samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.7742</td>
<td>0.053</td>
<td>0.183</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.7449</td>
<td>0.06</td>
<td>0.191</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.75</td>
<td>0.059</td>
<td>0.189</td>
</tr>
<tr>
<td>0.4</td>
<td>0.1</td>
<td>0.7551</td>
<td>0.058</td>
<td>0.188</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
<td>0.7654</td>
<td>0.055</td>
<td>0.186</td>
</tr>
<tr>
<td>0.6</td>
<td>0.1</td>
<td>0.7741</td>
<td>0.053</td>
<td>0.183</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1</td>
<td>0.7294</td>
<td>0.064</td>
<td>0.195</td>
</tr>
<tr>
<td>0.8</td>
<td>0.1</td>
<td>0.7319</td>
<td>0.063</td>
<td>0.196</td>
</tr>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0.7449</td>
<td>0.06</td>
<td>0.19</td>
</tr>
<tr>
<td>0.1</td>
<td>0.2</td>
<td>0.7645</td>
<td>0.055</td>
<td>0.187</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td>0.7696</td>
<td>0.054</td>
<td>0.186</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>0.7722</td>
<td>0.054</td>
<td>0.185</td>
</tr>
<tr>
<td>0.4</td>
<td>0.2</td>
<td>0.7094</td>
<td>0.068</td>
<td>0.201</td>
</tr>
<tr>
<td>0.5</td>
<td>0.2</td>
<td>0.7397</td>
<td>0.061</td>
<td>0.193</td>
</tr>
<tr>
<td>0.6</td>
<td>0.2</td>
<td>0.7468</td>
<td>0.06</td>
<td>0.19</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.7683</td>
<td>0.055</td>
<td>0.188</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>0.7626</td>
<td>0.056</td>
<td>0.188</td>
</tr>
<tr>
<td>0.9</td>
<td>0.2</td>
<td>0.7217</td>
<td>0.066</td>
<td>0.196</td>
</tr>
<tr>
<td>0.1</td>
<td>0.3</td>
<td>0.7085</td>
<td>0.069</td>
<td>0.202</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3</td>
<td>0.7399</td>
<td>0.061</td>
<td>0.193</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3</td>
<td>0.7422</td>
<td>0.061</td>
<td>0.193</td>
</tr>
<tr>
<td>0.4</td>
<td>0.3</td>
<td>0.7404</td>
<td>0.061</td>
<td>0.19</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3</td>
<td>0.7448</td>
<td>0.06</td>
<td>0.189</td>
</tr>
<tr>
<td>0.6</td>
<td>0.3</td>
<td>0.7451</td>
<td>0.06</td>
<td>0.191</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3</td>
<td>0.758</td>
<td>0.057</td>
<td>0.189</td>
</tr>
<tr>
<td>0.8</td>
<td>0.3</td>
<td>0.7272</td>
<td>0.064</td>
<td>0.201</td>
</tr>
<tr>
<td>0.9</td>
<td>0.3</td>
<td>0.8075</td>
<td>0.045</td>
<td>0.171</td>
</tr>
<tr>
<td>0.1</td>
<td>0.4</td>
<td>0.7352</td>
<td>0.062</td>
<td>0.191</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.7382</td>
<td>0.062</td>
<td>0.19</td>
</tr>
<tr>
<td>0.3</td>
<td>0.4</td>
<td>0.7406</td>
<td>0.061</td>
<td>0.189</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
<td>0.7429</td>
<td>0.061</td>
<td>0.188</td>
</tr>
<tr>
<td>0.5</td>
<td>0.4</td>
<td>0.747</td>
<td>0.06</td>
<td>0.187</td>
</tr>
<tr>
<td>0.6</td>
<td>0.4</td>
<td>0.7567</td>
<td>0.057</td>
<td>0.184</td>
</tr>
<tr>
<td>0.7</td>
<td>0.4</td>
<td>0.698</td>
<td>0.071</td>
<td>0.209</td>
</tr>
<tr>
<td>0.8</td>
<td>0.4</td>
<td>0.7138</td>
<td>0.067</td>
<td>0.206</td>
</tr>
<tr>
<td>0.9</td>
<td>0.4</td>
<td>0.7432</td>
<td>0.06</td>
<td>0.193</td>
</tr>
</tbody>
</table>
Table 7-5: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “one hidden layer” (Continued)

<table>
<thead>
<tr>
<th>Learning Parameters</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training dataset (121 samples)</td>
</tr>
<tr>
<td></td>
<td>( R^2 )</td>
</tr>
<tr>
<td>Learning Rate (LR)</td>
<td>Momentum (M)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
</tr>
<tr>
<td>0.6</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
</tr>
<tr>
<td>0.8</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
</tr>
</tbody>
</table>
Table 7-5: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “one hidden layer” (Continued)

<table>
<thead>
<tr>
<th>Learning Parameters</th>
<th>Results</th>
<th>Training dataset (121 samples)</th>
<th>Test dataset (37 samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R²</td>
<td>MSE*</td>
</tr>
<tr>
<td><strong>Learning Rate (LR)</strong></td>
<td><strong>Momentum (M)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.7209</td>
<td>0.066</td>
<td>0.2</td>
</tr>
<tr>
<td>0.2</td>
<td>0.722</td>
<td>0.065</td>
<td>0.2</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7221</td>
<td>0.065</td>
<td>0.2</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7214</td>
<td>0.066</td>
<td>0.201</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7197</td>
<td>0.066</td>
<td>0.202</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7081</td>
<td>0.069</td>
<td>0.205</td>
</tr>
<tr>
<td>0.7</td>
<td>0.6978</td>
<td>0.071</td>
<td>0.206</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7338</td>
<td>0.063</td>
<td>0.194</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7876</td>
<td>0.05</td>
<td>0.178</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Mean</th>
<th>Coefficient of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.698</td>
<td>0.736</td>
<td>3.351</td>
</tr>
<tr>
<td>0.045</td>
<td>0.062</td>
<td>9.409</td>
</tr>
<tr>
<td>0.171</td>
<td>0.209</td>
<td>4.302</td>
</tr>
<tr>
<td>0.079</td>
<td>0.590</td>
<td>13.859</td>
</tr>
<tr>
<td>0.093</td>
<td>0.210</td>
<td>14.508</td>
</tr>
<tr>
<td>0.249</td>
<td>0.346</td>
<td>5.471</td>
</tr>
</tbody>
</table>

● Mean Squared Error (MSE)
●● Mean Absolute Error (MAE)

Note: Shaded are the most preferred values of learning parameters, together with their resulting statistics.
Table 7-6: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “two hidden layers”

<table>
<thead>
<tr>
<th>Learning Rate (LR)</th>
<th>Momentum (M)</th>
<th>Training dataset (121 samples)</th>
<th>Test dataset (37 samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>R²</td>
<td>MSE</td>
</tr>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.7672</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.7652</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.7713</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.7773</td>
<td>0.052</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.7643</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.7723</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7655</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.7588</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.7639</td>
<td>0.056</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.7713</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.7696</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.7554</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.7669</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.7536</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.7635</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7276</td>
<td>0.064</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.7701</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.7409</td>
<td>0.061</td>
</tr>
<tr>
<td>0.3</td>
<td>0.1</td>
<td>0.7576</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.7645</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.7672</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.7727</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.7168</td>
<td>0.067</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.733</td>
<td>0.063</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7424</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.7625</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.7244</td>
<td>0.065</td>
</tr>
<tr>
<td>0.4</td>
<td>0.1</td>
<td>0.7683</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.7723</td>
<td>0.054</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.7603</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.7655</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.7381</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.7454</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7523</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.7261</td>
<td>0.065</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.8146</td>
<td>0.044</td>
</tr>
</tbody>
</table>
Table 7-6: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “two hidden layers” (Continued)

<table>
<thead>
<tr>
<th>Learning Parameters</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training dataset (121 samples)</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
</tr>
<tr>
<td>Learning Rate (LR)</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.7706</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7619</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7546</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7316</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7472</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7614</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7451</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7457</td>
</tr>
<tr>
<td>0.9</td>
<td><strong>0.8313</strong></td>
</tr>
<tr>
<td>0.6</td>
<td><strong>0.8313</strong></td>
</tr>
<tr>
<td>0.7</td>
<td><strong>0.8313</strong></td>
</tr>
<tr>
<td>0.8</td>
<td><strong>0.8313</strong></td>
</tr>
<tr>
<td>0.9</td>
<td><strong>0.8313</strong></td>
</tr>
<tr>
<td>0.1</td>
<td>0.7643</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7639</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7517</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7612</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7749</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7423</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7478</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7324</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8181</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7638</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7379</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7459</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7554</td>
</tr>
<tr>
<td>0.1</td>
<td>0.7545</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7416</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7508</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6732</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7559</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7357</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7436</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7535</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7515</td>
</tr>
<tr>
<td>0.1</td>
<td>0.7758</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7502</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7315</td>
</tr>
<tr>
<td>0.4</td>
<td>0.805</td>
</tr>
<tr>
<td>0.5</td>
<td>0.783</td>
</tr>
</tbody>
</table>
Table 7-6: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “two hidden layers” (Continued)

| Learning Parameters | Results | | | |
|---------------------|---------|---------|---------|---------|---------|---------|---------|---------|
|                     | Training dataset (121 samples) | Test dataset (37 samples) |
|                     | R²  | MSE* | MAE** | R²  | MSE* | MAE** |
| Learning Rate (LR) | Momentum (M) | | | | | |
|                     | 0.1  | 0.7401 | 0.061 | 0.197 | 0.5174 | 0.110 | 0.282 |
|                     | 0.2  | 0.748 | 0.059 | 0.194 | 0.5159 | 0.110 | 0.282 |
|                     | 0.3  | 0.7414 | 0.061 | 0.192 | 0.4992 | 0.114 | 0.289 |
|                     | 0.4  | 0.7551 | 0.058 | 0.187 | 0.5149 | 0.110 | 0.284 |
|                     | 0.5  | 0.7606 | 0.056 | 0.186 | 0.5677 | 0.098 | 0.262 |
|                     | 0.6  | 0.7535 | 0.058 | 0.191 | 0.5265 | 0.108 | 0.280 |
|                     | 0.7  | 0.7248 | 0.065 | 0.2 | 0.5490 | 0.103 | 0.271 |
|                     | 0.8  | 0.8052 | 0.046 | 0.167 | 0.5971 | 0.092 | 0.245 |
|                     | 0.9  | 0.8079 | 0.045 | 0.174 | 0.5175 | 0.110 | 0.276 |

Minimum | 0.673 | 0.040 | 0.152 | 0.446 | 0.061 | 0.193 |
Maximum | 0.831 | 0.077 | 0.214 | 0.734 | 0.126 | 0.311 |
Mean | 0.757 | 0.057 | 0.188 | 0.554 | 0.102 | 0.267 |

* Mean Squared Error (MSE)
** Mean Absolute Error (MAE)

Note: Shaded are the most preferred values of learning parameters, together with their resulting statistics.
Table 7-7: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “three hidden layers”

<table>
<thead>
<tr>
<th>Learning Parameters</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training dataset (121 samples)</td>
</tr>
<tr>
<td></td>
<td>R²</td>
</tr>
<tr>
<td>Learning Rate (LR)</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.7504</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7602</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7598</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7489</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7598</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7675</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7769</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7359</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7464</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7605</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7693</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7728</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7728</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7192</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7247</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7571</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7703</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7141</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7223</td>
</tr>
<tr>
<td>0.5</td>
<td>0.7248</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7253</td>
</tr>
<tr>
<td>0.7</td>
<td>0.752</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7347</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7142</td>
</tr>
<tr>
<td>0.4</td>
<td>0.7333</td>
</tr>
<tr>
<td>0.5</td>
<td>0.733</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7295</td>
</tr>
<tr>
<td>0.7</td>
<td>0.7588</td>
</tr>
<tr>
<td>0.8</td>
<td>0.7535</td>
</tr>
<tr>
<td>0.9</td>
<td>0.7532</td>
</tr>
</tbody>
</table>
Table 7-7: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “three hidden layers” (Continued)

<table>
<thead>
<tr>
<th>Learning Rate (LR)</th>
<th>Momentum (M)</th>
<th>Learning Parameters</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Training dataset (121 samples)</td>
<td>Test dataset (37 samples)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$R^2$</td>
<td>MSE</td>
</tr>
<tr>
<td>0.6</td>
<td>0.1</td>
<td>0.7372</td>
<td>0.062</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.7433</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.7489</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.7538</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.7582</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.7599</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7494</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.6884</td>
<td>0.073</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.7796</td>
<td>0.052</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1</td>
<td>0.7408</td>
<td>0.061</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.7492</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.7557</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.7591</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.7582</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.7014</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.8563</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.659</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.809</td>
<td>0.045</td>
</tr>
<tr>
<td>0.8</td>
<td>0.1</td>
<td>0.7509</td>
<td>0.059</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.7448</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.3</td>
<td>0.7544</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.7583</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.7549</td>
<td>0.058</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.7593</td>
<td>0.057</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0.7832</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.7663</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>0.9</td>
<td>0.7453</td>
<td>0.06</td>
</tr>
</tbody>
</table>
Table 7-7: Results of different combinations of learning parameters in the last ANN model (model number “seven”) by utilizing MLP networks with “three hidden layers” (Continued)

<table>
<thead>
<tr>
<th>Learning Parameters</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training dataset (121 samples)</td>
</tr>
<tr>
<td></td>
<td>R²</td>
</tr>
<tr>
<td>Learning Rate (LR)</td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>0.7467</td>
</tr>
<tr>
<td>0.2</td>
<td>0.7607</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7484</td>
</tr>
<tr>
<td>0.4</td>
<td>0.6716</td>
</tr>
<tr>
<td>0.5</td>
<td>0.6714</td>
</tr>
<tr>
<td>0.6</td>
<td>0.7565</td>
</tr>
<tr>
<td>0.7</td>
<td>0.6702</td>
</tr>
<tr>
<td>0.8</td>
<td>0.6778</td>
</tr>
<tr>
<td>0.9</td>
<td>0.8385</td>
</tr>
<tr>
<td>Minimum</td>
<td>0.659</td>
</tr>
<tr>
<td>Maximum</td>
<td>0.856</td>
</tr>
<tr>
<td>Mean</td>
<td>0.748</td>
</tr>
</tbody>
</table>

* Mean Squared Error (MSE)  
** Mean Absolute Error (MAE)  
Note: Shaded are the most preferred values of learning parameters, together with their resulting statistics.
Table 7-8: The most desirable set of networks (of different architectures) for each of ANN models studied

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>Network Architecture</th>
<th>No. of Hidden Neurons</th>
<th>Learning Parameters</th>
<th>Results Training dataset (121 samples)</th>
<th>Results Test dataset (37 samples)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LR</td>
<td>M</td>
<td>R²</td>
</tr>
<tr>
<td>1</td>
<td>One Hidden Layer</td>
<td>12</td>
<td>0.5</td>
<td>0.2</td>
<td>0.578</td>
</tr>
<tr>
<td></td>
<td>Two Hidden Layers</td>
<td>12</td>
<td>0.6</td>
<td>0.8</td>
<td>0.583</td>
</tr>
<tr>
<td></td>
<td>Three Hidden Layers</td>
<td>12</td>
<td>0.4</td>
<td>0.8</td>
<td>0.562</td>
</tr>
<tr>
<td>2</td>
<td>One Hidden Layer</td>
<td>12</td>
<td>0.5</td>
<td>0.9</td>
<td>0.624</td>
</tr>
<tr>
<td></td>
<td>Two Hidden Layers</td>
<td>12</td>
<td>0.4</td>
<td>0.9</td>
<td>0.683</td>
</tr>
<tr>
<td></td>
<td>Three Hidden Layers</td>
<td>12</td>
<td>0.7</td>
<td>0.9</td>
<td>0.608</td>
</tr>
<tr>
<td>3</td>
<td>One Hidden Layer</td>
<td>12</td>
<td>0.6</td>
<td>0.7</td>
<td>0.701</td>
</tr>
<tr>
<td></td>
<td>Two Hidden Layers</td>
<td>12</td>
<td>0.9</td>
<td>0.2</td>
<td>0.701</td>
</tr>
<tr>
<td></td>
<td>Three Hidden Layers</td>
<td>12</td>
<td>0.3</td>
<td>0.7</td>
<td>0.695</td>
</tr>
<tr>
<td>4</td>
<td>One Hidden Layer</td>
<td>12</td>
<td>0.7</td>
<td>0.9</td>
<td>0.746</td>
</tr>
<tr>
<td></td>
<td>Two Hidden Layers</td>
<td>12</td>
<td>0.6</td>
<td>0.6</td>
<td>0.734</td>
</tr>
<tr>
<td></td>
<td>Three Hidden Layers</td>
<td>12</td>
<td>0.8</td>
<td>0.1</td>
<td>0.687</td>
</tr>
<tr>
<td>5</td>
<td>One Hidden Layer</td>
<td>12</td>
<td>0.4</td>
<td>0.9</td>
<td>0.763</td>
</tr>
<tr>
<td></td>
<td>Two Hidden Layers</td>
<td>12</td>
<td>0.5</td>
<td>0.2</td>
<td>0.795</td>
</tr>
<tr>
<td></td>
<td>Three Hidden Layers</td>
<td>12</td>
<td>0.3</td>
<td>0.7</td>
<td>0.801</td>
</tr>
<tr>
<td>6</td>
<td>One Hidden Layer</td>
<td>14</td>
<td>0.9</td>
<td>0.8</td>
<td>0.796</td>
</tr>
<tr>
<td></td>
<td>Two Hidden Layers</td>
<td>14</td>
<td>0.8</td>
<td>0.9</td>
<td>0.812</td>
</tr>
<tr>
<td></td>
<td>Three Hidden Layers</td>
<td>14</td>
<td>0.4</td>
<td>0.7</td>
<td>0.774</td>
</tr>
<tr>
<td>7</td>
<td>One Hidden Layer</td>
<td>16</td>
<td>0.1</td>
<td>0.2</td>
<td>0.745</td>
</tr>
<tr>
<td></td>
<td>Two Hidden Layers</td>
<td>16</td>
<td>0.5</td>
<td>0.9</td>
<td>0.831</td>
</tr>
<tr>
<td></td>
<td>Three Hidden Layers</td>
<td>16</td>
<td>0.1</td>
<td>0.3</td>
<td>0.740</td>
</tr>
</tbody>
</table>

Note: For each ANN model, the best primary network is shaded.
Figure 7-6: The best learning parameters along with their respective $R^2$ values for “training” and “test” datasets (networks with “one hidden layer”)

LR= Learning Rate  
M= Momentum  
N= Number of Hidden Neurons

- LR=0.5  
  M=0.2  
  N=12

- LR=0.6  
  M=0.7  
  N=12

- LR=0.4  
  M=0.9  
  N=12

- LR=0.1  
  M=0.2  
  N=16

- LR=0.6  
  M=0.9  
  N=12

- LR=0.7  
  M=0.9  
  N=12

- LR=0.9  
  M=0.8  
  N=14
Figure 7-7: The best learning parameters along with their respective $R^2$ values for “training” and “test” datasets (networks with “two hidden layers”)

- **LR= Learning Rate**
- **M= Momentum**
- **N= Number of Hidden Neurons**

- **LR=0.6**, **M=0.8**, **N=12**
- **LR=0.9**, **M=0.2**, **N=12**
- **LR=0.5**, **M=0.2**, **N=12**
- **LR=0.5**, **M=0.9**, **N=16**
- **LR=0.4**, **M=0.9**, **N=12**
- **LR=0.6**, **M=0.9**, **N=12**
- **LR=0.8**, **M=0.9**, **N=14**
Figure 7-8: The best learning parameters along with their respective $R^2$ values for “training” and “test” datasets (networks with “three hidden layers”)

- **LR** = Learning Rate
- **M** = Momentum
- **N** = Number of Hidden Neurons

- **Training Dataset**
  - LR=0.4, M=0.8, N=12
  - LR=0.7, M=0.9, N=12
  - LR=0.8, M=0.1, N=12
- **Test Dataset**
  - LR=0.3, M=0.7, N=12
  - LR=0.3, M=0.7, N=12
  - LR=0.4, M=0.7, N=16
  - LR=0.1, M=0.3, N=16
Figure 7-9: Comparison of the “accuracy” ability of the networks with the best learning parameters within each model

Note: The networks are arranged in descending order of priority with respect to their relative “Accuracy” performance.
Figure 7-10: Comparison of the “generalization” ability of the networks with the best learning parameters within each model

Note: The networks are arranged in descending order of priority with respect to their relative “Generalization” performance.
For each model, the best primary network was selected according to the selection criterion, which was determined earlier in Section 7-6. Following this criterion, for each model, the network which exhibited the foremost level of generalization ability was selected. Figure 7-11 illustrates the performance of the selected networks over the training and test datasets. Table 7-9 summarizes the results of the preliminary development phase. This table presents the best architecture of each model, along with the most suitable values found for the learning parameters in each model.

Table 7-9: The best architectures of ANN models along with their corresponding most suitable learning parameters

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>Best architecture</th>
<th>Best learning parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Learning Rate (LR)</td>
</tr>
<tr>
<td>1</td>
<td>One hidden layer</td>
<td>0.5</td>
</tr>
<tr>
<td>2</td>
<td>Two hidden layers</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>Two hidden layers</td>
<td>0.9</td>
</tr>
<tr>
<td>4</td>
<td>Three hidden layers</td>
<td>0.8</td>
</tr>
<tr>
<td>5</td>
<td>One hidden layer</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>Two hidden layers</td>
<td>0.8</td>
</tr>
<tr>
<td>7</td>
<td>Two hidden layers</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Figure 7-11: Characteristics of the best primary networks along with their respective $R^2$ values for “training” and “test” datasets

- MLP-1 = MLP Network with One Hidden Layer
- MLP-2 = MLP Network with Two Hidden Layers
- MLP-3 = MLP Network with Three Hidden Layers
- LR = Learning Rate
- M = Momentum
- N = Number of Hidden Neurons
7.8.2 Final Development Phase

The purpose of the final phase was to determine the most appropriate number of hidden neurons in each of the best primary networks developed in the previous phase. This determination would allow these networks to improve further and reach their best performance in terms of the generalization ability.

As described in Chapter 6, the adjustment of hidden neurons is more of an art than science (Shtub and Versano, 1999), as neither a systematic approach nor formal quantitative method has been widely accepted as the best technique for determining the best number of hidden neurons. This number is therefore determined through the use of a trial-and-error procedure.

A wide range of hidden neurons was examined during the final development phase. The range of hidden neurons examined was from a minimum of 1 to a maximum of 50 for the best primary networks developed in the preliminary phase. For instance, Table 7-10 presents the results of this examination for the best primary network developed for the last ANN model. Notably, the networks shown in this table were all of the same architectural type (i.e., MLP network with two hidden layers) and their learning parameters were all set to the best values found in the previous phase (i.e., LR= 0.5, M= 0.9). In Table 7-10, for each number of hidden neurons examined, the values of the selection indicators are provided for both training and test datasets.
Table 7-10: Results of iteration of the number of hidden neurons on the performance of the best primary network developed for the last ANN model (model number “seven”)

(MLP network with two hidden layers; LR= 0.5; M= 0.9)

<table>
<thead>
<tr>
<th>No. of Hidden Neurons</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training dataset (121 samples)</td>
</tr>
<tr>
<td></td>
<td>$R^2$</td>
</tr>
<tr>
<td>2</td>
<td>0.5524</td>
</tr>
<tr>
<td>3</td>
<td>0.7276</td>
</tr>
<tr>
<td>4</td>
<td>0.669</td>
</tr>
<tr>
<td>5</td>
<td>0.7858</td>
</tr>
<tr>
<td>6</td>
<td>0.7825</td>
</tr>
<tr>
<td>7</td>
<td>0.813</td>
</tr>
<tr>
<td>8</td>
<td>0.6271</td>
</tr>
<tr>
<td>9</td>
<td>0.7095</td>
</tr>
<tr>
<td>10</td>
<td>0.8053</td>
</tr>
<tr>
<td>11</td>
<td>0.8243</td>
</tr>
<tr>
<td>12</td>
<td>0.8551</td>
</tr>
<tr>
<td>13</td>
<td>0.8136</td>
</tr>
<tr>
<td>14</td>
<td>0.8576</td>
</tr>
<tr>
<td>15</td>
<td>0.9043</td>
</tr>
<tr>
<td>16</td>
<td>0.8313</td>
</tr>
<tr>
<td>17</td>
<td>0.8266</td>
</tr>
<tr>
<td>18</td>
<td>0.6976</td>
</tr>
<tr>
<td>19</td>
<td>0.7083</td>
</tr>
<tr>
<td>20</td>
<td>0.7124</td>
</tr>
<tr>
<td>21</td>
<td>0.7082</td>
</tr>
<tr>
<td>22</td>
<td>0.7237</td>
</tr>
<tr>
<td>23</td>
<td>0.8766</td>
</tr>
<tr>
<td>24</td>
<td>0.7278</td>
</tr>
<tr>
<td>25</td>
<td>0.736</td>
</tr>
<tr>
<td>26</td>
<td>0.7098</td>
</tr>
<tr>
<td>27</td>
<td>0.8551</td>
</tr>
<tr>
<td>28</td>
<td>0.6832</td>
</tr>
<tr>
<td>29</td>
<td>0.7996</td>
</tr>
<tr>
<td>30</td>
<td>0.7006</td>
</tr>
<tr>
<td>31</td>
<td>0.7613</td>
</tr>
<tr>
<td>32</td>
<td>0.6823</td>
</tr>
<tr>
<td>33</td>
<td>0.755</td>
</tr>
<tr>
<td>34</td>
<td>0.7532</td>
</tr>
<tr>
<td>35</td>
<td>0.6932</td>
</tr>
<tr>
<td>36</td>
<td>0.7128</td>
</tr>
</tbody>
</table>
Table 7-10: Results of iteration of the number of hidden neurons on the performance of the best primary network developed for the last ANN model (model number “seven”) (MLP network with two hidden layers; LR= 0.5; M= 0.9) (Continued)

<table>
<thead>
<tr>
<th>No. of Hidden Neurons</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training dataset (121 samples)</td>
</tr>
<tr>
<td></td>
<td>(R^2)</td>
</tr>
<tr>
<td>37</td>
<td>0.7619</td>
</tr>
<tr>
<td>38</td>
<td>0.7523</td>
</tr>
<tr>
<td>39</td>
<td>0.7546</td>
</tr>
<tr>
<td>40</td>
<td>0.7706</td>
</tr>
<tr>
<td>41</td>
<td>0.7784</td>
</tr>
<tr>
<td>42</td>
<td>0.7454</td>
</tr>
<tr>
<td>43</td>
<td>0.7502</td>
</tr>
<tr>
<td>44</td>
<td>0.7551</td>
</tr>
<tr>
<td>45</td>
<td>0.8482</td>
</tr>
<tr>
<td>46</td>
<td>0.7514</td>
</tr>
<tr>
<td>47</td>
<td>0.8882</td>
</tr>
<tr>
<td>48</td>
<td>0.7528</td>
</tr>
<tr>
<td>49</td>
<td>0.7647</td>
</tr>
<tr>
<td>50</td>
<td>0.6755</td>
</tr>
</tbody>
</table>

| Minimum              | 0.552  | 0.023 | 0.113 | -0.117 | 0.061 | 0.193 |
| Maximum              | 0.904  | 0.105 | 0.232 | 0.733  | 0.254 | 0.363 |
| Mean                 | 0.758  | 0.057 | 0.182 | 0.530  | 0.107 | 0.266 |

* Mean Squared Error (MSE)  
** Mean Absolute Error (MAE)  

Note: Shaded is the most preferred number of hidden neurons, together with its resulting statistics.

Listed at the end of the table above are the summary measures of the selection indicators in both training and test datasets. By comparing these datasets in terms of their corresponding summary measures, it is evident that some networks should suffer from the problem of overlearning. This problem is just similar to the problem that happened in the preliminary development phase. In Table 7-10, for example, the C.O.V of \(R^2\) values in the test dataset equates to 24.42% which is 2.67 times greater than the C.O.V of \(R^2\) values in the training dataset (i.e., 9.14%). As in the preliminary development phase, such a difference in the final development phase arises because the relative variability of \(R^2\) values in the training dataset was much lower than the relative variability of \(R^2\) values in the test dataset. For instance, when the number of hidden neurons was set to 23, the third largest value of \(R^2\) in
the training dataset ($R^2 = 0.877$) was obtained. Nevertheless, the least $R^2$ value in the test dataset was also attributed to this number of hidden neurons ($R^2 = -0.117$). In fact, the network with this number of hidden neurons was unable to even predict the test samples as well as what could be predicted by just using the mean of these samples.

The variations of $R^2$ values in the training and test datasets over the whole range of the hidden neurons, which were examined for the best primary networks, are depicted in Figures 7-12 through 7-18. These figures are further examined in Section 7-13 to investigate the effect of the number of input variables on overlearning.

For each of the best primary networks developed in the preliminary development phase, the effect of the number of hidden neurons was investigated in accordance with the selection criterion. Following this criterion, the best number of hidden neurons which allowed a given network to reach its highest level of generalization ability was selected and tabulated in Table 7-11.

Table 7-11: The best numbers of neurons in the hidden layer(s) of ANN models

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>The best number of hidden neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>39</td>
</tr>
<tr>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>48</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
</tr>
</tbody>
</table>
Figure 7-12: Comparison of $R^2$ values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “one”)
Figure 7-13: Comparison of $R^2$ values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “two”)
Figure 7-14: Comparison of $R^2$ values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “three”).
Figure 7-15: Comparison of $R^2$ values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “four”).

The best architecture found for the 4th ANN model = MLP network with Three Hidden Layers
The best respective learning rate (LR) value = 0.8
The best respective momentum (M) value = 0.1
Figure 7-16: Comparison of $R^2$ values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “five”)

The best architecture found for the 5th ANN model = MLP network with One Hidden Layer
The best respective learning rate (LR) value = 0.4
The best respective momentum (M) value = 0.9
Figure 7-17: Comparison of $R^2$ values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “six”)

The best architecture found for the 6th ANN model = MLP network with Two Hidden Layers
- The best respective learning rate (LR) value = 0.8
- The best respective momentum (M) value = 0.9
Figure 7-18: Comparison of $R^2$ values in “training” and “test” datasets with respect to variation of the number of hidden neurons in the best primary network (model number “seven”).
7.9 The Final Best ANN Models

Following the accomplishment of the preliminary and final development phases, the final best ANN models were developed. The characteristics of these models, together with their performance when being applied to the training and test datasets are summarized in the following table.

Table 7-12: Characteristics of the final best ANN models

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>Best architecture</th>
<th>Best learning parameters</th>
<th>Best no. of hidden neurons</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>LR</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>One hidden layer</td>
<td>0.5</td>
<td>0.2</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>Two hidden layers</td>
<td>0.4</td>
<td>0.9</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>Two hidden layers</td>
<td>0.9</td>
<td>0.2</td>
<td>29</td>
</tr>
<tr>
<td>4</td>
<td>Three hidden layers</td>
<td>0.8</td>
<td>0.1</td>
<td>39</td>
</tr>
<tr>
<td>5</td>
<td>One hidden layer</td>
<td>0.4</td>
<td>0.9</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>Two hidden layers</td>
<td>0.8</td>
<td>0.9</td>
<td>48</td>
</tr>
<tr>
<td>7</td>
<td>Two hidden layers</td>
<td>0.5</td>
<td>0.9</td>
<td>16</td>
</tr>
</tbody>
</table>

7.10 Comparison of Architectures

The most suitable architectures of the final best ANN models were basically determined in the preliminary development phase. According to Table 7-12, the architecture which was composed of two hidden layers is dominant in four of the seven ANN models developed in this research (i.e., ANN models number two, three, six, and seven). The next dominant architectures were composed of one and three hidden layers, respectively. The former architecture was the most preferred architecture in two of the three remaining models (i.e., ANN models number one, and five), whereas the latter architecture was fitted best with just one model (i.e., ANN model number four).

Considering the performance of different architectures examined in this research, it is highly recommendable to take the advantage of MLP networks with more than one hidden layer in an ANN modelling. In this regard, the use of MLP networks with two hidden layers is of more benefit.
### 7.11 Appropriateness of the Initialized Numbers of Hidden Neurons

The results obtained during the fulfilment of the final development phase made it possible to explicitly examine the suitability of the methodology utilized in this research for the initialization of the number of hidden neurons. By using this methodology, the initial value of the hidden neurons for each ANN model was calculated and tabulated in the last column of Table 7-4. In each group of the networks examined for a given ANN model in the preliminary development phase, the number of hidden neurons was fixed to its initial value. In the final development phase, however, the number of hidden neurons was iterated and the most suitable number was determined for each of the best primary networks developed in the previous phase. For each of the best primary networks, the suitability of the initial number of hidden neurons was compared to that of other numbers examined during the final development phase. The results of this comparison are presented in Table 7-13.

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>Initial value of the no. of hidden neurons (being used in the preliminary phase)</th>
<th>Total no. of hidden neurons (examined in the final phase)</th>
<th>The suitability rank of the initial value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12</td>
<td>50</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>2</td>
<td>12</td>
<td>49</td>
<td>11&lt;sup&gt;th&lt;/sup&gt;</td>
</tr>
<tr>
<td>3</td>
<td>12</td>
<td>49</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>48</td>
<td>6&lt;sup&gt;th&lt;/sup&gt;</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>50</td>
<td>3&lt;sup&gt;rd&lt;/sup&gt;</td>
</tr>
<tr>
<td>6</td>
<td>14</td>
<td>49</td>
<td>2&lt;sup&gt;nd&lt;/sup&gt;</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>49</td>
<td>1&lt;sup&gt;st&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

According to the above table, out of the seven values utilized for the initialization of the hidden neurons, five values were ranked amongst the top third of the best performing values which secured the highest generalization ability for their respective ANN models. Therefore, the methodology used in Section 7-5-1-2 is robust for the identification of the most appropriate number of hidden neurons.
7.12 Learning Parameters vs. the Number of Hidden Neurons

Adjusting the designing parameters of a network to their most suitable values is an extremely time-consuming task. These parameters are normally represented by the network size parameter (in terms of the number of hidden neurons) and the network learning parameters (in terms of the learning rate and the momentum). To accelerate the procedure of an ANN modelling, it is sometimes indispensable to accomplish this adjustment for only either the network size parameter or the network learning parameters. Preferably, the parameter should be selected that the performance of a network is more sensitive to. This parameter should receive higher priority because it is more likely to have a greater influence on the network performance. This selection is an issue that has not been broadly investigated and little, if any, can be found in the literature. Therefore, it is of the interest of this research to also inspect the sensitivity of the performance of the networks to these designing parameters.

An extensive trial and error process was carried out to evaluate the effect of, first, various combinations of learning parameters, and, second, a wide range of numbers of hidden neurons on the overall performance of the networks. Table 7-14 presents the variability of the $R^2$ values in the networks obtained at the end of each development phase. This variability is measured using the coefficient of variation (i.e., C.O.V), and it is shown for both the training and test datasets. As described earlier, in the preliminary development phase, the values of learning parameters were iteratively changed, while the numbers of the hidden neurons were fixed to their initial values. In the final development phase, however, the numbers of the hidden neurons were iteratively changed, while the values of learning parameters were fixed to their best values, as obtained in the previous phase.
Table 7-14: Comparison of the networks sensitivity with respect to learning parameters and the number of hidden neurons

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>Architecture of the model</th>
<th>Sensitivity with respect to learning parameters (Studied in preliminary phase)</th>
<th>Sensitivity with respect to no. of hidden neurons (Studied in final phase)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>C.O.V of $R^2_{\text{Training}}$</td>
<td>C.O.V of $R^2_{\text{Test}}$</td>
</tr>
<tr>
<td>1</td>
<td>One hidden layer</td>
<td>3.05%</td>
<td>3.54%</td>
</tr>
<tr>
<td>2</td>
<td>Two hidden layers</td>
<td>3.52%</td>
<td>6.63%</td>
</tr>
<tr>
<td>3</td>
<td>Two hidden layers</td>
<td>3.29%</td>
<td>8.06%</td>
</tr>
<tr>
<td>4</td>
<td>Three hidden layers</td>
<td>4.22%</td>
<td>14.32%</td>
</tr>
<tr>
<td>5</td>
<td>One hidden layer</td>
<td>2.92%</td>
<td>9.69%</td>
</tr>
<tr>
<td>6</td>
<td>Two hidden layers</td>
<td>3.43%</td>
<td>5.89%</td>
</tr>
<tr>
<td>7</td>
<td>Two hidden layers</td>
<td>3.06%</td>
<td>7.74%</td>
</tr>
</tbody>
</table>

* Coefficient of Variation (C.O.V)

According to the above table, the $R^2$ values in the test dataset are more sensitive than their respective $R^2$ values in the training dataset to both the learning parameters and the number of hidden neurons. In addition, yet more importantly, it can be seen in the Table 7-14 that the variation of $R^2$ values in a given model is more dominated by the number of hidden neurons than the learning parameters. For instance, by iterating the values of learning parameters for the last model in the preliminary development phase, the C.O.Vs of $R^2$ in the training and test datasets amounted to 3.06% and 7.74%, respectively. These values are almost three times less than their respective values (i.e. 9.14% and 24.42%), obtained for the same model in the final development phase where the number of hidden neurons was iterated. Almost the same results were obtained for all models but two, being model number one and model number four. These networks had some particular features that make them distinct from other models. Model number one was the only model that appeared to be more competent in the detection of the underlying pattern of the test samples rather than the training samples. For instance, for every single number of hidden neurons examined during the final development phase, this model produced higher $R^2$ values for the test dataset than the training dataset (see Figure 7-12). In all other models, the reverse is true, as shown in Figures 7-13 to 7-18. Further, model number four is different from other models because of its unique architecture, which is composed of three hidden layers. Having become known at the end of the preliminary development phase, the architectures of other models included
dissimilar numbers of hidden layers. The architecture of these models was basically composed of one or two hidden layers, of which the latter prevailed.

What can be inferred from the discussion above are mainly two issues. First, the number of hidden neurons and the learning parameters affect more seriously the generalization ability than the accuracy ability. Second, these abilities are likely to be far more sensitive to the number of hidden neurons than the learning parameters.

7.13 Effect of the Number of Input Variables on Overlearning

As described in Chapter 6, perhaps the most important factor which mainly affects the overlearning is the number of hidden neurons. Therefore, the results of the final development phase were further explored to investigate the severity of the overlearning in ANN models developed in this study. To this end, the variation of $R^2$ values in the training and test datasets was first explored over the whole range of the hidden neurons examined during the final development phase. Then, the most unfavourable number of hidden neurons was selected for each model. In this selection, the number which resulted in the largest overlearning rate for a given model was considered as the most unfavourable number. For any number examined for the hidden neurons, the overlearning rate was defined as the difference between the $R^2$ values of the training and test datasets (see Equation (7.4)).

\[
\text{Overlearning Rate} = R^2_{\text{Training}} - R^2_{\text{Test}}
\] (7.4)

Table 7-15 lists the overlearning rates for all the hidden neurons examined in the last best primary model (see Table 7-10 and Figure 7-18).
Table 7-15: Comparison of overlearning rates within the last best primary model
(MLP network with two hidden layers; LR=0.5; M=0.9)

<table>
<thead>
<tr>
<th>No. of Hidden Neurons</th>
<th>Overlearning rate</th>
<th>No. of Hidden Neurons</th>
<th>Overlearning rate</th>
<th>No. of Hidden Neurons</th>
<th>Overlearning rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.0189</td>
<td>10</td>
<td>0.4777</td>
<td>18</td>
<td>0.1983</td>
</tr>
<tr>
<td>3</td>
<td>0.2077</td>
<td>11</td>
<td>0.3545</td>
<td>19</td>
<td>0.2538</td>
</tr>
<tr>
<td>4</td>
<td>0.2103</td>
<td>12</td>
<td>0.3215</td>
<td>20</td>
<td>0.2088</td>
</tr>
<tr>
<td>5</td>
<td>0.2538</td>
<td>13</td>
<td>0.2758</td>
<td>21</td>
<td>0.5913</td>
</tr>
<tr>
<td>6</td>
<td>0.2088</td>
<td>14</td>
<td>0.3844</td>
<td>22</td>
<td>0.6243</td>
</tr>
<tr>
<td>7</td>
<td>0.1983</td>
<td>15</td>
<td>0.3545</td>
<td>23</td>
<td>0.3215</td>
</tr>
<tr>
<td>8</td>
<td>0.0539</td>
<td>16</td>
<td>0.2747</td>
<td>24</td>
<td>0.3215</td>
</tr>
<tr>
<td>9</td>
<td>0.1983</td>
<td>17</td>
<td>0.3545</td>
<td>25</td>
<td>0.3215</td>
</tr>
<tr>
<td>10</td>
<td>0.3844</td>
<td>18</td>
<td>0.2747</td>
<td>26</td>
<td>0.3545</td>
</tr>
<tr>
<td>11</td>
<td>0.3545</td>
<td>19</td>
<td>0.2747</td>
<td>27</td>
<td>0.3545</td>
</tr>
<tr>
<td>12</td>
<td>0.2758</td>
<td>20</td>
<td>0.2747</td>
<td>28</td>
<td>0.2747</td>
</tr>
<tr>
<td>13</td>
<td>0.2747</td>
<td>21</td>
<td>0.2747</td>
<td>29</td>
<td>0.2747</td>
</tr>
<tr>
<td>14</td>
<td>0.2747</td>
<td>22</td>
<td>0.2747</td>
<td>30</td>
<td>0.2747</td>
</tr>
<tr>
<td>15</td>
<td>0.2747</td>
<td>23</td>
<td>0.2747</td>
<td>31</td>
<td>0.2747</td>
</tr>
<tr>
<td>16</td>
<td>0.0978</td>
<td>24</td>
<td>0.2747</td>
<td>32</td>
<td>0.2747</td>
</tr>
<tr>
<td>17</td>
<td>0.1399</td>
<td>25</td>
<td>0.2747</td>
<td>33</td>
<td>0.2747</td>
</tr>
<tr>
<td>18</td>
<td>0.1494</td>
<td>26</td>
<td>0.2747</td>
<td>34</td>
<td>0.2747</td>
</tr>
<tr>
<td>19</td>
<td>0.5018</td>
<td>27</td>
<td>0.2747</td>
<td>35</td>
<td>0.2747</td>
</tr>
<tr>
<td>20</td>
<td>0.1494</td>
<td>28</td>
<td>0.2747</td>
<td>36</td>
<td>0.2747</td>
</tr>
<tr>
<td>21</td>
<td>0.1494</td>
<td>29</td>
<td>0.2747</td>
<td>37</td>
<td>0.2747</td>
</tr>
<tr>
<td>22</td>
<td>0.1494</td>
<td>30</td>
<td>0.2747</td>
<td>38</td>
<td>0.2747</td>
</tr>
<tr>
<td>23</td>
<td>0.1494</td>
<td>31</td>
<td>0.2747</td>
<td>39</td>
<td>0.2747</td>
</tr>
<tr>
<td>24</td>
<td>0.1494</td>
<td>32</td>
<td>0.2747</td>
<td>40</td>
<td>0.2747</td>
</tr>
<tr>
<td>25</td>
<td>0.1494</td>
<td>33</td>
<td>0.2747</td>
<td>41</td>
<td>0.2747</td>
</tr>
<tr>
<td>26</td>
<td>0.1494</td>
<td>34</td>
<td>0.2747</td>
<td>42</td>
<td>0.2747</td>
</tr>
<tr>
<td>27</td>
<td>0.1494</td>
<td>35</td>
<td>0.2747</td>
<td>43</td>
<td>0.2747</td>
</tr>
<tr>
<td>28</td>
<td>0.1494</td>
<td>36</td>
<td>0.2747</td>
<td>44</td>
<td>0.2747</td>
</tr>
<tr>
<td>29</td>
<td>0.1494</td>
<td>37</td>
<td>0.2747</td>
<td>45</td>
<td>0.2747</td>
</tr>
<tr>
<td>30</td>
<td>0.1494</td>
<td>38</td>
<td>0.2747</td>
<td>46</td>
<td>0.2747</td>
</tr>
<tr>
<td>31</td>
<td>0.1494</td>
<td>39</td>
<td>0.2747</td>
<td>47</td>
<td>0.2747</td>
</tr>
<tr>
<td>32</td>
<td>0.1494</td>
<td>40</td>
<td>0.2747</td>
<td>48</td>
<td>0.2747</td>
</tr>
<tr>
<td>33</td>
<td>0.1494</td>
<td>41</td>
<td>0.2747</td>
<td>49</td>
<td>0.2747</td>
</tr>
<tr>
<td>34</td>
<td>0.1494</td>
<td>42</td>
<td>0.2747</td>
<td>50</td>
<td>0.2747</td>
</tr>
<tr>
<td>35</td>
<td>0.1494</td>
<td>43</td>
<td>0.2747</td>
<td>51</td>
<td>0.2747</td>
</tr>
<tr>
<td>36</td>
<td>0.1494</td>
<td>44</td>
<td>0.2747</td>
<td>52</td>
<td>0.2747</td>
</tr>
<tr>
<td>37</td>
<td>0.1494</td>
<td>45</td>
<td>0.2747</td>
<td>53</td>
<td>0.2747</td>
</tr>
<tr>
<td>38</td>
<td>0.1494</td>
<td>46</td>
<td>0.2747</td>
<td>54</td>
<td>0.2747</td>
</tr>
<tr>
<td>39</td>
<td>0.1494</td>
<td>47</td>
<td>0.2747</td>
<td>55</td>
<td>0.2747</td>
</tr>
<tr>
<td>40</td>
<td>0.1494</td>
<td>48</td>
<td>0.2747</td>
<td>56</td>
<td>0.2747</td>
</tr>
<tr>
<td>41</td>
<td>0.1494</td>
<td>49</td>
<td>0.2747</td>
<td>57</td>
<td>0.2747</td>
</tr>
<tr>
<td>42</td>
<td>0.1494</td>
<td>50</td>
<td>0.2747</td>
<td>58</td>
<td>0.2747</td>
</tr>
<tr>
<td>43</td>
<td>0.1494</td>
<td>51</td>
<td>0.2747</td>
<td>59</td>
<td>0.2747</td>
</tr>
<tr>
<td>44</td>
<td>0.1494</td>
<td>52</td>
<td>0.2747</td>
<td>60</td>
<td>0.2747</td>
</tr>
<tr>
<td>45</td>
<td>0.1494</td>
<td>53</td>
<td>0.2747</td>
<td>61</td>
<td>0.2747</td>
</tr>
<tr>
<td>46</td>
<td>0.1494</td>
<td>54</td>
<td>0.2747</td>
<td>62</td>
<td>0.2747</td>
</tr>
<tr>
<td>47</td>
<td>0.1494</td>
<td>55</td>
<td>0.2747</td>
<td>63</td>
<td>0.2747</td>
</tr>
<tr>
<td>48</td>
<td>0.1494</td>
<td>56</td>
<td>0.2747</td>
<td>64</td>
<td>0.2747</td>
</tr>
<tr>
<td>49</td>
<td>0.1494</td>
<td>57</td>
<td>0.2747</td>
<td>65</td>
<td>0.2747</td>
</tr>
<tr>
<td>50</td>
<td>0.1494</td>
<td>58</td>
<td>0.2747</td>
<td>66</td>
<td>0.2747</td>
</tr>
<tr>
<td>51</td>
<td>0.1494</td>
<td>59</td>
<td>0.2747</td>
<td>67</td>
<td>0.2747</td>
</tr>
<tr>
<td>52</td>
<td>0.1494</td>
<td>60</td>
<td>0.2747</td>
<td>68</td>
<td>0.2747</td>
</tr>
<tr>
<td>53</td>
<td>0.1494</td>
<td>61</td>
<td>0.2747</td>
<td>69</td>
<td>0.2747</td>
</tr>
<tr>
<td>54</td>
<td>0.1494</td>
<td>62</td>
<td>0.2747</td>
<td>70</td>
<td>0.2747</td>
</tr>
<tr>
<td>55</td>
<td>0.1494</td>
<td>63</td>
<td>0.2747</td>
<td>71</td>
<td>0.2747</td>
</tr>
<tr>
<td>56</td>
<td>0.1494</td>
<td>64</td>
<td>0.2747</td>
<td>72</td>
<td>0.2747</td>
</tr>
<tr>
<td>57</td>
<td>0.1494</td>
<td>65</td>
<td>0.2747</td>
<td>73</td>
<td>0.2747</td>
</tr>
<tr>
<td>58</td>
<td>0.1494</td>
<td>66</td>
<td>0.2747</td>
<td>74</td>
<td>0.2747</td>
</tr>
<tr>
<td>59</td>
<td>0.1494</td>
<td>67</td>
<td>0.2747</td>
<td>75</td>
<td>0.2747</td>
</tr>
<tr>
<td>60</td>
<td>0.1494</td>
<td>68</td>
<td>0.2747</td>
<td>76</td>
<td>0.2747</td>
</tr>
</tbody>
</table>

In essence, the higher the overlearning rate, the worse the model will be. Therefore, it is reasonable to name the largest overlearning rate in a given model as its worst overlearning rate.

\[
\text{The worst Overlearning Rate} = \max (R^2_{\text{training}} - R^2_{\text{Test}}) \quad (7.5)
\]

According to Table 7-15, the worst overlearning rate in case of the last best primary model is equal to 0.993. This overlearning rate is attributed to the network using twenty-three neurons in its hidden layers. Table 7-16 summarizes the worst overlearning rates obtained for the whole set of the best primary models, together with the numbers of hidden neurons associated with these rates. However, this information was not reported for the first model because, as described earlier, this model consistently predicted the test samples better than the training samples, irrespective of the number of hidden neurons (see Figure 7-12).
Table 7-16: Summary of the worst overlearning rates and their corresponding numbers of hidden neurons

<table>
<thead>
<tr>
<th>No. of ANN model</th>
<th>No. of input variables</th>
<th>Architecture of the model</th>
<th>Learning parameters</th>
<th>The worst overlearning rate</th>
<th>The associated no. of hidden neurons</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>LR</td>
<td>M</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>Two hidden layers</td>
<td>0.4</td>
<td>0.9</td>
<td>0.256</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>Two hidden layers</td>
<td>0.9</td>
<td>0.2</td>
<td>0.283</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>Three hidden layers</td>
<td>0.8</td>
<td>0.1</td>
<td>0.347</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>One hidden layer</td>
<td>0.4</td>
<td>0.9</td>
<td>0.409</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>Two hidden layers</td>
<td>0.8</td>
<td>0.9</td>
<td>0.713</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>Two hidden layers</td>
<td>0.5</td>
<td>0.9</td>
<td>0.993</td>
</tr>
</tbody>
</table>

As can be inferred from Table 7-16, the overlearning problem occurred when an excessive amount of neurons were introduced to the hidden layers. The numbers of hidden neurons in association with the worst overlearning rates in Table 7-16 were considerably greater than the numbers which were stated in Table 7-4 and initially utilized in the preliminary development phase. This holds true for all the models presented in Table 7-16, but model number four, which was described earlier to have a different architecture than other models. As a general conclusion, it can be stated that the methodology used in this research for the establishment of the initial number of hidden neurons would plausibly help an ANN model not to get trapped with the overlearning problem.

The variation of the worst overlearning rates with respect to the number of input variables (i.e., independent variables) is graphically shown in Figure 7-19. It should be noted that the number assigned to a given model in this figure is exactly the same as the number of input variables considered in that particular model. According to Figure 7-19, the severity of the overlearning problem sharply increases as the number of input variables increases. Therefore, it can be concluded that the overlearning problem is of more concern when dealing with more number of input variables.
Figure 7-19: Comparison of the worst overlearning rates with respect to the number of input variables
7.14 Effect of the Number of Input Variables on the Performance of the Networks

The indicative attributes of the final best models are illustrated in Figure 7-20. These attributes include the number of hidden layers, the number of hidden neurons, and the values of the learning rate and the momentum. To study how the performance of these models were influenced by the number of the input variables, their respective $R^2$ values over the training and test datasets are depicted in Figure 7-20.

As can be observed in Figure 7-20, the value of $R^2$ in the training dataset increases continuously as the number of input variables increases. This observation is just similar to what observed in Chapter 5 when developing generic MLR models. This similarity indicates that, like MLR models, the accuracy ability of the ANN models improved by taking more independent variables into account. As stated earlier, these variables were all found to be statistically significant predictors of the RNCC. Additionally, Figure 7-20 signifies that the rate of increase of $R^2$ values in the training dataset does not follow a regular pattern. For instance, the $R^2$ value obtained from the application of the first model to the training dataset was increased by nearly 15% when the second independent variable is added (i.e., $R^2$ values of the first and second models were 0.5868 and 0.6764, respectively). However, the inclusion of the last independent variable slightly increased the $R^2$ value of the sixth model by just 1.4% (i.e., $R^2$ values of the sixth and last model are 0.8198 and 0.8313, respectively). Therefore, as in the MLR models developed in Chapter 5, the input variables in the final best ANN models did reveal different levels of explanatory power, particularly when these variables were combined with their former variables.

Figure 7-20 shows that $R^2$ values of the test dataset bounce around over the models without a particular order. According to Figure 7-20, these values are constantly decreasing from the first model to the third model and reciprocally increasing from the fifth model to the last model. Model number four is between these two trends and represents another peak in the curve. In fact, this model impairs the $R^2$ values to follow an explicit pattern which could otherwise be possibly expressed by a parabolic curve. Similar to the training dataset, the test dataset is best predicted by the last model, having the largest $R^2$ value. However, the model which includes the least number of input variables (i.e., model number one) returns the second largest $R^2$ value for the test dataset. This model contains only one independent variable that is $X_1$ (i.e., Ln (area)). As a concluding remark, the results of the application of the ANN models to the test dataset suggest that “area” is a key, and probably the most
important, variable in the prediction of the RNCC. This conclusion is in complete agreement with the conclusion drawn earlier in Chapter 5.
Figure 7-20: Characteristics of the final best ANN models along with their respective $R^2$ values for “training” and “test” datasets.
7.15 Conclusion

The data from 158 framed structures were utilized to develop intelligent models for the prediction of the RNCC, using the ANN methodology. Seven ANN models were accordingly developed for seven different sets of independent variables whose effect on the determination of the RNCC was found to be statistically significant in Chapter 5. For each model, a series of back-propagation neural networks with various architectures and training parameters were tried. Seventy-five percent of the data were randomly selected for use in the training process. The remainder of data were reserved to build the test dataset against which the performance of the trained networks was examined. Through the training process, the calibration technique was desirably tailored to lessen the chance of the overlearning problem to occur. The generalization ability, rather than the accuracy ability, was taken as the criterion for the evaluation of the performance of different networks developed for each of the seven ANN models. On the basis of this criterion, the models which yielded the largest $R^2$ values on the test dataset were selected as the final best ANN models.

The final ANN models were created by following a two-phase development procedure. The first phase, called the preliminary development phase, was an iterative process through which different architectures in terms of the number of hidden layers, together with various combinations of learning parameters in terms of the learning rate and the momentum were examined for each of the ANN models. Upon the completion of this phase, the best performing architecture and the most appropriate values of the learning parameters were determined for each model. After the determination of the optimal architectures and learning parameters, they were held constant and additional experiments were carried out in the second phase, which was called the final development phase. In this phase, a wide range of hidden neurons was examined for each of the models resulted from the previous phase and, as a result, the most suitable number was selected for the hidden neurons of each model.

Important issues which have a pivotal role in the successful utilization of the ANN methodology were addressed in this chapter and their effect on the performance of the ANN models developed was investigated. The significance of this investigation is especially more pronounced for issues that have not yet been adequately studied in the literature. For instance, it was found that the performance of a given model is more affected by the number of hidden neurons rather than the values of the learning parameters. It was also found that the
generalization ability of a given ANN model is more sensitive than its accuracy ability to variations of both the number of hidden neurons and the learning parameters. Several experiments were also carried out to compare the performance of different architectures. Three types of MLP networks with different numbers of hidden layers were applied and accordingly the optimal network was selected for each of the ANN models studied. Based on the results obtained, it was suggested that MLP networks with more than one hidden layer are of great potential use in the ANN methodology. In this regard, an emphasis was given to the use of MLP networks with two hidden layers.

In this chapter, a methodology was used to initialize the number of hidden neurons. The evaluation of the results, obtained at the end of the final development phase, showed that this methodology was robust for the identification of the most appropriate number of hidden neurons. The application of this methodology was further justified by the fact that it effectively prevented the occurrence of the overlearning problem. The severity of this problem was also assessed in accordance with the variation in the number of input (i.e., independent) variables. The results of this assessment revealed that the overlearning problem becomes more severe when the number of input variables increases. It was therefore suggested that the overlearning problem should be carefully controlled, particularly when dealing with a greater number of input variables.

The results presented in this chapter are in good agreement with the findings presented in Chapter 5. Amongst all the final ANN models developed in the current chapter, model number seven with the inclusion of all the independent variables was found to exhibit the maximal performance with respect to the accuracy ability. Support for this finding was obtained by comparing the R² values of the final models, when being applied to the training dataset. Taking these values into account, it was observed that the accuracy ability of the models increased as the number of input variables increased. This observation implies that all the seven input variables, which were considered in the development of the ANN models, are meaningful predictors of the RNCC. However, as described in this chapter, the explanatory power of these variables was found to be dissimilar. This dissimilarity was also previously noted in Chapter 5, when the MLR models were applied to the development (i.e., training) dataset.

The final issue studied in this chapter was the comparison of the generalization abilities of the final ANN models developed. Similar to the training samples, the samples in
the test dataset were also most accurately predicted by the last ANN model. As shown, this model resulted in the largest $R^2$ value of 0.73 for the test dataset. In addition, based on the results obtained, it was found that the first ANN model with the inclusion of only one independent variable (i.e., $\text{Ln (area)}$) yielded the second largest $R^2$ value. According to this finding, it was concluded that “area” is a key, and probably the most important, variable in the prediction of the RNCC. This conclusion further supports the use of the most parsimonious regression model, which was developed in Chapter 5 in the form of $C=KA^B$.

In the next chapter, the performance of the ANN methodology is compared to that of offered by the conventional MLR modelling technique.
CHAPTER 8

COMPARISON OF MLR AND ANN METHODOLOGIES TO PREDICT RETROFIT NET CONSTRUCTION COST (RNCC) IN FRAMED STRUCTURES

8.1 Introduction

In previous chapters, based on a total of 158 earthquake-prone framed structures, seven MLR models and seven ANN models using the same combinations of statistically significant independent variables were developed and used for the prediction of the RNCC. 121 of these structures were randomly selected to develop these models while the remaining 37 structures were mainly used to validate the generalization ability of the models. Each of the models utilized the same set of structures for development and validation.

A final objective of this study derived from recognition that research has not yet been published which specifically compares the performance of the non-parametric modelling approach (i.e., ANN) against a more traditional parametric modelling approach (i.e., MLR) to predict RNCC. The results of this study will bridge this gap in the literature. This chapter starts with a concise discussion on important issues that may affect the credibility of each of these approaches and continues with a detailed review of previous comparative studies in the literature. The chapter concludes with evaluating the forecasting strengths of the ANN models, selected as superior in Chapter 7, against those derived from their respective MLR models in Chapter 5. Accuracy and generalization abilities are again considered as the principles for comparing the performance of these two modelling approaches. The measure utilized for this comparison is determined from the mean absolute percentage error (MAPE) with the lesser value being superior.

8.2 MLR Modelling Technique vs. the ANN Methodology

ANN models can justifiably be defined as a true tool because, unlike regression models, they allow all relevant variables to interact freely with one another (Ostberg, 2005) and implicitly detect all possible interaction terms (Doig, 1999). Assuming no forced constraints, ANN models are therefore adept at more accurately assessing the uncertainty in the estimation of the problem being modelled. However, although they are free of
mathematically explicit models, ANN models are basically statistical and stochastic in origin (Liu, 1998). Rather than defining neural networks as a radically different alternative to traditional statistical approaches, experts have noted that neural models are in reality statistics-based models as demonstrated by the mathematics-based activation functions which derive the neural processing, and the weight adjustment process which attempts to achieve a better fit to the data using specific algorithms (Somers, 1999). ANN’s ability to relate input and output variables can be used to tackle problems that have been conventionally handled by statistical methods such as regression analysis.

Although regression models have demonstrated acceptable predictive performance, some researchers believe that further attempts at refining existing systems will be worthless and that understanding the complex patterns associated with the problem under investigation may require the use of other mathematical approaches such as the ANN methodology, that has generated a lot of interest in many fields of science (Doig, 1999) including cost estimation. As a relatively new information-processing methodology, the ANN methodology serves as a non-parametric statistical estimator (Bode, 2000; De la Garza and Rouhana, 1995; Smith and Mason, 1997; White, 1990) which holds great promise for the development of reliable, valid, and accurate cost estimates (De la Garza and Rouhana, 1995). The ANN methodology is potentially able to overcome some of the major deficiencies associated with the MLR technique and therefore ANN models may provide better results than their corresponding MLR models. The general advantages of the ANN modelling technique are thoroughly outlined in Chapter 6 for various types of problems including cost estimation. However, a number of these advantages which specifically address the superiority of this technique over the MLR technique are highlighted below:

a. The most important advantage of ANN models over MLR models is their adaptivity (Boussabaine, 1996; Haykin, 1999). ANN models can be considered dynamic prediction models (Jeong, 2004) because they consistently improve during the training. The samples in the training dataset are gradually presented to ANN models and then the models automatically adjust their weights according to the learning algorithm to give the most appropriate output. By contrast, these samples are all simultaneously processed in MLR models (Liu, 1998). According to Rafiq et al., (2001), it is the learning ability of ANN models which gives them an advantage in solving complex problems whose analytic or numerical solutions are hard to be obtained by MLR models.
b. A major disadvantage of parametric methods is that the methods are based on the assumption of a defined mathematical functional form that best fits the available historical data (Creese and Li, 1995). Likewise, regression-based methods require the specification of a function type before the parameters of this function can be estimated (Bode, 2000). MLR models can fit the data only in certain pre-specified types of functions (e.g., linear, polynomial, double-log or exponential functions) which may not always be appropriate and consequently may limit the ability of the models to fit the data on which they are estimated. A sufficiently accurate cost function must therefore be known to the cost researcher a priori, which is obviously rarely the case, posing a major difficulty for implementation of the MLR technique. ANN models, in contrast, are non-parametric estimators in which no assumption about the shape of the approximation function has to be made prior to training (Adeli and Wu, 1998; Alex et al., 2010; Bode, 2000; De la Garza and Rouhana, 1995; Liu, 1998). ANN models attempt to fit curves through data without being provided a predetermined function. Having greater freedom to fit the data than regression models (Wilmot and Mei, 2005), ANN models may more effectively solve problems in which no explicit mathematical function in a multi-dimensional space of variables can be detected (Adeli and Wu, 1998). This detection represents a major advantage of the ANN methodology over the regression modelling technique, as the use of the former reduces the level of expertise required by users (Wang et al., 2000).

c. Another disadvantage of applying traditional regression analysis is its unsuitability to account for a large number of variables being present in a problem and the numerous interactions among them (Hegazy and Ayed, 1998). ANN models, by contrast, are not restricted by the number of variables. They are in fact more desirable for those cases in which more variables are taken into account (Attalla and Hegazy, 2003; Jeon, 2007; Liu, 1998), particularly if the size of the training dataset is sufficiently large. In such cases, ANN models become more diverse and are able to detect any patterns found in the data and provide larger opportunity to investigate different variables affecting the problem at hand (Attalla and Hegazy, 2003).

d. One of the most important advantages of ANN models over MLR models is their ability to capture the nonlinearity and discontinuity that inevitably will exist between variables (Alex et al., 2010; Doig, 1999; Emsley et al., 2002; Jeong, 2004; Smith and
Mason, 1997). ANNs are able to capture highly complex nonlinear relationships and \(n^{th}\)-order interactions. Although the nonlinear regression technique can also be used to account for nonlinearity, in this technique specifying the exact nature of nonlinearity between variables is a burdensome task which usually requires having a detailed knowledge about the appropriate nonlinear relationships between the independent and dependent variables, which is obviously rarely the case. When using ANN models these relationships do not need to be pre-specified as they are determined implicitly by ANN models (Emsley et al., 2002).

e. The ANN methodology is not constrained by the distributional and other assumptions necessary with the regression technique (Garson, 1998; McFadden et al., 2001; Swicegood, 1998) as the former technique belongs to the class of non-parametric statistical modelling. This feature gives ANN models the opportunity to be particularly used for investigating problems having the possibility of severe multicollinearity in the input variables as such models can readily accommodate and modify multicollinearity (Doig, 1999; Shehab et al., 2010; Smith and Mason, 1997). This modification is another major advantage of the ANN methodology, which in turn enables researchers to simply examine all independent variables which presumably have an effect on the problem regardless of whether or not these variables are correlated.

f. Another advantage of the ANN methodology over regression analysis is when missing or incomplete data are available. ANN models have the ability to handle this type of data more effectively than regression models do (Harding et al., 1999; Liu, 1998). This advantage is more pronounced when taking this fact into account that complete datasets will usually be unavailable (Alex et al., 2010; Harding et al., 1999; Smith and Mason, 1997).

g. ANN models can typically produce more accurate predictions with well-established and good quality dataset than can regression models (Dutta and Shekhar, 1988). In addition, ANN models are usually more beneficial where significant data exist (Liu, 1998). Some studies have revealed that to achieve a specific level of prediction accuracy, lower amounts of data together with fewer independent variables are
required to develop an ANN model than to develop a regression model (Jeon, 2007; Shehab et al., 2010; Stockton and Wang, 2004).

Although the ANN methodology has some advantages over conventional techniques, as stated above, it has relative weaknesses that should be acknowledged. These weaknesses account for the reasons why conventional techniques for prediction are well practiced and abound in comparison with the ANN methodology.

As discussed in Chapter 6, one of the largest impediments to widespread use of the ANN methodology is its black box nature, in addition to its inability to duplicate results (Smith and Mason, 1997). Unlike the MLR technique, causal analysis using the ANN methodology is extremely problematic and ANN models are practically unable to provide an obvious explanation of how problems are solved. The results are often uninterpretable and offer little insight to the analyst. In addition, the overall effect of individual input variables on the dependent variable cannot be estimated when using ANN models. These models are also limited in their ability to test the significance of individual input variables (Swicegood, 1998). Moreover, MLR models could normally be checked for validity using the practical experience of users. An ANN model however does not have this advantage as it has no relationship with normal parametric models. In order to maintain confidence in an ANN model, it is also necessary to ensure that the model use is frequently audited and at regular intervals the model is updating using actual new data. Another disadvantage of the ANN methodology arises due to the fact that learning curve in building ANN models is far more imposing than that of conventional models (Smith and Mason, 1997). Many different types of neural networks exist, whose performance is also typically controlled by a variable number of parameters. Deciding over the network type and setting the network parameters to their most suitable values is neither simple nor straightforward as ANN theory does not yet provide applicable rules for such issues (Bode, 2000). There is also no explicit set of rules to determine whether a given learning algorithm is suitable for a given application. A learning algorithm may work particularly well for a certain number of applications but not for others (Hegazy and Ayed, 1998). Therefore, an extensive amount of trial and error seems inevitable to arrive at the optimal composition of an ANN model and therefore a career could be devoted to this endeavour.
Drawbacks such as those stated above make it difficult to persuade individuals who do not have a broad understanding of the ANN methodology that this methodology is competent to produce reliable estimates. This methodology is often less clear to members of the construction industry and even some academics in the construction management are suspicious of artificial intelligence systems (Boussabaine, 1996). In particular, the common perception of ANN models as “mystical black boxes” (Somers, 1999) may increase concerns regarding the appropriateness and legal defensibility of these innovative models.

According to the study conducted by Smith and Mason (1997), it is worth emphasizing two particular points. First, the regression modelling technique is expected to outperform the ANN methodology when an analyst knows or can closely guess the underlying analytic relationship between dependent variable and its predictors. Second, while it is expected that the ANN technique will be used with increasing frequency as a viable alternative for the MLR technique, the concerns associated with the use of the former modelling technique may represent formidable impediments to widespread use and acceptance of ANN-based cost estimation relationships.

8.3 Comparative Studies in the Literature

Comparison of ANN models with more traditional statistical techniques has been the focus of many recent studies. Various ANN models have been proposed over the past decades and impressive results have been obtained when compared to those obtained using traditional models in general, and regression models in particular. The following articles are part of a growing body of literature that supports the claim above.

1- In one of the first structured comparisons of ANN and multivariate regression models, Dutta and Shekhar (1988) trained a series of back propagation neural networks with various architectures to predict bond ratings. Bond rates from 47 firms were selected at random. Thirty of these bonds were used to train the networks and the remaining ones were used to test their performance. Two MLR models were simultaneously developed using the same independent and dependent variables. The ANN models consistently outperformed their respective MLR models during both learning and test phases. On the training dataset, the MLR models predicted bond ratings correctly
from 63% to 67% of the time, while the ANN models predicted bond ratings correctly from 80% to 92% of the time. When the models were applied on the test dataset, the accuracy of the MLR models held at about 65%, while the accuracy of the ANN models fell to 77% to 82%. Additionally, the total squared error for the MLR models was, on average, a full order of magnitude greater than that of the ANN models.

2- A review by Sharda and Patil (1992) of forty-two studies, comparing ANN models to more common statistical techniques in operations research, suggested that ANN models tend to outperform conventional techniques. Comparisons included discriminant analysis, Box-Jenkins methodology, logistic regression, linear binary, and multiple regression techniques. In 30 (71%) of the 42 studies, ANN models outperformed the conventional techniques, matched their performance in 5 (12%) cases, and underperformed in 7 (17%) cases.

3- Shtub and Zimmerman (1993) demonstrated the potential value of using ANN models in cost estimation as opposed to traditional regression or engineering analysis. Network architecture was proposed by which the expected cost of six major types of assembly systems was estimated. The performances of the network were then compared to those of a regression model commonly used for cost estimation. The network consistently outperformed the regression model with respect to several performance measures.

4- Yoon et al., (1993) examined the capabilities and limitations of multivariate discriminant analysis (MDA) and the ANN methodology in classifying data. They compared a MDA model with back-propagation based neural network models to predict stock price performance. The dataset consisted of 151 firms. Half of the data were used in the training phase, while another half was separated out to validate the performance of the techniques. In their study, a four-layer network, containing four input neurons, two output neurons, and seven hidden neurons, was the best ANN model. This network outperformed the MDA model by 23% in the training dataset and 13% in the validation or test dataset.

5- Fletcher and Goss (1993) reported that they applied the logistic regression analysis and the back-propagation based neural networks to bankruptcy data. Data from 18 bankrupt companies were pair-matched to 18 non-bankrupt companies. Training efficiency, prediction accuracy, variance in errors, and prediction risk were the criteria
used to compare the performance of the two techniques. The results showed that back-
propagation neural networks outperformed logistic regression analysis. The research
concluded with demonstrating that ANN models are a possible alternative to more
traditional methods of estimating causal relationships in data.

6- McKim (1993a) developed an ANN model to predict project cost overruns. The
predictive error of the model was less than half when compared to the arithmetic
mean, which is a very elementary conventional method. McKim’s study showed that
the neural network approach outperformed the conventional, yet simple, method.

7- Openshaw (1993) examined the empirical performance of feed-forward neural
networks as the basis for representing the spatial interaction reflected in the journey to
work data. The performance of ANN models were compared with various types of
conventional models. Taking the goodness of fit index (i.e., $R^2$) into account, the
ANN models produced a level of performance better than the conventional models.

8- Faghri and Hua (1993) explored the application of neural networks for trip generation
analysis. They built two ANN models, trained them using the back-propagation
learning algorithm, and applied them to a real world database of 150 different sites. A
traditional MLR model was also developed using the same database. The prediction
performance of these three models was then compared. The results showed that both
ANN models performed better than the MLR model.

9- In his doctoral dissertation, Scarborough (1995) compared linear and nonlinear
multivariate regressions, and MLP networks, using the back-propagation learning
algorithm to predict normalized annual sales revenue among telephone sales agents.
He found that the MLP networks were significantly more accurate than linear
regression. However, no significant differences were found between the MLP
networks and nonlinear regression model at the 5% level of significance.

10- De La Garza and Rouhana (1995) used regression analysis as an example of
establishing a cost estimating relationship, and both linear and nonlinear regression
models were presented. The results of the regression models were then compared with
those obtained using an ANN methodology. For the numerical application of the ANN
methodology, a three-layered back-propagation network having a three-neuron input
layer, four-neuron hidden layer, and a single-neuron output layer was used. The
results obtained showed that the ANN methodology outperformed the parametric estimating techniques based on the linear and nonlinear regression analysis, implying strong predictive capabilities of the ANN models.

11- In the travel demand forecasting procedure, Faghri and Aneja (1996) proposed an ANN-based methodology using a database collected from 60 sites. Based on the same data, 26 regression models and 18 ANN models were developed. The results of these two types of models were compared, which showed that the ANN models generated more accurate results than did regression models.

12- Liu (1998) studied the application of the ANN methodology in capital project risk analysis and developed intelligent computer models to predict cost and time variations at the front end stages of projects. The models were used to evaluate the potential effects of risks and project decisions on outcomes. Results indicated that ANN models have the capability to capture general patterns by learning from samples of similar past projects. Furthermore, ANN models emerged to be superior to MLR models in the prediction of project cost and time variations. However, Liu (1998) argued that ANN models in combination with stepwise regression models provide more accurate estimations than stand-alone ANN models.

13- In contrast to most previous studies that used MLR regression models for developing physiological models, Abou-Elseoud (1998) used a three-layer feed-forward ANN paradigm with the back-propagation learning algorithm to study the suitability of utilizing the ANN methodology to establish such models. Given an appropriate network architecture and training algorithm, the ANN models developed have demonstrated highly accurate prediction and generalization abilities. Using the same training and test datasets, the results showed that the ANN models, by taking full advantage of inherent nonlinearities, outperformed the existing MLR models, particularly in their predictive power for new data.

14- To describe the relationship between the percentage of pass-by trips and average daily traffic, Faghri et al., (1999) developed three regression models and also five ANN models which were all trained by the back-propagation learning algorithm. Comparing the predictive power of the models, they found that the ANN models captured the relationship more satisfactorily than did the regression models.
15- ANN models trained by the back-propagation learning algorithm were developed by Bullock (1999) as a tool to predict corporate failure and non-failure. 54 ANN models with different configurations were tested for forecasting bankruptcy. The predictive accuracy of the ANN model selected as superior was then examined and compared to a conventional model developed by implementing a discriminant analysis methodology. The results showed that the ANN methodology is more desirable in general. The forecasting performance of the ANN model was almost 30% more accurate than that of the discriminant analysis formula.

16- The predictive performance of a series of logistic regression models to a corresponding series of back-propagation ANN models was also compared by Doig (1999) while studying the severity of illness scoring in the intensive care unit (ICU) using a database of 1,181 patients. Data were collected on day one and day three of stay in the ICU. The results of this study revealed that there was no significant difference between logistic regression models and ANN models developed using day one data. However, the ANN model developed using day three data performed significantly better than the corresponding logistic regression model. The time dependent ANN model also performed significantly better than the corresponding regression model. On the basis of the acquired results, Doig (1999) concluded that ANN models demonstrated significantly better predictive performance.

17- Shtub and Versano (1999) developed a cost estimating system that can be linked to a computer aided design (CAD) system producing cost estimates based on the CAD data. The parameters used as input to the cost estimation model were therefore objective and easily available. The proposed system was used to predict the cost of steel pipe bending and was based on an ANN analysis. Four types of pipes were studied. A MLR model was also developed for each of the four groups. The accuracy of the ANN models was tested and compared to the accuracy of the traditional approach of regression analysis. A comparative study revealed that the proposed ANN models outperformed the traditional MLR models used for cost estimation.

18- Bode (2000) compared the performance of ANN-based cost models with that of conventional methods, i.e., linear and nonlinear parametric regression. The results indicated the superiority of ANN models when compared with corresponding regression models, as ANN models achieve lower deviations in their cost estimations.
19- On the basis of data from nearly 300 building projects, Emsley et al., (2002) examined the application of both ANN and regression approaches to the prediction of total construction costs. Models based on the linear regression technique were used as a benchmark for evaluation of the ANN models. The results of this study showed that where linear regression and ANN models were developed using the same variables, ANN models always outperformed their regression counterparts.

20- Jeong (2004) utilized regression analysis, the ANN methodology, and a historical-based method to predict bus arrival time. Different prediction models were developed and tested. It was found that the ANN models gave the smallest prediction errors in terms of prediction accuracy and, therefore, performed considerably better than either historical-based models or MLR models. It was then hypothesized that the ANN technique was able to identify the complex nonlinear relationship between travel time and the independent variables and this led to superior results.

21- Kim et al., (2004) examined the performance of three approaches commonly used for cost estimating. The examinations were based on MLR analysis, the ANN methodology, and the case-base reasoning (CBR) method applied for estimating construction costs of Korean residential buildings. These three approaches used a dataset containing 530 historical costs. The results showed that the best ANN model gave more accurate estimation results than did the MLR or CBR models. However, the results indicated that the CBR model was more effective than other models with respect to (1) the time and accuracy tradeoffs, (2) its clarity of explanation in estimating construction costs, and (3) its ease of updating and consistency in the variables stored.

22- Using the mean absolute error (MAE) value as the measure of accuracy, Stockton and Wang (2004) compared the ability of ANN and MLR techniques in the development of cost models. Three sets of training dataset were used to develop both ANN models and MLR models. The number of data points in the training datasets was 150, 450, and 750. In addition, one, two, four, six, nine, and sixteen independent variables were also considered for each of these training datasets. Using the same combination of training size and independent variables, the results indicated the superiority of the best ANN models developed over their counterpart MLR models in all arrangements studied.
Two methodological studies were undertaken by Ostberg (2005) to empirically demonstrate the value of applying the ANN methodology to predict employee job performance criteria on the basis of the data gathered from 4,299 hourly employees. The MLP networks, in addition to MLR models, were developed. The results suggested that the MLP networks significantly outperformed the MLR models in predicting voluntary or involuntary termination status and eligibility for rehire. However, the performance of MLP networks on predicting the employee tenure criterion was just equivalent to that of the MLR models. This study, however, showed that the different modelling techniques may vary in usefulness for different prediction contexts.

Wilmot and Mei (2005) developed ANN-based and MLR-based models to estimate the escalation of highway construction costs over time. Data used for establishing these models and also testing the predictive performance of the developed models was the same and consisted of information on a total of 2,827 highway and bridge contracts for the period from 1981 to 1997. Results demonstrate that the ANN model is more accurate than the model generated using regression technique.

Liu (2006) employed the ANN methodology to investigate the potential effects of influential independent factors on trip behaviour. Sixty-two variables that may potentially influence trip behaviour were studied. Different architectures of ANN models were tested. The ANN models were then compared with MLR models in addition to cross-classification models using the same data. The results showed that ANN models were better than their counterpart linear regression models and cross-validation models in terms of residual mean squared error (RMSE).

Introducing a simple feed-back neural network trained by the back-propagation learning algorithm, Peng (2006) concluded that the neural network shows promise for forecasting bids even with extremely limited information. Both linear regression and neural network methods were applied to predict bids. The results showed that linear methods fail to adequately describe bidding behaviour while the neural network exhibits some improvement.

Chen and Huang (2006) developed regression and ANN models to predict the cost and duration of projects for the reconstruction of schools which must be quickly rebuilt. Data for the school reconstruction projects in central Taiwan, which received
the most serious damage from the Chi-Chi Earthquake, were collected and analysed. The analytical results demonstrated that the neural network model trained by the back-propagation learning algorithm is a feasible approach that yields better prediction results than the regression model for school reconstruction projects. Although the prediction abilities of the ANN models and the regression models may differ from one type of reconstruction project to another, the results of this study offered that further application of the ANN methodology to other types of reconstruction projects is highly warranted.

28- Shehab et al., (2010) described the development of cost estimating models for sewer and water network repair projects being undertaken in the United States to improve their condition. To develop these models, data from a set of 54 projects were used. Data pertaining to these projects were first processed to identify the factors that highly impact the overall cost. These factors were then further processed using two approaches, namely, the ANN methodology and MLR analysis, to develop the cost estimating models. Using the same information to develop both models, the study found that the ANN model produced more accurate results when compared to the MLR model.

The studies described above implicitly demonstrate that the ANN methodology tends to provide a method for successfully overcoming the constraints of other conventional techniques (particularly the MLR technique) and therefore has better potential to predict the problem behaviour. However, in contrast to these studies, many studies do also exist which clearly illustrate that the ANN methodology is not in every case superior to traditional statistical modelling approaches. For instance, Mason and Sweeney (1992) found that for estimating a simple linear function, regression analysis could be better than the ANN methodology in terms of both average and maximum error metrics. Some other studies with similar implication are summarized as follows:

1- Williams (1994) developed a back-propagation ANN model to predict the changes in construction cost indexes. Construction cost indexes provide a comparison of cost changes from period to period for a fixed quantity of goods or services. A set of macroeconomic data was tailored for the period from 1967 to 1991. Almost 75% of all cases were used to train the ANN models and the remainder 25% were used to test
the predictive performance of these models. The output from the ANN models was then compared with predictions made by an exponential smoothing and simple linear regression models. The author found that the prediction produced by the ANN model gave a greater error than either exponential smoothing or simple linear regression. Williams concluded that the movement of the cost indexes is amongst the problems that cannot be predicted accurately by a back-propagation ANN model. According to Boussabaine (1996), the failure of this model could be attributed to the selection and design of input variables or failure to find an optimum network topology and fine-tune the network structure (weights, neurons and layers) to obtain a suitable model.

2- Berends (1998) empirically evaluated the robustness of hedonic pricing by developing a standard econometric MLR model and a feed-forward ANN model trained by the back-propagation learning algorithm. Hedonic pricing is a concept that states that the price for a good or service is based upon its physical characteristics. Results indicated that the two models are similar in price prediction. It was then concluded that the MLR modelling technique is a robust technique for evaluating farmland prices.

3- The ability of an ANN model to predict mortality post-cardiac surgery was also compared by Tu et al., (1998) with logistic regression. This study used an extensive dataset of 4,782 patients who underwent coronary artery bypass surgery for model development. Two separate test dataset of 5,309 and 5,517 patients each were used to compare predictive performance. Both ANN and logistic regression models were developed using 11 independent variables that had previously been shown to predict mortality post-cardiac surgery. The ANN model evaluated in this study was a back-propagation network developed with a learning rate of 0.1 and a momentum term of 0.1. The results obtained indicated that the performance of the ANN model and the logistic regression model did not differ significantly.

4- Elhag and Boussabaine (1999) demonstrated the development of cost estimation models using neural networks and regression analysis. Thirty-six newly constructed office buildings were extracted from an online database which were further separated arbitrarily into a training dataset consisting of twenty-seven projects, and a test dataset consisting of the remaining nine projects. The findings showed that the two models were able to map the underlying relationship between the cost factors and the tender price during the training stage and maintained average accuracy percentages of 93.4%
and 91.5% for ANN and MLR, respectively. In the validation stage good
generalization capability was obtained by both models and average accuracy
percentages of 90.9% and 93.3% were achieved by ANN and MLR, respectively.
These results indicate that there was no significant difference in the average accuracy
achieved by the two techniques. The high level of accuracy obtained by the two
models was attributed to the high correlation coefficients between the tender price and
the cost factors. However, it was recognised that the ANN methodology is a data
intensive technique, and therefore to enhance the reliability of ANN models it was
recommended to exploit more projects for future development.

5- McFadden et al., (2001) evaluated two back-propagation ANN models to predict the
operating speed profile of passenger cars. They compared the results with those
obtained from MLR models. The comparison showed that the explanatory and the
prediction powers of ANN models and regression-based models were similar.

6- The effects of number of variables and number of training dataset size on the
performance of ANN models were studied by Wang et al., (2000). The best and the
poorest ANN models were then compared with MLR models for the same sets of
number of variables and training dataset size. The results showed that the poorest
ANN models, in comparison with MLR models, yielded estimating models that were
widely inaccurate in their predictive ability measured by mean absolute error (MAE).
However, the best ANN models compared favourably with MLR models.

7- ANN and MLR models were also developed by Williams (2002) to predict the
completed cost of competitively bid highway projects. Data studied included the low
bid, median bid, standard deviation of the bids, expected project duration, and the
number of bids. Two hundred and seventy-six projects were used to develop both the
MLR and ANN models. Twenty-six projects were also used to serve as an
independent test dataset to identically test these models. The simplest regression
model that used only the natural log of the low bid as the independent variable to
predict the natural log of the completed cost as the dependent variable, was found to
be the best performing predictive model. This regression model produced superior
predictions to the best performing ANN model.

8- Attalla and Hegazy (2003) developed two MLR models and two ANN models to
predict the overall performance of reconstruction projects. Based on a survey of
construction professionals, information was obtained on the reasons behind cost overruns and poor quality from 50 reconstruction projects. The results showed that both the regression and neural network models produced relatively close results in terms of the coefficient of determination ($R^2$) and correlation coefficient ($r$). The difference, however, was that the ANN model was able to develop these results while using the 18 independent variables, whereas the regression model used only 5 independent variables. Hence, it was argued that the ANN sensitivity to all variables in predicting the outcome of a future project may be advantageous because it would give the user of the ANN model a larger opportunity to investigate different project control techniques.

9- Williams et al., (2005) constructed ratios relating the second lowest bid, mean bid, median bid, and maximum bid to the low bid for highway construction projects to study if there are useful patterns in project bids that are indicators of the project completion cost. Regression and ANN models were developed to predict the completed cost of Texas highway projects using the bidding ratio data. A total of 2,766 projects were analysed. Twenty-five percent of these projects were used to test the accuracy of both the regression and ANN models. The input data used were transformed using the natural logarithm. Both the ANN and regression models achieved mixed results when predicting the completed project cost. The model that produced the highest number of accurate predictions was the regression model that included three independent variables. Yet, the best performing model in terms of minimizing mean squared error (MSE) and mean absolute percentage error (MAPE) was the most parsimonious ANN model using only one independent variable to predict the completed cost. The results of this study indicate that the performance of ANN models was unlikely to be significantly different from that obtained for the MLR models.

Based on the above studies, it is concluded that different modelling techniques may vary in usefulness for different prediction contexts. Therefore, it is not yet appropriate to make specific claims regarding the superiority or inferiority of the ANN methodology in comparison to more commonly utilized linear and nonlinear statistical methodologies. Rather, it is likely, based on each technique’s unique advantages and disadvantages, that neural
processing and conventional approaches are complementary and problem-specific (Scarborough, 1995). ANN and conventional modelling approaches are not in competition but they complement each other. In this regard, the ANN methodology may offer a viable alternative to conventional approaches, producing sound results if used sensibly. As described in Chapter 6, the ANN methodology may be particularly useful when relationships within the data are unknown or highly nonlinear, when input information is incomplete, highly unstructured, noisy, or of different types, when underlying data distributions vary, and when assumptions of normality and linearity cannot be assured. It is therefore of great worth to explore the potential of the ANN methodology to complement and improve the effectiveness of other possible methodologies.

8.4 Comparison Measure

The criteria used to compare the predictive performance of atypical modelling techniques vary with different applications. However, in the literature, the error ratio is broadly accepted as the most commonly used tool to examine this performance. This acceptance arises from the fact that the goal of any technique mainly used for forecasting purposes is to identify input-output mappings which minimize the difference between actual and predicted values of the dependent variable. Error ratio is successful in this regard, implying how close a model comes to fulfilling this goal.

While there are a number of measures to be taken for calculating error ratio, mean absolute percentage error (MAPE) is amongst the most popular measures being most often used by researchers in their comparative studies (Emsley et al., 2002; Kim et al., 2004; Stockton and Wang, 2004; Wang et al., 2000; Williams, 2002; Williams et al., 2005). MAPE is a quantitative measure that gives some insight into how effective different models are performing with respect to each other and therefore can be regarded as the measure of effectiveness (MOE). In addition, this measure is often found to be useful for the purpose of reporting the results, because it is expressed in generic percentage term which in turn illustrates how precise an individual model is in predicting the problem that is being investigated. The smaller MAPE, the more precise a model is.
For the reasons stated above, in addition to the coefficient of determination, MAPE was accepted in this research to serve as the comparison measure. The equation, which governs the MAPE, is as follows:

\[
MAPE = \frac{1}{n} \sum_{i=1}^{n} \frac{|RNCC_{pi} - RNCC_{Ai}|}{RNCC_{Ai}} \times 100\% \tag{8.1}
\]

The equation above represents the average percentage difference between the actual value of RNCC of case \(i\) (denoted by \(RNCC_{Ai}\)) and its predicted value (denoted by \(RNCC_{pi}\)). Also, the parameter \(n\) represents the number of samples in either the training dataset (i.e., \(n=121\)) or test dataset (i.e., \(n=37\)).
8.5 Results

MLR models with the contribution of statistically significant independent variables were tailored for the purpose of providing comparisons to their respective ANN models developed using the same sets of independent variables. The characteristics of these models are summarized in Table 8-1.

<table>
<thead>
<tr>
<th>No. of model</th>
<th>MLR Equation</th>
<th>ANN Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( Y = -1.273 + 0.771X_1 )</td>
<td>One hidden layer 0.5 0.2 7</td>
</tr>
<tr>
<td>2</td>
<td>( Y = -0.621 + 0.627X_1 + 0.376X_2 )</td>
<td>Two hidden layers 0.4 0.9 43</td>
</tr>
<tr>
<td>3</td>
<td>( Y = -0.763 + 0.617X_1 + 0.378X_2 + 0.246X_4 )</td>
<td>Two hidden layers 0.9 0.2 29</td>
</tr>
<tr>
<td>4</td>
<td>( Y = -1.025 + 0.648X_1 + 0.362X_2 + 0.226X_4 + 0.136X_8 )</td>
<td>Three hidden layers 0.8 0.1 39</td>
</tr>
<tr>
<td>5</td>
<td>( Y = -1.085 + 0.643X_1 + 0.368X_2 + 0.235X_4 + 0.118X_7 + 0.131X_{11} )</td>
<td>One hidden layer 0.4 0.9 11</td>
</tr>
<tr>
<td>6</td>
<td>( Y = -1.242 + 0.674X_1 + 0.293X_2 + 0.172X_4 + 0.233X_8 + 0.120X_7 + 0.129X_{11} )</td>
<td>Two hidden layers 0.8 0.9 48</td>
</tr>
<tr>
<td>7</td>
<td>( Y = -1.238 + 0.674X_1 + 0.273X_2 + 0.195X_4 + 0.218X_8 + 0.117X_7 + 0.118X_9 + 0.132X_{11} )</td>
<td>Two hidden layers 0.5 0.9 16</td>
</tr>
</tbody>
</table>

\( Y = \ln(\text{RNCC (thousands U.S. $)}) \)
\( X_1 = \ln(\text{Area (m}^2)) \)
\( X_2 = \ln(\text{No. of stories}) \)
\( X_4 = \ln(\text{Weight per unit (ton/m}^2)) \)
\( X_6 = \text{Seismicity} \)
\( X_7 = \text{Soil type} \)
\( X_8 = \text{Plan Irregularity} \)
\( X_{11} = \text{Seismicity} \)
\(^1 \text{LR} = \text{Learning Rate} \)
\(^2 \text{M} = \text{Momentum} \)
\(^3 \text{HN} = \text{Hidden Neurons} \)

There are two main features that make a predictive cost model a very useful tool for solving cost estimation problems. These features are “accuracy” and “generalization” abilities. The performance of the models in Table 8-1 was therefore compared from two perspectives as described below. This comparison and the results taken from it are presented and discussed in the following sections.

- Accuracy ability: It is of interest to measure the accuracy ability of the models developed in this research. This ability addresses the question of how successful the models are at recognizing the patterns of input-output data in the training (i.e., development) dataset. For this purpose, a group of 121 structures was randomly selected from a total of 158 structures and placed in this dataset. Each of the MLR and
ANN models in Chapters 5 and 7 was respectively developed and trained, using this dataset.

- Generalization ability: In addition to the accuracy ability, there is also a need to compare the generalization ability of the models. This ability is the main performance measure that should be carefully examined as it determines how well the models perform on data not included in the training dataset. Also, the generalization ability works as a measure of the sufficiency of the training data to cover much of the problem’s solution space (Hegazy et al., 1994). The remaining 37 structures were used to form the test dataset and examine the models’ generalization ability.

8.5.1 Accuracy Ability

The $R^2$ values obtained from each of the MLR models over the training dataset as well as those that resulted from their respective ANN models are provided in Table 8-2. These values are also graphically illustrated in Figure 8-1.

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Modelling Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
</tr>
<tr>
<td>1</td>
<td>0.5884</td>
</tr>
<tr>
<td>2</td>
<td>0.6310</td>
</tr>
<tr>
<td>3</td>
<td>0.6603</td>
</tr>
<tr>
<td>4</td>
<td>0.6790</td>
</tr>
<tr>
<td>5</td>
<td>0.6915</td>
</tr>
<tr>
<td>6</td>
<td>0.6998</td>
</tr>
<tr>
<td>7</td>
<td>0.7099</td>
</tr>
</tbody>
</table>
Figure 8-1: Comparison of $R^2$ values of the MLR and ANN models over the training dataset
According to Figure 8-1, in both the MLR and the ANN models, the values of $R^2$ are monotonically increasing as the number of independent variables increases by one at a time from the first to the last model. However, the rate of increase is different in the two modelling techniques. The $R^2$ value ranges from 0.59 to 0.71 across the seven MLR models, indicating a 21% increase. This range in the ANN models is from 0.59 to 0.83, indicating a 42% increase which is clearly twice as large as the improvement observed in the MLR models. Furthermore, when the number of independent variables is more than one (i.e., models two through seven), the ANN models always outperform their respective MLR models in terms of $R^2$. This superiority means that in the training dataset the former models more successfully explain the proportion of the total variation in the dependent variable (i.e., Ln (RNCC)). In the first model, however, the performance of both the MLR and the ANN models is almost the same.

In the group of MLR-based models, the last model represents the largest $R^2$ value. The same was also obtained for the group of ANN-based models. The $R^2$ values of the last MLR and ANN models are 0.71 and 0.83, respectively, suggesting that when taking all the seven independent variables into account, the ANN methodology improves the $R^2$ value of the MLR technique by 17%. These independent variables are the ones that were found to be statistically significant in the determination of the RNCC.

The accuracy ability of the MLR and the ANN models is further compared taking the comparison measure into account; that is, MAPE. It is an important performance measure of predictive models enabling comparison of the results of MLR models with those offered by their respective ANN models. Measuring the quality of prediction, MAPE represents the average absolute percentage difference between the actual values of RNCC and those values predicted by either parametric MLR models or non-parametric ANN models. For instance, in Table 8-3 the actual and estimated values of RNCC and the absolute percentage errors produced by the last ANN model are provided along with the same parameters produced by the last MLR model. It should be noted that all the estimated values obtained by both ANN and MLR models were in the natural logarithm format. Using the exponential operator, these values were first transformed to their original format and then used for further comparison.
Table 8-3: Actual and estimated values of RNCC and the absolute percentage errors produced by the last set of MLR and ANN models in the training dataset

<table>
<thead>
<tr>
<th>No. of observation in the training dataset</th>
<th>Actual RNCC (U.S.$)</th>
<th>Estimated RNCC (U.S.$)</th>
<th>Absolute percentage error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MLR</td>
<td>ANN</td>
</tr>
<tr>
<td>1</td>
<td>67368</td>
<td>87401</td>
<td>87737</td>
</tr>
<tr>
<td>2</td>
<td>72391</td>
<td>96909</td>
<td>99582</td>
</tr>
<tr>
<td>3</td>
<td>83084</td>
<td>131487</td>
<td>102410</td>
</tr>
<tr>
<td>4</td>
<td>118877</td>
<td>102558</td>
<td>97110</td>
</tr>
<tr>
<td>5</td>
<td>126622</td>
<td>106918</td>
<td>107322</td>
</tr>
<tr>
<td>6</td>
<td>121372</td>
<td>91069</td>
<td>95230</td>
</tr>
<tr>
<td>7</td>
<td>99229</td>
<td>94458</td>
<td>98968</td>
</tr>
<tr>
<td>8</td>
<td>97013</td>
<td>87345</td>
<td>85071</td>
</tr>
<tr>
<td>9</td>
<td>215216</td>
<td>144630</td>
<td>152947</td>
</tr>
<tr>
<td>10</td>
<td>93998</td>
<td>78159</td>
<td>83199</td>
</tr>
<tr>
<td>11</td>
<td>199279</td>
<td>180782</td>
<td>168993</td>
</tr>
<tr>
<td>12</td>
<td>113559</td>
<td>116309</td>
<td>123875</td>
</tr>
<tr>
<td>13</td>
<td>126061</td>
<td>134015</td>
<td>150832</td>
</tr>
<tr>
<td>14</td>
<td>90222</td>
<td>80264</td>
<td>80888</td>
</tr>
<tr>
<td>15</td>
<td>69871</td>
<td>75982</td>
<td>79154</td>
</tr>
<tr>
<td>16</td>
<td>58888</td>
<td>62513</td>
<td>72542</td>
</tr>
<tr>
<td>17</td>
<td>74957</td>
<td>99576</td>
<td>82212</td>
</tr>
<tr>
<td>18</td>
<td>66882</td>
<td>75564</td>
<td>79011</td>
</tr>
<tr>
<td>19</td>
<td>21469</td>
<td>26428</td>
<td>25630</td>
</tr>
<tr>
<td>20</td>
<td>59242</td>
<td>54625</td>
<td>70174</td>
</tr>
<tr>
<td>21</td>
<td>119184</td>
<td>91572</td>
<td>92527</td>
</tr>
<tr>
<td>22</td>
<td>125977</td>
<td>84152</td>
<td>86708</td>
</tr>
<tr>
<td>23</td>
<td>119065</td>
<td>120099</td>
<td>127594</td>
</tr>
<tr>
<td>24</td>
<td>116537</td>
<td>82131</td>
<td>82003</td>
</tr>
<tr>
<td>25</td>
<td>83588</td>
<td>82426</td>
<td>83031</td>
</tr>
<tr>
<td>26</td>
<td>100160</td>
<td>79281</td>
<td>81690</td>
</tr>
<tr>
<td>27</td>
<td>87508</td>
<td>107087</td>
<td>88974</td>
</tr>
<tr>
<td>28</td>
<td>61264</td>
<td>87094</td>
<td>83046</td>
</tr>
<tr>
<td>29</td>
<td>150651</td>
<td>144708</td>
<td>153177</td>
</tr>
<tr>
<td>30</td>
<td>158202</td>
<td>108786</td>
<td>89829</td>
</tr>
<tr>
<td>31</td>
<td>123900</td>
<td>108329</td>
<td>127961</td>
</tr>
<tr>
<td>32</td>
<td>47111</td>
<td>68461</td>
<td>62898</td>
</tr>
<tr>
<td>33</td>
<td>74862</td>
<td>81058</td>
<td>78656</td>
</tr>
<tr>
<td>34</td>
<td>87956</td>
<td>83848</td>
<td>79781</td>
</tr>
</tbody>
</table>
Table 8-3: Actual and estimated values of RNCC and the absolute percentage errors produced by the last set of MLR and ANN models in the training dataset (Continue)

<table>
<thead>
<tr>
<th>No. of observation in the training dataset</th>
<th>Actual RNCC (U.S.$)</th>
<th>Estimated RNCC (U.S.$)</th>
<th>Absolute percentage error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
<td>ANN</td>
<td>MLR</td>
</tr>
<tr>
<td>35</td>
<td>72910</td>
<td>77921</td>
<td>6.87</td>
</tr>
<tr>
<td>36</td>
<td>31224</td>
<td>57513</td>
<td>84.19</td>
</tr>
<tr>
<td>37</td>
<td>115646</td>
<td>62927</td>
<td>45.59</td>
</tr>
<tr>
<td>38</td>
<td>55298</td>
<td>109481</td>
<td>97.98</td>
</tr>
<tr>
<td>39</td>
<td>111061</td>
<td>103889</td>
<td>6.46</td>
</tr>
<tr>
<td>40</td>
<td>104681</td>
<td>90132</td>
<td>13.90</td>
</tr>
<tr>
<td>41</td>
<td>92689</td>
<td>82293</td>
<td>11.22</td>
</tr>
<tr>
<td>42</td>
<td>51061</td>
<td>65485</td>
<td>28.25</td>
</tr>
<tr>
<td>43</td>
<td>114921</td>
<td>95128</td>
<td>17.22</td>
</tr>
<tr>
<td>44</td>
<td>99175</td>
<td>63287</td>
<td>36.19</td>
</tr>
<tr>
<td>45</td>
<td>59924</td>
<td>72698</td>
<td>21.32</td>
</tr>
<tr>
<td>46</td>
<td>72131</td>
<td>52617</td>
<td>27.05</td>
</tr>
<tr>
<td>47</td>
<td>142231</td>
<td>136379</td>
<td>4.11</td>
</tr>
<tr>
<td>48</td>
<td>144199</td>
<td>128948</td>
<td>10.58</td>
</tr>
<tr>
<td>49</td>
<td>67298</td>
<td>53733</td>
<td>20.16</td>
</tr>
<tr>
<td>50</td>
<td>71379</td>
<td>100036</td>
<td>40.15</td>
</tr>
<tr>
<td>51</td>
<td>174936</td>
<td>129266</td>
<td>26.11</td>
</tr>
<tr>
<td>52</td>
<td>61320</td>
<td>58848</td>
<td>4.03</td>
</tr>
<tr>
<td>53</td>
<td>96079</td>
<td>89397</td>
<td>6.95</td>
</tr>
<tr>
<td>54</td>
<td>67688</td>
<td>72292</td>
<td>6.80</td>
</tr>
<tr>
<td>55</td>
<td>61472</td>
<td>69427</td>
<td>12.94</td>
</tr>
<tr>
<td>56</td>
<td>61472</td>
<td>69427</td>
<td>12.94</td>
</tr>
<tr>
<td>57</td>
<td>61472</td>
<td>69427</td>
<td>12.94</td>
</tr>
<tr>
<td>58</td>
<td>59875</td>
<td>68234</td>
<td>13.96</td>
</tr>
<tr>
<td>59</td>
<td>80067</td>
<td>81862</td>
<td>2.24</td>
</tr>
<tr>
<td>60</td>
<td>93288</td>
<td>88094</td>
<td>5.57</td>
</tr>
<tr>
<td>61</td>
<td>98081</td>
<td>106692</td>
<td>8.78</td>
</tr>
<tr>
<td>62</td>
<td>58345</td>
<td>103159</td>
<td>76.81</td>
</tr>
<tr>
<td>63</td>
<td>131014</td>
<td>179072</td>
<td>36.68</td>
</tr>
<tr>
<td>64</td>
<td>56639</td>
<td>69286</td>
<td>22.33</td>
</tr>
<tr>
<td>65</td>
<td>76330</td>
<td>94166</td>
<td>23.37</td>
</tr>
<tr>
<td>66</td>
<td>78730</td>
<td>98373</td>
<td>24.95</td>
</tr>
<tr>
<td>67</td>
<td>81371</td>
<td>100335</td>
<td>23.31</td>
</tr>
<tr>
<td>68</td>
<td>79232</td>
<td>83123</td>
<td>4.91</td>
</tr>
</tbody>
</table>
Table 8-3: Actual and estimated values of RNCC and the absolute percentage errors produced by the last set of MLR and ANN models in the training dataset (Continue)

<table>
<thead>
<tr>
<th>No. of observation in the training dataset</th>
<th>Actual RNCC (U.S.$)</th>
<th>Estimated RNCC (U.S.$)</th>
<th>Absolute percentage error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
<td>ANN</td>
<td>MLR</td>
</tr>
<tr>
<td>69</td>
<td>85710</td>
<td>92741</td>
<td>93788</td>
</tr>
<tr>
<td>70</td>
<td>163970</td>
<td>141704</td>
<td>132828</td>
</tr>
<tr>
<td>71</td>
<td>92527</td>
<td>78901</td>
<td>88531</td>
</tr>
<tr>
<td>72</td>
<td>111848</td>
<td>85272</td>
<td>93141</td>
</tr>
<tr>
<td>73</td>
<td>34156</td>
<td>42279</td>
<td>35866</td>
</tr>
<tr>
<td>74</td>
<td>61768</td>
<td>57643</td>
<td>63001</td>
</tr>
<tr>
<td>75</td>
<td>197652</td>
<td>143172</td>
<td>153525</td>
</tr>
<tr>
<td>76</td>
<td>66466</td>
<td>71547</td>
<td>78951</td>
</tr>
<tr>
<td>77</td>
<td>73198</td>
<td>78594</td>
<td>76359</td>
</tr>
<tr>
<td>78</td>
<td>58632</td>
<td>49005</td>
<td>63939</td>
</tr>
<tr>
<td>79</td>
<td>142805</td>
<td>143552</td>
<td>144181</td>
</tr>
<tr>
<td>80</td>
<td>216199</td>
<td>174424</td>
<td>182199</td>
</tr>
<tr>
<td>81</td>
<td>197335</td>
<td>199691</td>
<td>189031</td>
</tr>
<tr>
<td>82</td>
<td>107965</td>
<td>91290</td>
<td>85510</td>
</tr>
<tr>
<td>83</td>
<td>98090</td>
<td>96818</td>
<td>90664</td>
</tr>
<tr>
<td>84</td>
<td>63391</td>
<td>51518</td>
<td>58995</td>
</tr>
<tr>
<td>85</td>
<td>58700</td>
<td>78803</td>
<td>73795</td>
</tr>
<tr>
<td>86</td>
<td>98381</td>
<td>94293</td>
<td>102584</td>
</tr>
<tr>
<td>87</td>
<td>98381</td>
<td>94303</td>
<td>102583</td>
</tr>
<tr>
<td>88</td>
<td>120155</td>
<td>133332</td>
<td>125337</td>
</tr>
<tr>
<td>89</td>
<td>252607</td>
<td>162223</td>
<td>149018</td>
</tr>
<tr>
<td>90</td>
<td>125947</td>
<td>117336</td>
<td>109217</td>
</tr>
<tr>
<td>91</td>
<td>58823</td>
<td>21981</td>
<td>58808</td>
</tr>
<tr>
<td>92</td>
<td>92854</td>
<td>86911</td>
<td>87873</td>
</tr>
<tr>
<td>93</td>
<td>111233</td>
<td>130729</td>
<td>119087</td>
</tr>
<tr>
<td>94</td>
<td>150074</td>
<td>113254</td>
<td>116491</td>
</tr>
<tr>
<td>95</td>
<td>123156</td>
<td>140706</td>
<td>120436</td>
</tr>
<tr>
<td>96</td>
<td>64780</td>
<td>106591</td>
<td>100631</td>
</tr>
<tr>
<td>97</td>
<td>123156</td>
<td>139850</td>
<td>120328</td>
</tr>
<tr>
<td>98</td>
<td>24776</td>
<td>45895</td>
<td>33207</td>
</tr>
<tr>
<td>99</td>
<td>157390</td>
<td>137159</td>
<td>134378</td>
</tr>
<tr>
<td>100</td>
<td>75373</td>
<td>88219</td>
<td>96188</td>
</tr>
<tr>
<td>101</td>
<td>106057</td>
<td>116534</td>
<td>120342</td>
</tr>
<tr>
<td>102</td>
<td>94096</td>
<td>117523</td>
<td>103723</td>
</tr>
</tbody>
</table>
Table 8-3: Actual and estimated values of RNCC and the absolute percentage errors produced by the last set of MLR and ANN models in the training dataset (Continue)

<table>
<thead>
<tr>
<th>No. of observation in the training dataset</th>
<th>Actual RNCC (U.S.$)</th>
<th>Estimated RNCC (U.S.$)</th>
<th>Absolute percentage error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
<td>ANN</td>
<td>MLR</td>
</tr>
<tr>
<td>103</td>
<td>93154</td>
<td>97207</td>
<td>98047</td>
</tr>
<tr>
<td>104</td>
<td>52232</td>
<td>42037</td>
<td>61081</td>
</tr>
<tr>
<td>105</td>
<td>125723</td>
<td>146948</td>
<td>121284</td>
</tr>
<tr>
<td>106</td>
<td>138267</td>
<td>124310</td>
<td>125887</td>
</tr>
<tr>
<td>107</td>
<td>181360</td>
<td>141240</td>
<td>149147</td>
</tr>
<tr>
<td>108</td>
<td>53119</td>
<td>72473</td>
<td>78578</td>
</tr>
<tr>
<td>109</td>
<td>110546</td>
<td>120415</td>
<td>115869</td>
</tr>
<tr>
<td>110</td>
<td>146737</td>
<td>148400</td>
<td>153208</td>
</tr>
<tr>
<td>111</td>
<td>52920</td>
<td>43276</td>
<td>52142</td>
</tr>
<tr>
<td>112</td>
<td>110032</td>
<td>81464</td>
<td>83961</td>
</tr>
<tr>
<td>113</td>
<td>62380</td>
<td>49948</td>
<td>63495</td>
</tr>
<tr>
<td>114</td>
<td>81708</td>
<td>65562</td>
<td>78779</td>
</tr>
<tr>
<td>115</td>
<td>74233</td>
<td>113517</td>
<td>97282</td>
</tr>
<tr>
<td>116</td>
<td>62608</td>
<td>61002</td>
<td>64849</td>
</tr>
<tr>
<td>117</td>
<td>103482</td>
<td>87129</td>
<td>88658</td>
</tr>
<tr>
<td>118</td>
<td>21462</td>
<td>34304</td>
<td>22512</td>
</tr>
<tr>
<td>119</td>
<td>11288</td>
<td>17397</td>
<td>12036</td>
</tr>
<tr>
<td>120</td>
<td>154742</td>
<td>145700</td>
<td>144792</td>
</tr>
<tr>
<td>121</td>
<td>152703</td>
<td>101846</td>
<td>102801</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Minimum</th>
<th>Maximum</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>11288</td>
<td>252607</td>
<td>97613</td>
<td>43401</td>
</tr>
<tr>
<td>17397</td>
<td>199691</td>
<td>94685</td>
<td>34443</td>
</tr>
<tr>
<td>12036</td>
<td>189031</td>
<td>94660</td>
<td>32380</td>
</tr>
<tr>
<td>0.52</td>
<td>97.98</td>
<td>20.84</td>
<td>18.62</td>
</tr>
<tr>
<td>0.02</td>
<td>61.57</td>
<td>15.42</td>
<td>13.36</td>
</tr>
</tbody>
</table>

Table 8-4 presents the results of minimum, maximum, mean, and standard deviation of absolute percentage errors of RNCC in the training dataset for all models. The results obtained for MAPE are also explicitly depicted in Figure 8-2.
Table 8-4: Comparison of the MLR and ANN models with respect to different measures of absolute percentage errors of RNCC in the training dataset

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Modelling Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
</tr>
<tr>
<td></td>
<td>Max</td>
</tr>
<tr>
<td>1</td>
<td>168.91</td>
</tr>
<tr>
<td>2</td>
<td>132.06</td>
</tr>
<tr>
<td>3</td>
<td>138.77</td>
</tr>
<tr>
<td>4</td>
<td>125.13</td>
</tr>
<tr>
<td>5</td>
<td>108.04</td>
</tr>
<tr>
<td>6</td>
<td>104.89</td>
</tr>
<tr>
<td>7</td>
<td>97.98</td>
</tr>
</tbody>
</table>

According to Figure 8-2, in both the MLR and the ANN models, the values of MAPE are monotonically decreasing, implying that the inclusion of more independent variables being statistically significant improves the accuracy ability of the models. As observed for $R^2$, the rate of this improvement is more rapid within the ANN models than within the MLR models. However, the incremental improvement of the MAPE values is more even across the MLR models than across the ANN models. The value of MAPE in the former models varies in a range between 25.59% and 20.84%, while this range gets wider in the case of the latter models, moving from 27.48% to 15.42%. These results imply that the ANN modelling technique is more sensitive to the number of independent variables than is the MLR technique.

In either of the two modelling techniques, the best model resulting in the lowest values of MAPE over the training dataset is the last one that included all the seven statistically significant predictors. The MAPE values obtained in the last MLR and ANN models are 20.84% and 15.42%, respectively. These values compare favourably with past studies that have shown that traditional methods of cost estimation are less accurate, as evidenced by reported values of MAPE between 20.8% (Skitmore et al., 1990) and 27.9% (Lowe, 1996) for MLR models and 16.6% (Emsley et al., 2002) for ANN models.
Figure 8-2: Comparison of MAPE values of the MLR and ANN models over the training dataset
8.5.1.1 Effect of the Number of Input Variables

Listed in Table 8-5 are the percentages of improvement that the ANN models achieved over their respective MLR models in terms of MAPE. The following equation is used to calculate this improvement.

\[
\text{Percentage of Improvement} = \frac{MAPE_{\text{MLR}} - MAPE_{\text{ANN}}}{MAPE_{\text{MLR}}} \times 100
\]  

Table 8-5: Desirability of ANN modelling technique over MLR technique in the training dataset

<table>
<thead>
<tr>
<th>No. of model</th>
<th>MAPE (%)</th>
<th>Percentage of improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
<td>ANN</td>
</tr>
<tr>
<td>1</td>
<td>25.59</td>
<td>27.48</td>
</tr>
<tr>
<td>2</td>
<td>24.41</td>
<td>24.08</td>
</tr>
<tr>
<td>3</td>
<td>23.03</td>
<td>22.68</td>
</tr>
<tr>
<td>4</td>
<td>22.28</td>
<td>22.29</td>
</tr>
<tr>
<td>5</td>
<td>21.96</td>
<td>19.51</td>
</tr>
<tr>
<td>6</td>
<td>21.65</td>
<td>17.01</td>
</tr>
<tr>
<td>7</td>
<td>20.84</td>
<td>15.42</td>
</tr>
</tbody>
</table>

The ANN models are in general better than their respective MLR models in terms of MAPE. The first and the fourth models are however the exceptions in this regard. The first MLR and ANN models resulted in a MAPE of 25.59% and 27.48%, respectively. These percentages suggest that using the ANN methodology for the first model, being comprised of only one independent variable, would decline the accuracy ability of the MLR technique by 7.38%. For the fourth model, however, no practical difference was found as both models resulted in a MAPE of 22.3%.

The dominant performance of the ANN models over MLR models in the training dataset is more pronounced if the number of independent variables increases. From Table 8-5 it may be seen that the ANN methodology improves the MAPE values of the second and third MLR models by just 1.34% and 1.54%, respectively. However, this improvement increases rapidly by increasing the number of independent variables such that the errors obtained for the last three MLR models are reduced by approximately 11%, 21%, and 26%, respectively.
Also, fewer independent variables are required by the ANN methodology to achieve a specific level of accuracy ability. For instance, the lowest MAPE value in the group of MLR-based models is achieved for the last model (i.e., 20.84%), using all the seven independent variables. This value is just comparable to the MAPE value of the fifth ANN model (i.e., 19.51%), having two variables less.

Based on the evidence above supporting the general superiority of the ANN models over the MLR models in the training dataset, it is reasonable to hypothesis that the accuracy ability of the ANN modelling technique is higher than that of the MLR technique in the prediction of the RNCC. However, the significance of this hypothesis needs to be statistically tested by using an appropriate statistical test.

The statistical tests fall into two main categories, namely parametric and non-parametric tests. Parametric tests are based on a specific distributional assumption. The most widely used distribution in classical statistical analysis is the normal distribution. This distribution is the basis of much parametric statistical analysis. Parametric tests, such as the t-test, often assume that the sample being studied is from a population having a normal distribution. If using a parametric test, it is therefore vital to confirm that a normal distribution assumption is in fact justified. If this assumption is violated, then it is inadvisable to use the parametric test. Rather, non-parametric tests need to be employed as they make minimal assumptions about the underlying distribution of the data. These tests are called non-parametric because they make no assumptions about the parameters of a distribution, nor do they assume that any particular distribution is being used. Therefore, the non-parametric tests are likely to perform well under a wide range of distributional assumptions. However, in spite of this performance, parametric techniques are in general more powerful than their equivalent non-parametric counterparts as they can detect differences with smaller sample sizes, or detect smaller differences with the same sample size. Therefore, if the distributional assumption can be confirmed, the parametric techniques are generally preferred.

Figures 8-3 through 8-16 depict the normal distributions that are best fitted to absolute percentage errors of RNCC in the training dataset. The first seven figures display the errors produced by the first to the last MLR model, respectively. The last seven figures provide the same information, however this time, for the first to the last ANN model, respectively. The figures also include a number of descriptive statistics including mean, median, mode, 5th, and
95th percentiles of the errors. To examine the suitability of the normal distributions, shown in these figures, the following hypothesis was set:

\[ H_0: \text{The absolute percentage errors of RNCC in the training dataset follow a normal distribution} \]

\[ H_1: \text{The absolute percentage errors of RNCC in the training dataset do not follow a normal distribution} \]

The chi-square (i.e., \( \chi^2 \)) test is used to reject or accept the hypothesis above. For all the models, the \( \chi^2 \) values and their corresponding p-values are tabulated in Table 8-6. The p-value is known as the level of significance and shows the characteristic of the test. When the data are from a normal distribution, the p-value is large and the hypothesis is accepted; when the data are from other distributions, the p-value is smaller than the pre-specified significance level and therefore the hypothesis of an underlying normal distribution is rejected at that significance level. The level of significance adopted for all tests is 5%.

Table 8-6: Results of the \( \chi^2 \) test utilized to assess the null hypothesis in different models that their corresponding absolute percentage errors of RNCC in the training dataset follow a normal distribution

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Modelling Technique</th>
<th>( \chi^2 ) value</th>
<th>p-value of the ( \chi^2 ) test</th>
<th>( \chi^2 ) value</th>
<th>p-value of the ( \chi^2 ) test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
<td></td>
<td></td>
<td>ANN</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>( \chi^2 ) value</td>
<td>p-value of the ( \chi^2 ) test</td>
<td>( \chi^2 ) value</td>
<td>p-value of the ( \chi^2 ) test</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>53.45</td>
<td>0.000*</td>
<td>86.57</td>
<td>0.000*</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>32.42</td>
<td>0.000*</td>
<td>84.59</td>
<td>0.000*</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>45.31</td>
<td>0.000*</td>
<td>84.79</td>
<td>0.000*</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>48.09</td>
<td>0.000*</td>
<td>35.79</td>
<td>0.000*</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>46.50</td>
<td>0.000*</td>
<td>35.99</td>
<td>0.000*</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>31.23</td>
<td>0.000*</td>
<td>42.14</td>
<td>0.000*</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>43.53</td>
<td>0.000*</td>
<td>39.56</td>
<td>0.000*</td>
</tr>
</tbody>
</table>

* Significant at 5% level of significance

According to Table 8-6, the chi-square test led to rejection of the normality hypothesis. The test revealed that for all models the null hypothesis can be confidently
rejected, as the p-values obtained were all extremely small (i.e., p<0.001). Therefore, the use of parametric tests may not be warranted for evaluating the relative performance of MLR and ANN modelling techniques over the training dataset. For this purpose, an alternative non-parametric test should be used.

The Mann-Whitney U test is amongst the most popular non-parametric tests which, similar to other tests of this type, do not assume normality. This test is used in this research to test the following hypothesis:

\[ H_0: \text{There is no significant difference between the absolute percentage errors of RNCC produced by the ANN model in the training dataset and the errors produced by the MLR model in the same dataset using the same number of independent variables.} \]

\[ H_1: \text{There is a significant difference between the absolute percentage errors of RNCC produced by the ANN model in the training dataset and the errors produced by the MLR model in the same dataset using the same number of independent variables.} \]

For each set of the MLR and the ANN models, the results of the Mann-Whitney U test are summarized in Table 8-7. Based on the results presented in this table, the null hypothesis can be rejected for the last two sets of models at the 5% level of significance (i.e., p<0.05). As indicated in the last two rows of Table 8-7, the resulted p-values for the sixth and seventh set of models are equal to 0.037 and 0.016, respectively. In other sets of models, however, the null hypothesis cannot be rejected at the same significance level as the p-values obtained for the first five sets of models are 0.854, 0.386, 0.334, 0.993, and 0.445, respectively.
Table 8-7: Results of the non-parametric Mann-Whitney U test utilized to test the null hypothesis that there is no significant difference between the absolute percentage errors of RNCC produced by the ANN models and the errors produced by their respective MLR models in the training dataset

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Test Statistics</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean Rank</td>
<td>Mann-Whitney U</td>
<td>Wilcoxon W</td>
<td>Z value</td>
<td>Corrected for ties 2-tailed p-value</td>
</tr>
<tr>
<td></td>
<td>MLR (121 Cases)</td>
<td>ANN (121 Cases)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>120.67</td>
<td>122.33</td>
<td>7220.0</td>
<td>14601.0</td>
<td>-0.185</td>
</tr>
<tr>
<td>2</td>
<td>125.40</td>
<td>117.60</td>
<td>6848.0</td>
<td>14229.0</td>
<td>-0.868</td>
</tr>
<tr>
<td>3</td>
<td>125.84</td>
<td>117.16</td>
<td>6795.0</td>
<td>14176.0</td>
<td>-0.965</td>
</tr>
<tr>
<td>4</td>
<td>121.46</td>
<td>121.54</td>
<td>7316.0</td>
<td>14697.0</td>
<td>-0.008</td>
</tr>
<tr>
<td>5</td>
<td>124.93</td>
<td>118.07</td>
<td>6905.0</td>
<td>14286.0</td>
<td>-0.763</td>
</tr>
<tr>
<td>6</td>
<td>130.89</td>
<td>112.11</td>
<td>6184.0</td>
<td>13565.0</td>
<td>-2.087</td>
</tr>
<tr>
<td>7</td>
<td>132.35</td>
<td>110.65</td>
<td>6008.0</td>
<td>13389.0</td>
<td>-2.410</td>
</tr>
</tbody>
</table>

* Significant at 5% level of significance
Figure 8-3: Best fitted normal distribution in the training dataset (MLR model number one)

Figure 8-4: Best fitted normal distribution in the training dataset (MLR model number two)

Figure 8-5: Best fitted normal distribution in the training dataset (MLR model number three)

Figure 8-6: Best fitted normal distribution in the training dataset (MLR model number four)
Figure 8-7: Best fitted normal distribution in the training dataset (MLR model number five)

Figure 8-8: Best fitted normal distribution in the training dataset (MLR model number six)

Figure 8-9: Best fitted normal distribution in the training dataset (MLR model number seven)
Figure 8-10: Best fitted normal distribution in the training dataset (ANN model number one)

Figure 8-11: Best fitted normal distribution in the training dataset (ANN model number two)

Figure 8-12: Best fitted normal distribution in the training dataset (ANN model number three)

Figure 8-13: Best fitted normal distribution in the training dataset (ANN model number four)
Figure 8-14: Best fitted normal distribution in the training dataset (ANN model number five).

Figure 8-15: Best fitted normal distribution in the training dataset (ANN model number six).

Figure 8-16: Best fitted normal distribution in the training dataset (ANN model number seven).
8.5.2 Generalization Ability

Table 8-8 presents the performance of each of the MLR and ANN models when applied to the test dataset, in terms of their associated $R^2$ values. These values are also graphically illustrated in Figure 8-17.

Table 8-8: $R^2$ values of the MLR and ANN models applied to the test dataset

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Modelling Technique</th>
<th>MLR</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.6890</td>
<td>0.6715</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>0.6352</td>
<td>0.6189</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>0.5590</td>
<td>0.5901</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>0.6561</td>
<td>0.6579</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>0.6058</td>
<td>0.6123</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>0.6390</td>
<td>0.6558</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>0.6609</td>
<td>0.7335</td>
</tr>
</tbody>
</table>

According to Figure 8-17, $R^2$ values appear to fluctuate over the MLR models in a similar manner as these values do over the ANN models, implying that adding a new predictor to the model would pose a similar effect in both modelling techniques. However, as opposed to the behaviour observed in the training dataset, $R^2$ values in the test dataset are changing across the models, following no fixed pattern. The value of $R^2$ in the MLR models ranges from 0.56 to 0.69, while this range is shifted upward in the ANN models by almost 0.03, resulting in a range from 0.59 to 0.73. In both modelling techniques, the lowest value of $R^2$ pertains to the third model. However, the largest value in the MLR models belongs, rather surprisingly, to the first model while in the ANN models the largest $R^2$ value is held by the last model, as expected. Comparing the largest values of $R^2$ in both modelling techniques shows the superiority of the ANN methodology by almost 6.5%, which is quite less than that obtained in the training dataset (i.e. 17%).

Further, when the number of independent variables is more than two (i.e., models three through seven) the ANN models slightly outperform their respective MLR models in terms of $R^2$. The largest improvement happens in the last model where the $R^2$ value of the MLR model (i.e., $R^2 = 0.66$) increases by approximately 11%, as compared to that of its
corresponding ANN model (i.e., $R^2 = 0.77$). Again, this improvement is well less than that reported when comparing the $R^2$ values of these models in the training dataset (i.e., 17%).
Figure 8-17: Comparison of $R^2$ values of the MLR and ANN models over the test dataset
The generalization ability of the MLR and ANN models is further examined with respect to the comparison measure. This time MAPE is used to measure the average absolute percentage difference between the actual and predicted values of RNCC in the test dataset. As an illustrative example, the actual values of RNCC in the test dataset are contrasted with the estimated values obtained from the last MLR and ANN models and the produced absolute percentage errors are presented in Table 8-9. Similar to the procedure used in the training dataset, all the estimated values obtained by both modelling techniques in the test dataset are also back-transformed from natural logarithm space into linear space, using the exponential operator.

Table 8-9: Actual and estimated values of RNCC and the absolute percentage errors produced by the last set of MLR and ANN models in the test dataset

<table>
<thead>
<tr>
<th>No. of observation in the test dataset</th>
<th>Actual RNCC (U.S.$)</th>
<th>Estimated RNCC (U.S.$)</th>
<th>Absolute percentage error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MLR</td>
<td>ANN</td>
</tr>
<tr>
<td>1</td>
<td>66930</td>
<td>93887</td>
<td>89733</td>
</tr>
<tr>
<td>2</td>
<td>71421</td>
<td>58415</td>
<td>64112</td>
</tr>
<tr>
<td>3</td>
<td>45941</td>
<td>48875</td>
<td>37828</td>
</tr>
<tr>
<td>4</td>
<td>56518</td>
<td>67200</td>
<td>68953</td>
</tr>
<tr>
<td>5</td>
<td>63628</td>
<td>61490</td>
<td>75497</td>
</tr>
<tr>
<td>6</td>
<td>78556</td>
<td>103732</td>
<td>108734</td>
</tr>
<tr>
<td>7</td>
<td>117379</td>
<td>135363</td>
<td>145118</td>
</tr>
<tr>
<td>8</td>
<td>103224</td>
<td>94423</td>
<td>100202</td>
</tr>
<tr>
<td>9</td>
<td>73928</td>
<td>92160</td>
<td>97009</td>
</tr>
<tr>
<td>10</td>
<td>30439</td>
<td>50196</td>
<td>33135</td>
</tr>
<tr>
<td>11</td>
<td>101367</td>
<td>119026</td>
<td>127398</td>
</tr>
<tr>
<td>12</td>
<td>101504</td>
<td>118431</td>
<td>126771</td>
</tr>
<tr>
<td>13</td>
<td>89928</td>
<td>98026</td>
<td>92358</td>
</tr>
<tr>
<td>14</td>
<td>164787</td>
<td>106391</td>
<td>164136</td>
</tr>
<tr>
<td>15</td>
<td>273777</td>
<td>301552</td>
<td>214784</td>
</tr>
<tr>
<td>16</td>
<td>61936</td>
<td>38727</td>
<td>60090</td>
</tr>
<tr>
<td>17</td>
<td>70939</td>
<td>55671</td>
<td>65394</td>
</tr>
<tr>
<td>18</td>
<td>68605</td>
<td>48141</td>
<td>62583</td>
</tr>
<tr>
<td>19</td>
<td>62865</td>
<td>46062</td>
<td>64033</td>
</tr>
<tr>
<td>20</td>
<td>48457</td>
<td>43824</td>
<td>61833</td>
</tr>
<tr>
<td>21</td>
<td>68355</td>
<td>65128</td>
<td>68982</td>
</tr>
</tbody>
</table>
Table 8-9: Actual and estimated values of RNCC and the absolute percentage errors produced by the last set of MLR and ANN models in the test dataset (Continue)

<table>
<thead>
<tr>
<th>No. of observation in the test dataset</th>
<th>Actual RNCC (U.S.$)</th>
<th>Estimated RNCC (U.S.$)</th>
<th>Absolute percentage error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>MLR</td>
<td>ANN</td>
</tr>
<tr>
<td>22</td>
<td>52900</td>
<td>38543</td>
<td>67144</td>
</tr>
<tr>
<td>23</td>
<td>147489</td>
<td>113668</td>
<td>105975</td>
</tr>
<tr>
<td>24</td>
<td>139020</td>
<td>170013</td>
<td>146450</td>
</tr>
<tr>
<td>25</td>
<td>120383</td>
<td>102279</td>
<td>110472</td>
</tr>
<tr>
<td>26</td>
<td>69259</td>
<td>114875</td>
<td>110339</td>
</tr>
<tr>
<td>27</td>
<td>78673</td>
<td>48873</td>
<td>56375</td>
</tr>
<tr>
<td>28</td>
<td>61566</td>
<td>50252</td>
<td>63693</td>
</tr>
<tr>
<td>29</td>
<td>115083</td>
<td>84099</td>
<td>100330</td>
</tr>
<tr>
<td>30</td>
<td>45102</td>
<td>50904</td>
<td>63843</td>
</tr>
<tr>
<td>31</td>
<td>45102</td>
<td>51426</td>
<td>64005</td>
</tr>
<tr>
<td>32</td>
<td>46650</td>
<td>51328</td>
<td>63975</td>
</tr>
<tr>
<td>33</td>
<td>55789</td>
<td>51361</td>
<td>63985</td>
</tr>
<tr>
<td>34</td>
<td>293320</td>
<td>163018</td>
<td>154262</td>
</tr>
<tr>
<td>35</td>
<td>66661</td>
<td>67960</td>
<td>62824</td>
</tr>
<tr>
<td>36</td>
<td>66697</td>
<td>84691</td>
<td>67110</td>
</tr>
<tr>
<td>37</td>
<td>57982</td>
<td>88840</td>
<td>97334</td>
</tr>
</tbody>
</table>

Table 8-10 summarises the performance of all models applied to the test dataset in terms of various statistics calculated for absolute percentage errors of RNCC. The results obtained for MAPE are specifically depicted in Figure 8-18.
Table 8-10: Comparison of the MLR and ANN models with respect to different measures of absolute percentage errors of RNCC in the test dataset

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Modelling Technique</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
<td>Max</td>
<td>Min</td>
<td>Mean</td>
<td>Std.D</td>
<td>ANN</td>
<td>Max</td>
</tr>
<tr>
<td>1</td>
<td>65.63</td>
<td>0.62</td>
<td>22.90</td>
<td>15.45</td>
<td></td>
<td>67.88</td>
<td>0.82</td>
</tr>
<tr>
<td>2</td>
<td>82.16</td>
<td>0.11</td>
<td>24.74</td>
<td>17.55</td>
<td></td>
<td>81.54</td>
<td>1.35</td>
</tr>
<tr>
<td>3</td>
<td>89.37</td>
<td>0.42</td>
<td>24.78</td>
<td>19.67</td>
<td></td>
<td>94.73</td>
<td>3.21</td>
</tr>
<tr>
<td>4</td>
<td>72.94</td>
<td>0.29</td>
<td>22.29</td>
<td>18.18</td>
<td></td>
<td>67.57</td>
<td>0.60</td>
</tr>
<tr>
<td>5</td>
<td>79.98</td>
<td>1.65</td>
<td>24.04</td>
<td>17.87</td>
<td></td>
<td>74.58</td>
<td>0.55</td>
</tr>
<tr>
<td>6</td>
<td>71.44</td>
<td>1.63</td>
<td>23.33</td>
<td>16.99</td>
<td></td>
<td>62.60</td>
<td>0.38</td>
</tr>
<tr>
<td>7</td>
<td>65.86</td>
<td>1.95</td>
<td>23.22</td>
<td>15.87</td>
<td></td>
<td>67.87</td>
<td>0.39</td>
</tr>
</tbody>
</table>

According to Figure 8-18, MAPE values in the test dataset are varying across the models with no discernible trend. This variation is again dissimilar to the variation observed for MAPE values in the training dataset. As shown earlier in Figure 8-2, MAPE values in the training dataset were constantly decreasing from the first to the last MLR and ANN models. In spite of this dissimilarity, however, the two datasets bear a striking resemblance to each other. This similarity comes into view when examining the sensitivity of the MLR and ANN modelling techniques to the number of independent variables. According to Table 8-10, MAPE values of the MLR models vary in a range from 22.29% to 24.78%, while these values in the ANN models bounce in a range from 20.60% to 28.93%, which is clearly more than three times wider. These results further justify the argument, made previously in Sections 8-5-1 and 8-5-1-1, that the ANN methodology is more sensitive to the number of independent variables than is the MLR technique.

The best model amongst the MLR-based models resulting in the lowest value of MAPE (i.e., 22.29%) is the fourth model. This model slightly outperforms the first MLR model which was found to excel in terms of $R^2$. However, in ANN-based models, the model which surpasses other models in terms of both MAPE and $R^2$ is the last one. This model yields the minimum MAPE of 20.60%. The comparison of the lowest values of MAPE shows a 7.6% improvement in the generalization ability of the MLR analysis. This improvement is however minute in comparison with the improvement achieved for the accuracy ability of the MLR analysis in the training dataset (i.e., 26%).
Figure 8-18: Comparison of MAPE values of the MLR and ANN models over the test dataset
8.5.2.1 Effect of the Number of Input Variables

Table 8-11 compares the relative prediction performance of the MLR models with their corresponding ANN models on the test dataset, with respect to MAPE.

Table 8-11: Desirability of ANN modelling technique over MLR technique in the test dataset

<table>
<thead>
<tr>
<th>No. of model</th>
<th>MAPE (%)</th>
<th>Percentage of improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
<td>ANN</td>
</tr>
<tr>
<td>1</td>
<td>22.90</td>
<td>25.60</td>
</tr>
<tr>
<td>2</td>
<td>24.74</td>
<td>28.93</td>
</tr>
<tr>
<td>3</td>
<td>24.78</td>
<td>27.85</td>
</tr>
<tr>
<td>4</td>
<td>22.29</td>
<td>23.97</td>
</tr>
<tr>
<td>5</td>
<td>24.04</td>
<td>24.66</td>
</tr>
<tr>
<td>6</td>
<td>23.33</td>
<td>25.03</td>
</tr>
<tr>
<td>7</td>
<td>23.22</td>
<td>20.60</td>
</tr>
</tbody>
</table>

According to the table above, all the MLR models show a better predictive performance on the test dataset, except the last one. The last MLR and ANN models resulted in a MAPE of 23.22% and 20.60%, respectively. These values suggest that using the ANN methodology for the last model, in which all the statistically significant predictors are simultaneously entered, increases the generalization ability of the MLR technique by approximately 11%.

Further, the relative performance indicator, measured in terms of the percentage of improvement, signifies that the dominant performance of the first six MLR models over their respective ANN models is in general lessening as the number of independent variables increases. According to the last column of Table 8-11, the ANN methodology undesirably increases the MAPE values of the first and second models developed by utilising MLR analysis by 11.80% and 16.92%, respectively. Nevertheless, the former technique becomes less and less unfavourable in the next three models where the increase rate of MAPE diminishes to 12.39%, 7.51%, and 2.56%, respectively. This improvement is however interrupted in the sixth model where the increase rate of MAPE gets larger, with a value of 7.28%. As mentioned earlier, the general behaviour of the ANN and MLR techniques is
totally reversed in the last model where the former appears to be more precise than the latter by 11.25%.

It is shown so far that the generalization ability of the ANN and MLR modelling techniques is likely to be dissimilar as they produce different values of MAPE when being applied to the test dataset. To examine whether this discrepancy is merely a coincidence or instead is statistically significant, an appropriate statistical test is required to be carried out for each set of models. As discussed in Section 8-5-1-1, the type of the statistical test, being often classified as either parametric or non-parametric, should be decided upon based on whether the absolute percentage errors in the test dataset follow the normal distribution. If the normal distribution assumption is satisfied, then the parametric t-test is preferable; otherwise, it is better to use the non-parametric Mann-Whitney U test.

Figures 8-19 through 8-25 depict the normal distributions that are best fitted to absolute percentage errors resulted from applying the first through the last MLR model over the test dataset, respectively. Figures 8-26 through 8-32 show the same information for the first through the last ANN model, respectively. The figures also include a number of descriptive statistics including mean, median, 5th, and 95th percentiles of the errors. To examine the suitability of the normal distributions, being overlaid to the errors of each model, the following hypothesis was set and tested:

H₀: The absolute percentage errors of RNCC in the test dataset follow a normal distribution

H₁: The absolute percentage errors of RNCC in the test dataset do not follow a normal distribution

The chi-square (i.e., $\chi^2$) test was used to reject or accept the hypothesis above. For all the models, the $\chi^2$ values and their corresponding p-values are tabulated in Table 8-12.
Table 8-12: Results of the $\chi^2$ test utilized to assess the null hypothesis in different models that their corresponding absolute percentage errors of RNCC in the test dataset follow a normal distribution

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Modelling Technique</th>
<th>MLR</th>
<th>ANN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$ value</td>
<td>p-value of the $\chi^2$ test</td>
<td>$\chi^2$ value</td>
</tr>
<tr>
<td>1</td>
<td>2.92</td>
<td>0.571</td>
<td>5.19</td>
</tr>
<tr>
<td>2</td>
<td>2.54</td>
<td>0.637</td>
<td>1.78</td>
</tr>
<tr>
<td>3</td>
<td>10.49</td>
<td>0.033*</td>
<td>19.57</td>
</tr>
<tr>
<td>4</td>
<td>3.68</td>
<td>0.452</td>
<td>8.22</td>
</tr>
<tr>
<td>5</td>
<td>17.30</td>
<td>0.002*</td>
<td>5.57</td>
</tr>
<tr>
<td>6</td>
<td>5.95</td>
<td>0.203</td>
<td>5.57</td>
</tr>
<tr>
<td>7</td>
<td>5.57</td>
<td>0.234</td>
<td>8.97</td>
</tr>
</tbody>
</table>

* Significant at 5% level of significance

According to Table 8-12, the null hypothesis can be rejected in three out of the fourteen MLR and ANN models. These three models are the third and the fifth MLR models and the third ANN model, as their p-values were less than the pre-specified significance level of 5%. Therefore, use of the parametric t-test is warranted in all sets, except the third and the fifth ones. For these sets, the generalization ability of the MLR models is statistically compared with that of their associated ANN models using the Mann-Whitney U test. To conduct this comparison, the following hypothesis, initially established in Section 8-5-1-1, was tested again. The results of the t-test and Mann-Whitney U test are reported in Table 8-13 and 8-14, respectively. It should be noted that besides the normality assumption, another assumption of the t-test is that both groups under the test have equal variances. This assumption can be tested using Levene’s Test of Equality of Variances. The results of this test are also tabulated in Table 8-13, indicating that the equal variances assumption can be confirmed in all sets of models listed in this table at the 5% level of significance.

$H_0$: There is no significant difference between the absolute percentage errors of RNCC produced by the ANN model in the test dataset and those produced by the MLR model in the same dataset using the same number of independent variables.

$H_1$: There is a significant difference between the absolute percentage errors of RNCC produced by the ANN model in the test dataset and those produced by the MLR model in the same dataset using the same number of independent variables.
Table 8-13: Results of the parametric t-test utilized to test the null hypothesis that there is no significant difference between the absolute percentage errors of RNCC produced by the ANN models and those produced by their respective MLR models in the test dataset (All sets of models excluding the third and the fifth models)

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Levene’s Test for equality of variances</th>
<th>t-test for equality of means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-value</td>
<td>P-value</td>
</tr>
<tr>
<td>1</td>
<td>0.370</td>
<td>0.545</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>0.399</td>
<td>0.530</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>0.839</td>
<td>0.363</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.287</td>
<td>0.594</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>0.544</td>
<td>0.463</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8-14: Results of the non-parametric Mann-Whitney U test utilized to test the null hypothesis that there is no significant difference between the absolute percentage errors of RNCC produced by the ANN models and those produced by their respective MLR models in the test dataset (The third and the fifth sets of models)

<table>
<thead>
<tr>
<th>No. of model</th>
<th>Mean Rank</th>
<th>Test Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR (37 Cases)</td>
<td>AN (37 Cases)</td>
</tr>
<tr>
<td>3</td>
<td>35.92</td>
<td>39.08</td>
</tr>
<tr>
<td>5</td>
<td>38.00</td>
<td>37.00</td>
</tr>
</tbody>
</table>
As shown in the last column of Tables 8-13 and 8-14, the p-values obtained for the first to the last set of models equate to 0.473, 0.327, 0.527, 0.704, 0.841, 0.672 and 0.498, respectively. These values indicate that in all sets of models there is insufficient evidence to reject the null hypothesis of equality between the generalization ability of the MLR and ANN modelling techniques (i.e., p>0.05).
Figure 8-19: Best fitted normal distribution in the test dataset (MLR model number one)

Figure 8-20: Best fitted normal distribution in the test dataset (MLR model number two)

Figure 8-21: Best fitted normal distribution in the test dataset (MLR model number three)

Figure 8-22: Best fitted normal distribution in the test dataset (MLR model number four)
Figure 8-23: Best fitted normal distribution in the test dataset (MLR model number five)

Figure 8-24: Best fitted normal distribution in the test dataset (MLR model number six)

Figure 8-25: Best fitted normal distribution in the test dataset (MLR model number seven)
Figure 8-26: Best fitted normal distribution in the test dataset (ANN model number one)

Figure 8-27: Best fitted normal distribution in the test dataset (ANN model number two)

Figure 8-28: Best fitted normal distribution in the test dataset (ANN model number three)

Figure 8-29: Best fitted normal distribution in the test dataset (ANN model number four)
Figure 8-30: Best fitted normal distribution in the test dataset (ANN model number five)

Figure 8-31: Best fitted normal distribution in the test dataset (ANN model number six)

Figure 8-32: Best fitted normal distribution in the test dataset (ANN model number seven)
8.6 Conclusion

In this chapter, the seven best ANN models developed in Chapter 7 were respectively compared with their MLR counterparts developed in Chapter 5. Accuracy and generalization abilities, as measured by the mean absolute percentage error (MAPE), were the criteria used for this comparison. As defined in this study, the former ability corresponds to the performance of the models over the training dataset, whereas the latter ability attempts to elucidate this performance over the test dataset.

Considering the training dataset, the ranges over which the MAPE values were situated were 20.84% to 25.59% in the case of the MLR models and 15.42% to 27.48% in the case of the ANN models. These ranges imply that the ANN methodology is more sensitive to the number of independent variables than is the MLR technique. In both the MLR and ANN modelling techniques, the best model resulting in the lowest values of MAPE was the last model where all seven statistically significant predictors were simultaneously included. The comparison of these MAPE values showed that the ANN model outperformed its respective MLR model by approximately 26%. This improvement was tested using the non-parametric Mann-Whitney U test and proved to be statistically significant (i.e., p<0.05). The same conclusion emerged when comparing the accuracy ability of the models in the sixth set, where an almost 21% reduction in error was achieved for the ANN model. Again, this reduction was found to be statistically significant at the 5% level of significance. However, the results of the other ANN models with fewer independent variables showed that their accuracy ability did not significantly differ from that of their respective MLR models (i.e., p>0.05). It was then concluded that the significance of the superiority of the ANN methodology over the MLR analysis is more pronounced in situations where a greater number of independent variables are to be taken into account. Consequently, it was concluded that the greater the number of independent variables, the more statistically significant the difference between the accuracy ability of MLR and ANN modelling techniques will be and, as a result, the more favourable the latter will be.

In the test dataset, MAPE values of the ANN models varied in a range from 20.60% to 28.9%. This range was more than three times wider than that of the MLR models, as previously shown to be between 22.29% and 24.78%. This finding provides further support to the same conclusion put forward in the training dataset that the ANN methodology is more sensitive to the number of independent variables than is the MLR analysis. The best
performing MLR and ANN-based models representing the lower bound of the ranges above were the fourth MLR and the last ANN models respectively, while the latter outperformed the former by 7.6%. The last ANN model also outperformed its respective MLR model, using the same number of independent variables, by almost 11%. In the rest of the models, however, the MLR models revealed a better predictive performance. In general, the dominant performance of these MLR models was observed to be depreciating as the number of independent variables increased. This observation further strengthens the idea of applying the ANN methodology to cases where more variables are present. Further, by using an appropriate statistical test, it was found that there is insufficient evidence to reject the null hypothesis of equality between the generalization ability of the MLR and ANN modelling techniques at the 5% level of significance. This particular finding was inconsistent with the conclusion drawn for the accuracy ability of these techniques. One of the possible reasons behind this incompatibility was envisaged to at least partly arise from the fact that the training and test datasets were of different size. The size of the training dataset, being composed of 121 data points, was nearly three times larger than that of the test dataset, consisting of 37 data points. It was then concluded that, firstly, the relative performance of the MLR and ANN modelling techniques can be strongly dependent on the number of data points; and secondly, the latter technique is likely to be more statistically desirable when dealing with a larger amount of data.
Attempts to predict construction cost represent a problem of continual concern and interest to both researchers and practitioners. Such an attempt was undertaken in this research to predict the construction cost associated with retrofitting existing structures. The main reasons accounting for the importance of this prediction were outlined in Chapter 1. As described in this chapter, there has been an increasing need in earthquake-prone countries to evaluate and possibly enhance the seismic performance of existing structures as the presence of seismically vulnerable buildings has posed serious threats to residents and the public at large. Consequently, owners of such buildings are frequently faced with the decision of whether or not to retrofit their buildings in order to lower the potential losses due to earthquakes. Retrofit cost estimation is one of the major factors having a crucial impact on this critical decision (FEMA 227 and 228, 1992). This estimation is also essential to improve decision-makers’ understanding of the likely condition of their inventory and to enable them to establish sound retrofit programs. Similar to other construction activities, an accurate estimation of the retrofit construction cost is highly desirable in the early stage of development of retrofit projects. In this research, the areas for which such an early accurate estimation is of particular significance were identified and classified into five main categories, being (1) socio-economic feasibility of taking retrofit into account; (2) funding retrofit programs; (3) setting an optimal retrofit compliance interval; (4) improving the accuracy of decision-making analysis outcomes; and (5) framing rational retrofit policy and regulation.

Being recognized as a novel area of research in the engineering and construction industry, retrofit cost estimation has received little attention thus far. Only a few retrofit cost estimating (RCE) models were found to be available in the literature, whose implementation has been mainly restricted for retrofit projects in the US. The literature search conducted in this research revealed that no mathematically stringent RCE models have been previously proposed for retrofit projects in other earthquake-prone countries. In addition, the reliability of the available RCE models is uncertain because of a number of factors, as described in Chapter 2. For instance, one of the sources of this uncertainty is related to the approaches utilized in the development of these models. These approaches utilized subjective judgement
and regression analysis. The accuracy of the available RCE models is essentially limited by the human errors and certain assumptions, which are respectively inherent in the nature of the former and latter approaches. The uncertainty raised by these limitations is further heightened by the fact that, when developing the RCE models, no effort was made to verify the prediction accuracy of these models on datasets other than those from which the models were originally derived.

RCE modelling is a multidisciplinary task which is subject to variations arising from a variety of influential variables associated with the various engineering areas of expertise. To develop reliable RCE models, these variables were identified and classified into four main groups, being (1) building-based, (2) site-based, (3) social-based, and (4) retrofit policy-based categories. Variables pertaining to three of these four categories were studied in this research. The total number of these variables was fourteen, nearly half of which have never been studied in the literature. Variables associated with the social-based category were held fixed as they were not varied with the attributes they could possibly take on in this research. Besides this classification, an innovative framework was developed for activities, which may be performed through the whole retrofit procedure. This procedure normally starts from visual rapid screening and continues to retrofit construction work. The developed framework was utilized in this research to estimate the cost of those activities that together constitute the retrofit construction cost. This estimation was performed on the basis of the data collected in Iran within a seven-month survey period.

This research is a response to the need for a greater amount of highly reliable data. The database generated in this research is one of the largest databases that can be identified in the literature for cost modelling purposes. Besides the database size, the reliability of this database was also ensured because the most plausible documents were used in this research to elicit the required data. The plausibility of these documents was examined and ascertained by managerial authorities through an extensive refereeing process. Notably, a similar data collection effort was initially attempted in New Zealand by creating an exclusive website (www.retrofitcost.net). However, this effort was hampered for the reasons explained in Chapter 3. The data collection effort in Iran resulted in the creation of a database of 158 data points, each pertaining to a particular earthquake-prone public school with framed structures (i.e., concrete or steel structure). Each data point consisted of the retrofit net construction cost (RNCC) of a given school, together with data on the identified variables influencing the RNCC of that particular school. The RNCC represented, by definition, the basic retrofit
construction cost which was then multiplied by a number of coefficients to comprise the retrofit construction tender price. The descriptive statistical properties of the RNCC variable and its influential variables were explored and the results were provided in Chapter 4. In this research, the RNCC and its influential variables served as the dependent and independent variables, respectively. These variables were tailored to develop parametric and non-parametric RCE models, using the multi-linear regression (MLR) analysis and the artificial neural network (ANN) methodologies, respectively.

Through utilization of the modelling approaches stated above, the available database was randomly partitioned into two datasets, being (1) the development or training dataset, and (2) the hold-out or test dataset. Of the total 158 samples in the research database, 75% (i.e., 121 samples) were randomly separated out to constitute the former dataset, while the remaining 25% (i.e., 37 samples) were held to constitute the latter dataset. Consequently, the predictive performance of the models, whether developed using the MLR analysis or the ANN methodology, was investigated in two respects: (1) when the models were applied to the development or training samples, and (2) when the models were applied to the hold-out or testing samples. The predictive performance obtained from the former and the latter applications was defined as the accuracy ability and the generalization ability, respectively.

An attempt was made in this research to address and resolve an issue which has hitherto been one of the major deficits of the previous studies conducted in the field of RCE modelling. In this attempt, the backward elimination (BE) regression technique was employed to properly explore the extent of the influence of independent variables on the RNCC and, as a result, to identify those variables that made a statistically significant contribution to the prediction of the RNCC. Of the fourteen independent variables examined in this research, seven variables were found to be critical predictors of the RNCC at the 5% significance level (i.e., p<0.05). These variables pertained to either the building-based category (i.e., building area, building number of stories, building weight indicator, building configuration in terms of plan irregularity, and building structural type) or the site-based category (i.e., seismicity, and soil type). The remaining seven variables, including those which pertained to the retrofit policy-based category, appeared to have an inappreciable influence on the RNCC determination, when taking the same significance level into account (i.e., p>0.05).
Using the BE technique, fourteen different regression models were developed. The accuracy ability of these models was compared and the most accurate model was defined on the basis of the selection criteria examined in this research, including the leave-one-out cross validation technique, minimum absolute t-value, and residual mean squares of errors, together with the Akaike, modified Akaike, and Bayes information criteria. The results of this comparison revealed that the model with the inclusion of all seven statistically significant variables resulted in the highest accuracy ability. This model was exactly the same as the model obtained from the implementation of the BE technique, when the cut-off p-value for removing the insignificant variables was set to 5%. Consequently, it was concluded that the BE technique, if being used with the 5% cut-off p-value, is a robust analytical technique for the development of reliable MLR models. Besides its robustness, this technique has the great advantage of reducing the computational time. This advantage is more pronounced for cases where a large number of independent variables are present as this presence makes the examination of all feasible regression models impractical.

The MLR model with the highest accuracy ability allowed for explanation of 71% of the RNCC variation in the development dataset. No evidence was found that the underlying assumptions of the regression analysis were violated by this model. This finding led to another major contribution of this research, which was the interpretation of the causal relationships between the RNCC and each of its independent variables in the MLR model stated above. According to the functional form specified for this model, it was concluded that, holding all other variables constant, if the building area and number of stories increases by 1%, the RNCC mean value rises by approximately 0.67% and 0.27%, respectively. The former percentage implies that the RNCC increases proportionally less than the increase in the building area. Consequently, the retrofit cost indicator (RCI), specified on a dollar-per-square-metre basis (U.S. $/m^2), would decrease as the building area increases. This relationship suggests that as the building increases in size, the importance of the building area in determining RCI values or, alternatively, in describing the RCI total variability is proportionally decreasing. The latter percentage (i.e., 0.27%) indicates that the difference between the RNCC values of high-rise buildings would be less significant in comparison with the difference occurring between the RNCC values of low-rise buildings. This indication suggests that the RNCC of buildings with a larger number of stories will be subject to less variation than those of low-rise buildings. The causal relationship between the RNCC and the building weight indicator (measured as the total building seismic weight in ton/m^2) was also
investigated and it was found that, holding all other variables constant, a one percent increase in the latter leads to a 0.19% increase in the mean value of the former. In addition, locating the building in very high seismic zones, the absence of an earthquake-resistant system in an existing building, the presence of soil types with poor seismic characteristics, and the presence of plan irregularity increased the mean value of the RNCC by almost 24%, 14%, 12%, and 12%, respectively.

The last and probably the most important contribution of this research to the field of parametric RCE modelling was made when comparing the generalization ability of the last seven regression models derived from the implementation of the BE technique. These models were selected because they were all developed using the most statistically significant independent variables. The results of this comparison, as measured in terms of determination of coefficient (R²), indicated that the most parsimonious regression model having only one independent variable (i.e., area) produced the highest R² value. Taking this indication into account, a novel model was established for the general purpose of estimating the retrofit construction cost. The generic form of this model was formulated in the form of C=KA^B, where the terms K and B are constants, and the terms C and A represent the retrofit construction cost in thousands of U.S. $ and the building area in square metre, respectively. The value of these constants for estimating the RNCC of the retrofit projects studied in this research was determined to be 0.280 and 0.771, respectively. The utilization of this model is highly advantageous to building owners, project managers, and cost planners, as the model enables them to simply predict the construction cost of a retrofit project based on just one variable (i.e., the building area) whose value is almost always known before the project commences. The significance of this advantage is best appreciated in cases where no detailed information about the building characteristics, site conditions, and also retrofit solution plans is available at the time of prediction. Although the MLR models developed in this research are valuable in terms of offering simple parametric solutions for the problem of RCE modelling, the predictive performance of these models was subject to limitations imposed by certain assumptions inherent in the MLR analysis. To alleviate these limitations, an alternative modelling approach was employed in this research which holds great promise for solving complex problems, such as RCE modelling.

In this research, intelligent non-parametric RCE models were developed for the first time in the literature by means of the ANN methodology. This methodology is a new breed of non-parametric estimators that has recently evolved based on artificial intelligence. The ANN
methodology is theoretically well-suited to predict a complex dependent variable, such as the RNCC, for several reasons. As described in Chapter 6, ANN models are able to detect any patterns that can be found in data. This detection enables better prediction of a dependent variable, particularly when a precise knowledge of all contributing independent variables is lacking, when a large number of independent variables are to be studied, when the relationships amongst variables are unknown, complex, or nonlinear, when data are flawed or missing, and when data are sparse and noisy. The ANN paradigms utilized in this research were the feed-forward multi-layer perceptron (MLP) network, sigmoid logistic function, and the back-propagation learning algorithm as the network architecture, activation function and the learning algorithm, respectively.

Using the paradigms stated above, a new procedure was proposed for the successful development of ANN models. Three specific objectives were aimed at the application of this procedure, being (1) to determine the most appropriate number of hidden layers, (2) to determine the most appropriate value for the learning parameters including the learning rate (LR) and momentum (M), and (3) to determine the most appropriate number of hidden neurons. The criterion utilized for these determinations was geared more towards the generalization ability, rather than the accuracy ability. The development procedure consisted of two phases, which were named the preliminary development phase and the final development phase. The first two and the last of the aforementioned objectives were accomplished through implementation of the former and the latter phase, respectively. Upon completion of these phases, for each of the last seven MLR models derived from the implementation of the BE technique, the best correspondent ANN model was developed.

The application of the best ANN models to the training and test datasets resulted in the finding that the last ANN model, including seven input (i.e., independent) variables, maintained the highest accuracy and generalization ability. This finding confirmed the importance of the seven variables, which were shown to be statistically significant in the development of the MLR analysis. However, these variables were of different importance for the prediction of the RNCC. Taking the generalization ability of the ANN models into consideration, the most dominant predictor of the RNCC was observed to be the building area. This important observation further supported the use of the most parsimonious regression model, proposed in the form of \( C = K A^B \).
In addition to the development of unprecedented RCE models using the ANN methodology, this dissertation provided three valuable contributions to the literature to assist in better utilization of this methodology. First, it was found that the number of hidden neurons and the values of the learning parameters affect more the generalization ability than the accuracy ability. Second, these abilities were shown to be more sensitive to the number of hidden neurons than to the values of the learning parameters. Third, the severity of the overlearning problem when developing ANN models increased as the number of independent variables increased. Therefore, it was suggested that the overlearning problem should be carefully controlled, particularly when dealing with a greater number of input variables.

Besides the contributions stated above, it was found that an increase in the number of hidden layers may lead to an improvement in the performance of ANN models. In this regard, an emphasis was given to the use of MLP networks with two hidden layers. This finding is in good agreement with results previously reported in the literature (e.g., Attalla and Hegazy, 2003; Surkan and Singleton, 1990; Swales and Yoon, 1992; Yoon et al., 1993). In addition, a number of equations were examined in this research for the purpose of initializing the number of hidden neurons. Of the equations examined, the following equation was found to be of great potential use for this purpose.

\[ \text{Number of hidden neurons} = \frac{m + n}{2} + \sqrt{p} \]  

(9.1)

In the equation above, \(m\), \(n\), and \(p\) stand for the number of input neurons (i.e., the number of independent variables), the number of output neuron (i.e., the number of dependent variables), and the number of samples in the training dataset, respectively.

A final objective of this research derived from recognition that no study has yet been conducted to evaluate the forecasting strength of parametric RCE models against that derived from their competing non-parametric RCE models. An attempt was made in this research to bridge this gap in the literature. In this attempt, MLR models with the contribution of statistically significant independent variables were tailored for the purpose of providing comparisons to their respective ANN models developed using the same sets of independent variables. Accuracy and generalization abilities were again considered as the principles for this comparison. As stated earlier, the former ability was related to the performance of the models when being applied to the development (i.e., training) dataset, while the latter ability
was related to the performance of the models when being applied to the hold-out (i.e., test) dataset. The samples in the development and test datasets remained unaltered during the implementation of this research. The MLR models were compared to their respective ANN models in terms of $R^2$ and mean absolute percentage error (i.e., MAPE).

In the development dataset, the $R^2$ values obtained in both the MLR and the ANN models were monotonically increasing from the first to the last model. The $R^2$ values ranged from 0.59 to 0.71 across the seven MLR models, indicating a 21% increase. This range in the ANN models was from 0.59 to 0.83, indicating a 42% increase, which is clearly twice as large as the improvement observed in the MLR models. The comparison of the highest $R^2$ values within these ranges indicated that the last ANN model outperformed the last MLR model by 17%. A monotonic decrease was also obtained in the values of MAPE, implying that the accuracy ability of the models was continually increasing from the first to the last model. As observed for the $R^2$ values, the rate of this increase was more rapid within the ANN models than within the MLR models. The values of MAPE in the MLR models varied in a range between 25.59% and 20.84%, indicating a 19% improvement. This range in the case of ANN models was moving from 27.48% to 15.42%, indicating a 44% improvement. In addition, the results of this research illustrated that the superiority of the accuracy ability of the ANN models over the accuracy ability of the MLR models was more evident when more independent variables were included in the models. As reported in this dissertation, the last two ANN models advanced the accuracy ability of their corresponding MLR models (as measured in terms of MAPE) by approximately 21% and 26%, respectively. The results of the Mann-Whitney U test proved that these improvements were in fact statistically significant (i.e., $p<0.05$). The increase in observed performance implies that the ANN models encapsulated aspects of the relationships between the variables that their respective MLR models failed to do, suggesting that the ANN methodology is capable of improving upon the regression analysis where nonlinear relationships do exist. It should be noted that although the other ANN models with a reduced number of independent variables were also in general more superior when compared to their respective MLR models, their superiority was proven to be statistically insignificant.

As opposed to the behaviour observed in the development dataset, the $R^2$ values obtained by application of the MLR and the ANN models to the test dataset were fluctuating across these models with no discernible pattern. These values in the former models ranged from 0.56 to 0.69, whilst this range was from 0.59 to 0.73 for the latter models. The lowest
values in these ranges were obtained for the third MLR model and the third ANN model, while the highest values were obtained for the first MLR model and the last ANN model with the inclusion of one and seven independent variables, respectively. The comparison of the highest $R^2$ values showed that the last ANN model excelled by almost 6.5% when compared to the first MLR model. In addition, this ANN model outperformed its respective MLR model, having the same number of independent variables, by almost 11%. Both of these improvements were well less than the improvement obtained when comparing the $R^2$ values of the MLR and the ANN models in the development dataset (i.e., 17%). Similar to the $R^2$ values, the MAPE values in the test dataset varied across the MLR and the ANN models with no fixed pattern. The MAPE values obtained for the MLR models varied in a range from 22.29% to 24.78%, while these values in the ANN models bounced in a wider range from 20.60% to 28.93%. The best models amongst the MLR and the ANN models, resulting in the lowest values of MAPE (i.e., 22.29% and 20.60%, respectively) were the fourth MLR model and the last ANN model, while the latter model outperformed the former model by 7.6%. The last ANN model was also superior to the last MLR model as the MAPE value obtained for the former was almost 11% less than the MAPE value obtained for the latter (i.e., 23.22%). However, this superiority was found to be statistically insignificant. Similarly, in the rest of the models, no statistically significant differences were found between the MAPE values of the ANN models and those obtained for their respective MLR counterparts. This finding suggested that the relative generalization ability of the MLR and ANN modelling techniques was significantly different from what was observed when comparing the accuracy ability of these techniques. One of the possible reasons behind this difference was, at least in part, attributed to the number of data points comprising the development and test datasets. The size of the former dataset which consisted of 121 data points was nearly three times larger than that of the latter dataset consisting of 37 data points. Thus, it is believed that the inclusion of more data in the test dataset would yield significant improvements in the generalization ability that could be achieved by the ANN models.

The salient conclusions from the comparative study of the MLR and the ANN models can be summarised by the following three points. First, the ANN methodology is more sensitive to the number of independent variables than is MLR analysis. Second, the significance of the superiority of the ANN methodology over the MLR analysis is more pronounced in situations where a greater number of independent variables are taken into account. Third, the relative performance of the MLR and ANN modelling techniques is likely
to be strongly dependent on the number of data points. The latter technique is expected to be more statistically desirable when dealing with a larger amount of data.

9.1 Limitations and Directions for Further Research

The topic of RCE modelling is a fairly new research field. Little research has been undertaken in this topic despite its increasing importance and the large investments being directed towards seismic retrofit projects. In addition, developing non-parametric RCE models is a previously unexplored field of study. This research can be considered as the first comprehensive research in the topic of RCE modelling. The implications of this research may lead to new opportunities to be investigated in future studies. The most important extensions to this research are recommended as follows:

1- The methods employed in this research to develop MLR models (i.e., the BE, FS, and stepwise) are arbitrary and restrictive, as these methods can only assess a limited number of regression models. One solution to overcome this limitation is implementation of the “best subset regression” (BSR) method. This method helps determine which predictors have to be incorporated in a MLR model. The BSR method allows for the assessment of all possible regression models that can be generated from all possible combinations of predictors. In the BSR method, all models that have one predictor included are first assessed and the best model is selected based on specified selection criteria such as those utilised in this research. Then, all models that have two predictors included are assessed and the best model is selected as per the selection criteria. This selection procedure continues until all combinations of independent variables are taken into account. As a result, the end outcome of the BSR method is a number of best-fitting models that accommodate one predictor, two predictors, and so on.

2- The ANN models in this research were all developed using the most widely utilized nonlinear activation function, known as the logistic activation function. The impact of using other activation functions (e.g., the hyperbolic tangent activation function, Gaussian activation function, and linear activation function) on the predictive performance of the ANN models is a topic for further investigation.
3- In this research, 75% and 25% of the samples in the research database were randomly singled out to constitute the training and test datasets. Different combinations of training and test samples need to be examined in order to find the combination which leads to the best results, taking the generalization ability into account. This examination also allows for (1) exploration of the sensitivity of the generalization ability to changes in the size of training and test datasets, and (2) adoption of the “voting system” technique in order to improve this ability while eliminating bias. This technique involves the creation of a number of models each using different training, and test datasets with the output being taken as the average of these models (Lowe et al., 2006).

4- The ANN models developed in this research were all trained by the back-propagation learning algorithm, which is the most popular and most utilized type of supervised learning algorithm. Despite its capabilities, this learning algorithm suffers from a number of problems due to the lack of adequate rules for the development of an optimal ANN model. In this research, the optimal ANN models were developed through the design and implementation of a novel heuristic process which, although being successful, required considerable time and effort to implement. Therefore, further research may be warranted to investigate the suitability of other alternative learning algorithms, particularly the genetic algorithm (GA). This algorithm is an artificially intelligent optimization technique that can be utilized to optimize network parameters including the number of layers and their respective number of neurons, together with the values of the network weights and biases (Hegazy et al., 1994). Using the GA technique, an ANN model is optimized with an objective function to minimize the errors in the test dataset through the use of a trial and error approach (Hegazy et al., 1994). According to Kim et al., (2005), the advantages of this optimization are as follows: (1) the difficulty in determining the number of neurons in the hidden layer is resolved; (2) it is not necessary to spend time and effort in determining the parameters of the back-propagation algorithm by trial and error; and (3) whereas trial and error uses integers, GAs use real numbers in optimization, thereby minimizing the possibility of trapping local minima and improving the accuracy of estimation.
5- It may be of future interest to develop an advanced hybrid neuro-fuzzy system for the purpose of estimating the retrofit construction cost. The neuro-fuzzy system may be treated as an effective computational tool for this purpose as it combines the approximate reasoning of the fuzzy logic with the learning ability of the ANN methodology. A further attempt can be also made to develop a hybrid RCE model that utilizes a regression model output as an input to an ANN mode. Such hybrid models were utilized by Williams (2002) to predict the completed cost of competitively bid highway projects and were found to produce reasonable predictions.

6- Due to data availability challenges, the scope of this research was restricted to the development of parametric and non-parametric RCE models for earthquake-prone public schools having a framed structure (i.e., concrete and steel structure), using the MLR and the ANN methodology, respectively. Although the predictive abilities of these methodologies may differ from one type of retrofit project to another, the results of this research offered that the further application of these methodologies to estimate the retrofit construction cost of other earthquake-prone buildings with different structural types (e.g., unreinforced masonry structures) and different functionalities (e.g., hospitals) is highly warranted.

7- This research used data that were collected in Iran and, as a result, the RCE models developed may not be directly applicable to retrofit projects in other countries due to different retrofit practices. However, these countries may benefit from the results of this research. For instance, the identified double log cost-area formula (i.e., $C=KA^B$) provides an important benchmark for further research to verify whether the utilization of this formula can be extended to estimate the construction cost of retrofit projects undertaken in other countries. In addition, the results of this study may be used to facilitate comparison of the retrofit construction cost at an international level.

8- This research made use of the retrofit construction cost data that were extracted from the tender documents prepared by the engineering consultants after completion of the retrofit design phase. These data did not reflect the final construction cost to the client. As described in this dissertation, the final cost of
complex projects, such as retrofit projects, is subject to inevitable variations caused by different factors such as additions to the original work or modifications to the original work due to design changes. Thus, there is an authentic need to initiate additional research in order to develop RCE models that can provide predictions of the magnitude of the final retrofit construction cost.

9- Finally, the findings of the key studies by Skitmore and Cheung (2007) and by Cheung et al., (2008) raise two important implications for future studies, being (1) the specification of an appropriate loss function to compensate for the potential bias induced by OLS RCE models, and (2) the incorporation of this function into the development and assessment of more realistic non-OLS RCE models.
REFERENCES


Decision-Makers, Prepared by EQE International, Inc. for the California Seismic Safety Commission, Oakland, California.


Surkan, A. J., & Singleton, J. C. (1990). *Neural Networks for Bond Rating Improved by Multiple Hidden Layers*. Paper presented at the International Joint Conference on Neural Networks (IJCNN), San Diego, California, USA.


APPENDIX A

WEB SURVEY QUESTIONNAIRE

(WWW.RETROFITCOST.NET)
Figure A-1: Home page of designed web site (www.retrofitcost.net)
STEP ONE
(Login and contributor information)

**Contributor type**

**Name**

**Address**

**Email**

**Contact number**

**User name**

**Password**

**Retype password**

STEP TWO
(General information of the retrofitted building)

**City where the building is located**

**Heritage building control**

**Year of original design and construction**

**Occupancy and social classification**

**Building importance level**

**Building total area**

**Building total height**

**Building model type**

STEP THREE
(Seismic assessment and retrofit information)

**Seismic load level considered in seismic assessment**

**Assessed ductility of existing structure**

**Site soil type**

**Structural retrofit method(s)**

Figure A-2: Fundamental questionnaire (essential information for cost modelling)
STEP FOUR
(Construction cost information)

Base year for cost

Construction cost details

Construction cost compared to “Building Replacement Cost”

Duration of retrofit construction work

STEP FIVE
(Complementary cost information)

Non-construction cost details

Indirect cost details

Condition of occupancy during construction work

Cost reference

Figure A-2: Fundamental questionnaire (essential information for cost modelling)

(Continued)
**STEP ONE**
(General information of the retrofitted building)

- **Owner type**
- **Maintenance condition**
- **Alteration before retrofit construction work**
- **Availability of building “as-built” documents**

**STEP TWO**
(Structural and geotechnical information)

- **Foundation and geologic site hazard**
- **Structural deficiencies before retrofit**

**STEP THREE**
(Complementary information and documents)

- **Code or design guideline used for retrofit design**
- **Fundamental period of vibration**
- **Photos and sketch of building**

Figure A-3: Supplemental questionnaire (complementary information for cost modelling)