



저작자표시-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

Doctor of Philosophy

**FRONT-END COST ESTIMATION BY
SELECTIVE CASE-BASED REASONING FOR
BUILDING CONSTRUCTION PROJECTS**

February 2016

Department of Architecture & Architectural Engineering
The Graduate School
Seoul National University

Joseph Ahn

**FRONT-END COST ESTIMATION BY
SELECTIVE CASE-BASED REASONING FOR
BUILDING CONSTRUCTION PROJECTS**

**A dissertation submitted to the Graduate School of
Seoul National University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy**

by

Joseph Ahn

December 2015

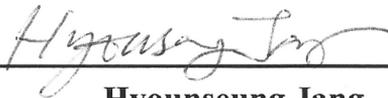
Approval Signatures of Dissertation Committee



Moonseo Park



Bo-Sik Son



Hyounseung Jang



Seokho Chi



Hyun-Soo Lee

Abstract

Front-End Cost Estimation by Selective Case-Based Reasoning for Building Construction Projects

Joseph Ahn

Department of Architecture & Architectural Engineering

The Graduate School

Seoul National University

A successful building construction project can be achieved by estimating construction cost with high level of accuracy, which is particularly crucial in the front-end stage due to the influence on cost reduction and effective cost management. However, since there are problems regarding inaccurate budgeting for building construction projects, limited information availability, limited usage of unit price of actual construction cost, and lack of flexibility of cost estimation models for diverse building project, owners and cost estimators need to establish an effective cost estimation countermeasures.

To deal with the aforementioned problems, this dissertation aims to develop a front-end cost estimation methodology by selective case-based

reasoning (CBR) for building construction projects for improving 1) cost estimation accuracy, 2) reliability of human trust on estimated costs, and 3) transparency of cost estimation process.

More specifically, this research has objectives of developing three modules that are 1) *Module 1: Case-Base Development*, 2) *Module 2: CBR Method Selection*, and 3) *Module 3: CBR Cost Estimation*. The *Module 2* again comprises 1) *Sub-Module 1: Normalization Method Selection* (interval, Gaussian distribution-based, Z-score, logistic function-based, and ratio normalizations), 2) *Sub-Module 2: Attribute Weighting Method Selection* (Attribute Impact, entropy, feature counting, and genetic algorithms), and 3) *Sub-Module 3: Similarity Measurement Method Selection* (Mahalanobis distance-based, Euclidean distance-based, arithmetic summation-based, fractional function-based).

The proposed front-end cost estimation methodology by selective case-based reasoning was validated using leave-one-out cross validation method for multi-family housing (100 cases), military barrack (117 cases), and government office (52 cases) projects. Accuracy (under mean absolute error rate, mean squared deviation, and mean absolute deviation), stability (under standard deviation), and appropriateness (using kernel density estimation) of cost estimation results were examined. More importantly, the level of flexibility of the selective CBR model which provides the most accurate and stable normalization, attribute weighting, and similarity measurement method according to different types of building projects was tested.

The results of case studies for the validation of the proposed methodology are summarized as below: For the multi-family housing project, ratio normalization method, GA attribute weighting method, and arithmetic summation similarity measurement method-based CBR cost model was proposed to be the most accurate and stable. For the military barrack project, interval/ratio normalization method, GA attribute weighting method, and Euclidean distance similarity measurement method-based CBR cost model was suggested to be the most accurate and stable. For the government office project, ratio normalization method, AI attribute weighting method, and fractional function similarity measurement method-based CBR cost model was derived to be the most accurate and stable.

As contributions of the research, the suggested data preprocessed case-base development procedures are expected to improve transparency and reliability of cost estimate results. Also, this research performed validations of the improved estimate accuracy and explanatory power of the selective CBR models for different characteristics of case-bases. Consequently, accurate front-end cost estimations with enhanced flexibility responding to various building construction projects are expected.

Keyword: Front-End Cost Estimation, Building Construction Project, Selective Case-Based Reasoning, Data Preprocessing, Normalization, Attribute Weighting, Similarity Measurement

Student Number: 2011-30176

TABLE of CONTENTS

Chapter 1. Introduction	1
1.1 Research Background.....	1
1.2 Problem Statement	3
1.3 Research Objectives, Methodology and Scope.....	8
Chapter 2. Issues in Front-End Cost Estimation	15
2.1 Overview of Cost Planning	16
2.1.1 Role of Front-End Cost Estimation in Cost Planning.....	16
2.1.2 Importance of Front-End Cost Estimation in Cost Planning...	18
2.2 Requirements for Front-End Cost Estimation.....	21
2.2.1 Cost Estimate Accuracy.....	21
2.2.2 Reliability of Human Trust.....	23
2.2.3 Transparency of Estimation Process.....	25
2.3 Current Practice Reviews on Front-End Cost Estimation....	27
2.3.1 Inaccurate Budgeting for Building Construction Projects.....	27
2.3.2 Limited Information Availability.....	30
2.3.3 Limited Usage of Unit Price of Actual Construction Cost	33
2.3.4 Lack of Flexibility for Diverse Building Projects	35
2.4 Literature Reviews on Front-End Cost Estimation	39
2.5 Summary	42
Chapter 3. Case-Based Reasoning Approach	45
3.1 Principles of CBR	46
3.1.1 Overview of CBR.....	46

3.1.2	CBR Problem-Solving Process	47
3.1.3	CBR and Rule-Based Reasoning.....	50
3.2	CBR Advantages, Limitations, and Issues	52
3.2.1	Advantages of CBR.....	52
3.2.2	Limitations of CBR	55
3.2.3	Challenging Issues in CBR.....	56
3.3	CBR Model Components	60
3.3.1	Normalization for Case Representation.....	60
3.3.2	Attribute Weighting for Case Indexing.....	65
3.3.3	Similarity Measurement for Case Retrieval	68
3.4	Summary	71

Chapter 4. CBR Model Design Experiment for Improving

Cost Estimation 73

4.1	Normalization Method and Accuracy	73
4.1.1	Normalization Issue.....	73
4.1.2	Comparative Experimental Design.....	74
4.1.3	Results and Discussions	76
4.2	Attribute Weighting Method and Accuracy	80
4.2.1	Attribute Weighting Issue.....	80
4.2.2	Concept of Attribute Impact.....	82
4.2.3	Comparative Experimental Design.....	90
4.2.4	Results and Discussions	98
4.3	Similarity Measurement Method and Accuracy.....	102
4.3.1	Covariance Effect Issue.....	102
4.3.2	Comparative Experimental Design.....	103

4.3.3	Simulation Data Test	106
4.3.4	Applicability Test	112
4.4	Summary	116

Chapter 5. Cost Estimation Methodology by Selective Case-Based Reasoning119

5.1	Overview of Methodology Development	119
5.2	Case-Base Development	124
5.3	CBR Method Selection and Cost Estimation.....	134
5.3.1	Sub-Module 1: Normalization Method Selection.....	134
5.3.2	Sub-Module 2: Attribute Weighting Method Selection.....	141
5.3.3	Sub-Module 3: Similarity Measurement Method Selection ..	147

Chapter 6. Case Studies 155

6.1	Validation Methods and Process	155
6.2	Multi-Family Housing	161
6.2.1	Case-base Profile.....	161
6.2.2	Results and Discussions	167
6.3	Military Barrack.....	180
6.3.1	Case-base Profile.....	180
6.3.2	Results and Discussions	183
6.4	Government Office	195
6.4.1	Case-base Profile.....	195
6.4.2	Results and Discussions	198
6.5	Summary	211

Chapter 7. Conclusions	219
7.1 Research Results	220
7.2 Research Contributions	223
7.3 Limitations and Future Research	225
Bibliography	227
Appendix	245
Abstract (Korean).....	281

List of Tables

Table 3-1	Literature Reviews on Normalization in CBR Cost Estimation...63
Table 4-1	Results of MAER for Normalization Method77
Table 4-2	Results of MSD for Normalization Method77
Table 4-3	Results of MAD for Normalization Method.....78
Table 4-4	Results of SD for Normalization Method.....79
Table 4-5	Correlation Analysis (Pearson Correlation Coefficient).....86
Table 4-6	Derived Value of Attribute Impact & Correlation Coefficient.....88
Table 4-7	Ranking Comparison between AI & CC89
Table 4-8	Profile of Test Cases93
Table 4-9	Parametric Equations Using Multiple Regression Analysis ..94
Table 4-10	Correlation Analysis (Type 84).....95
Table 4-11	Weights of the Attributes Obtained by AI, RC, FC, and GA.97
Table 4-12	Comparison of Absolute Error Ratio (AER)99
Table 4-13	CBR Applicability Test.....101
Table 4-14	MAER Comparison for Simulation Test 1109
Table 4-15	MAER Comparison for Simulation Test 2110
Table 4-16	MAER Comparison for Simulation Test 3111
Table 4-17	MAER and SD Comparison for Case Study (MFH)113
Table 5-1	Design of Case-Base Structure125
Table 6-1	Validation Methods and Process for Case Studies.....156
Table 6-2	Attributes for Multi-Family Housing163
Table 6-3	Attribute Weights by AI, Entropy, FC, and GA (MFH).....164
Table 6-4	MAER for Normalization Methods (MFH).....167

Table 6-5	MSD for Normalization Methods (MFH)	168
Table 6-6	MAD for Normalization Methods (MFH).....	169
Table 6-7	SD for Normalization Methods (MFH).....	170
Table 6-8	MAER for Attribute Weighting Method (MFH).....	173
Table 6-9	MSD for Attribute Weighting Method (MFH)	174
Table 6-10	MAD for Attribute Weighting Method (MFH).....	174
Table 6-11	SD for Attribute Weighting Method (MFH).....	175
Table 6-12	MAER for Similarity Measurement Methods (MFH).....	177
Table 6-13	MSD for Similarity Measurement Methods (MFH).....	177
Table 6-14	MAD for Similarity Measurement Methods (MFH)	178
Table 6-15	SD for Similarity Measurement Methods (MFH)	178
Table 6-16	Attributes for Military Barrack.....	180
Table 6-17	Attribute Weights by AI, Entropy, FC, and GA (MB)	181
Table 6-18	MAER for Normalization Methods (MB)	183
Table 6-19	MSD for Normalization Methods (MB).....	184
Table 6-20	MAD for Normalization Methods (MB)	184
Table 6-21	SD for Normalization Methods (MB)	185
Table 6-22	MAER for Attribute Weighting Method (MB).....	189
Table 6-23	MSD for Attribute Weighting Method (MB).....	190
Table 6-24	MAD for Attribute Weighting Method (MB)	190
Table 6-25	SD for Attribute Weighting Method (MB)	191
Table 6-26	MAER for Similarity Measurement Methods (MB)	192
Table 6-27	MSD for Similarity Measurement Methods (MB)	193
Table 6-28	MAD for Similarity Measurement Methods (MB).....	193
Table 6-29	SD for Similarity Measurement Methods (MB).....	194
Table 6-30	Attributes for Government Office	195

Table 6-31	Attribute Weights by AI, Entropy, FC, and GA (GO).....	196
Table 6-32	MAER for Normalization Methods (GO)	198
Table 6-33	MSD for Normalization Methods (GO)	199
Table 6-34	MAD for Normalization Methods (GO).....	200
Table 6-35	SD for Normalization Methods (GO).....	200
Table 6-36	MAER for Attribute Weighting Method (GO)	204
Table 6-37	MSD for Attribute Weighting Method (GO)	205
Table 6-38	MAD for Attribute Weighting Method (GO).....	205
Table 6-39	SD for Attribute Weighting Method (GO).....	206
Table 6-40	MAER for Similarity Measurement Methods (GO).....	208
Table 6-41	MSD for Similarity Measurement Methods (GO).....	208
Table 6-42	MAD for Similarity Measurement Methods (GO)	209
Table 6-43	SD for Similarity Measurement Methods (GO)	210
Table 6-44	Summary of Norm. Method Selection (MFH)	212
Table 6-45	Summary of AW Method Selection (MFH).....	213
Table 6-46	Summary of SM Method Selection (MFH).....	213
Table 6-47	Summary of Norm. Method Selection (MB).....	214
Table 6-48	Summary of AW Method Selection (MB)	215
Table 6-49	Summary of SM Method Selection (MB)	216
Table 6-50	Summary of Norm. Method Selection (GO)	217
Table 6-51	Summary of AW Method Selection (GO).....	217
Table 6-52	Summary of SM Method Selection (GO).....	218

List of Figures

Figure 1-1	Methodological Approach and Cost Estimation Aspects.....	9
Figure 1-2	Definition of Front-End/Conceptual Stage.....	12
Figure 1-3	Dissertation Outline.....	13
Figure 2-1	Cost Reduction Potential through the Project Life Cycle.....	19
Figure 3-1	CBR Problem-Solving Process.....	48
Figure 3-2	Task-Method Decomposition of CBR.....	49
Figure 4-1	Experimental Design for Normalization Method.....	75
Figure 4-2	Results of MAER, MSD, MAD, and SD for Normalization.....	76
Figure 4-3	Experimental Design for Attribute Weighting Methods.....	90
Figure 4-4	Experimental Design for Similarity Measurement Method.....	104
Figure 4-5	MAER Comparison for Simulation Test 1.....	109
Figure 4-6	MAER Comparison for Simulation Test 2.....	110
Figure 4-7	MAER Comparison for Simulation Test 3.....	111
Figure 4-8	MAER and SD Comparison for Case Study (MFH).....	114
Figure 5-1	Framework of Front-End Cost Estimation by Selective CBR.....	122
Figure 5-2	Mechanism of Front-End Cost Estimation by Selective CBR.....	123
Figure 5-3	Attribute Extraction Process.....	126
Figure 5-4	Procedures of Acquiring Preprocessed Data.....	129
Figure 5-5	Sub-Module 1: Normalization Method Selection.....	136
Figure 5-6	Sub-Module 2: Attribute Weighting Method Selection.....	142
Figure 5-7	Sub-Module 3: Similarity Measurement Method Selection.....	148
Figure 5-8	Comparison of Mahalanobis Distance and Euclidean Distance.....	151
Figure 6-1	Matrix Plot (MFH).....	165

Figure 6-2	MAER, MSD, MAD, and SD for Norm. Methods (MFH)..	171
Figure 6-3	Kernel Density Estimation for Normalized Case-Bases (MFH)	172
Figure 6-4	MAER, MSD, MAD, and SD for AW Method (MFH)	176
Figure 6-5	MAER, MSD, MAD, and SD for SM Methods (MFH)	179
Figure 6-6	Matrix Plot (MB).....	182
Figure 6-7	MAER, MSD, MAD, and SD for Norm. Methods (MB)	186
Figure 6-8	Kernel Density Estimation for Normalized Case-Bases (MB)...	188
Figure 6-9	MAER, MSD, MAD, and SD for AW Method (MB).....	191
Figure 6-10	MAER, MSD, MAD, and SD for SM Methods (MB).....	194
Figure 6-11	Matrix Plot (GO)	197
Figure 6-12	MAER, MSD, MAD, and SD for Norm. Methods (GO)	201
Figure 6-13	Kernel Density Estimation for Normalized Case-Bases (GO) ...	203
Figure 6-14	MAER, MSD, MAD, and SD for AW Method (GO)	207
Figure 6-15	MAER, MSD, MAD, and SD for SM Methods (GO).....	210

Chapter 1. Introduction

1.1 Research Background

Cost management in a construction project summarizes all the financial activities in an organization that are aimed at ensuring the cost of a given building is maintained within the period agreed upon in the construction contract. This process is a managerial process that gathers information about the construction of a building so as to support the decision making process. The process in addition is aimed at improving the value as well as stimulate reduction of costs in an organization. In the process, management of costs in an organization is aimed at ensuring stability in funds at the designing stages of the construction and after the project has started. It thus covers the myriad areas which composed of estimating, planning and controlling processes of costs in an organization (Kim 2005; Kirkham 2014). Cost management requires innovation, creativity, learning, involvement and commitment of people so as to be implemented.

More specifically, cost planning is a process that deals with refining costs that were made before so as to estimate and provide a cash flow of the project as well as addition of more information so as to cater for the construction of the project as argued by Towey (2013) and Gerrard (2000). Cost planning is aimed at accommodating all the expenses the project is likely to encounter during the period of construction including the production plans and those that are

recurrent to the project. Cost planning is very important as it ensures the accomplishment of the business goals is achieved due to the support the plan gives on decision making process.

Construction projects have always met a problem of insufficient funds to facilitate the payment of workers, suppliers among other sectors that determines the completion of the project. All these problem originates from making inaccurate estimations of costs for the construction project. During the process, the cost estimation might have been made by making predictions about the possible costs of the project prematurely. This is termed premature when the estimation of costs for a project is made without consultation of the main sectors of the construction project (Blocher et al. 2012).

In order for this to be done, cost estimation need to be arrived upon after consultation of the qualified contractors, risk estimation for the entire period of costs among other factors that will lead to a arrive to an effective cost for the project. Having an accurate cost estimation is very crucial as it determines the overall success of a construction project. On the other side, under estimation of costs leads to many problems in the completion of the construction project. The project may thus delay to start to in construction or begin and fail to complete hence unable to meet the business goals.

1.2 Problem Statement

A successful one-off construction project can be achieved by estimating construction cost with high level of accuracy, which is particularly crucial in the conceptual stage due to the influence on cost reduction (Doğan et al. 2006; Sevgi et al. 2008; Ji et al. 2012). Since a considerable amount of building cost is required with the increased expectations for higher quality with lower or stable budgeting, construction costs need to be accurately estimated and managed as one of the key decision-making elements (Kim 2005; Kirkham 2014). However, since there are problems regarding inaccurate budgeting for building construction projects, limited information availability, limited usage of unit price of actual construction cost, and lack of flexibility of cost estimation models for diverse building project, owners and cost estimators need to establish an effective cost estimation counter measure.

Inaccurate Budgeting for Building Construction Projects

Public institutions recognize the importance of cost management from the planning stage; however, in most cases, they do not have an organized construction cost estimation and management system. Thus, at the stage of planning a new public construction project, those in charge of budgeting should estimate construction cost based on existing data and experiences, compare estimated construction cost with budget after the detailed design and the documentation stage, and then decide whether to continue the project or change the design according to the budgets.

Moreover, social organizations have been raising the construction cost bubble theory; hence, there are increasing concerns over what would be the appropriate building construction cost. However, when such issues are approached simply through the logic of budget saving, it may cause another social problems such as the destruction of fair competition through dumping. Also, the current situation needs to be considered that high-quality construction is hardly to be expected because of prevalent abnormally low bidding in the current construction market due to oversupply.

In overall, inaccurate budgeting brings critical problems in constructing the project to completion as it brings management loop holes. Consequentially, this inaccurate budgeting causes many issues that affects both the quality and success of building construction projects.

Limited Information Availability

Cost estimation in the conceptual stage requires a large number of consultation so as to equip the planners with information on how to formulate an accurate cost estimation. Since information is limited during this stage, an organization or individual may end up making an inaccurate cost for building construction projects hence transferring the problem to the construction process (Robinson et al. 2015).

Limited information during cost estimation process makes the decision making process difficult due to less participation of participant that aids in cost

evaluation. When this happens, the personnel responsible in formulating an accurate cost for the project is deprived enough. For this case, cost estimation is not reached upon under evaluation of economic, political and physical factors that determines the cost of a construction project (Robinson et al. 2015).

Importantly, the process of decision-making is information based hence limited information may lead to making poor decisions about the project. In this case, this may hinder the organization from obtaining high level of cost estimate accuracy. With only limited information available in the early design phases, public owners require cost estimation with accuracy, reliability of human trust, and transparency of estimation process. Thus owners and cost estimators require highly effective strategies.

Limited Usage of Unit Price of Actual Construction Cost

In order to estimate an accurate cost for construction projects, standards containing unit price of construction materials, labors, equipment, and etc. are used, mostly in detailed design or documentation stages. When this values of costs is used, then the final estimation of costs of construction for the project does not have a greater deviation from what the project will need in actual sense. In some instances, the cost estimated based on unit prices are usually slightly higher hence facilitating the completion of the construction project.

When the unit price of actual construction is used on limited scale, then there is a possibility of the estimated cost running below what was expected.

According to Gerrard (2000) and Smith (1995), this will lead to inaccurate estimation of costs resulting to a problem in determining the budget for the construction project.

Inadequate usage of the unit price of actual construction cost creates a vacancy in determining the success of the project. For instance, when unit price of actual construction cost is used on limited basis, then cost of the project may be overestimated leading to an exorbitant cost. The result may lead to the project not been approved for construction especially if the construction project belongs to a company or the government. On the other hand, the cost estimated may be too little ending up being approved easily then becoming difficult to complete the project.

More importantly, most of cost management related regulations and standards are focused on or after documentation stage. Especially, unit price for actual construction cost and standard of construction estimate can be utilized after detailed design or documentation stages. Therefore, it is necessary to estimate accurate cost with reliable data resource that can support decision-makers in early design stages.

Lack of Flexibility for Diverse Building Projects

With the trend of the complication, diversification and enlargement of the construction industry, the construction industry needs to estimate an accurate construction cost in the conceptual/planning stages of projects and to manage

cost effectively and consistently. Construction costs estimation should be dynamic and prone to accommodate various types of buildings. When this happens, then the public institutions or companies are likely to successfully achieve its goals and aims.

However, building construction projects have characteristics of dynamic and uniqueness. A fixed cost model has limitations in dealing with various construction projects and their database characteristics. Hence, it requires selective and flexible approach; and a cost model need to be designed accordingly considering the unique and dynamic characteristics of various building construction projects.

Otherwise, several other cost models need to be developed and utilized so that accurate cost estimation can be achieved. However, developing and utilizing several different models may require significant amount of time and effort to deal with various building types; and ineffective utilization and low flexibility of these models to deal with various building projects are expected. When flexibility of cost estimation model for various building projects is not achieved, then the cost estimate results are likely to be very low; hence construction projects are likely to be faced by a number of problems that will drag the rate for achieving its successful completion.

1.3 Research Objectives, Methodology and Scope

Every building construction project has its own unique characteristics, and these can relate to building type, scope, quality, cost, and/or duration. Contractors and consultants must meet all client requirements satisfactorily while, significantly, meeting increasing demands for higher quality, shorter duration, and lower costs. In general, a considerable amount of cost is incurred to accomplish a one-off construction project successfully. As asserted by Kim (2005), those involved should regard entire or partial construction costs as one of the key decision-making elements. The success of a construction project can be evaluated by how accurately the project cost is estimated at initial/conceptual phase and managed throughout the life cycle (Kim 2005; Schuette and Liska 1994).

However, as elaborated in *Chapter 1.2. Problem Statement*, there are four main issues that need to be addressed to improve conceptual cost estimation:

- Inaccurate Budgeting for Building Construction Projects
- Limited Information Availability
- Limited Usage of Unit Price of Actual Construction Cost
- Lack of Flexibility of a Cost Model for Diverse Building Projects

As an effort to enhance the estimate outputs, plenty of efforts have been made using various methods such as regression analysis, artificial neural network (ANN), case-based reasoning (CBR), and so on. Recently, CBR

methodology has been progressively used for construction cost estimation because of its benefits (Doğan et al. 2006; An et al. 2007; Sevgi et al. 2008; Kim and Kim 2010; Koo et al. 2011; Ji et al. 2012; Jin et al. 2012; Ahn et al. 2014).

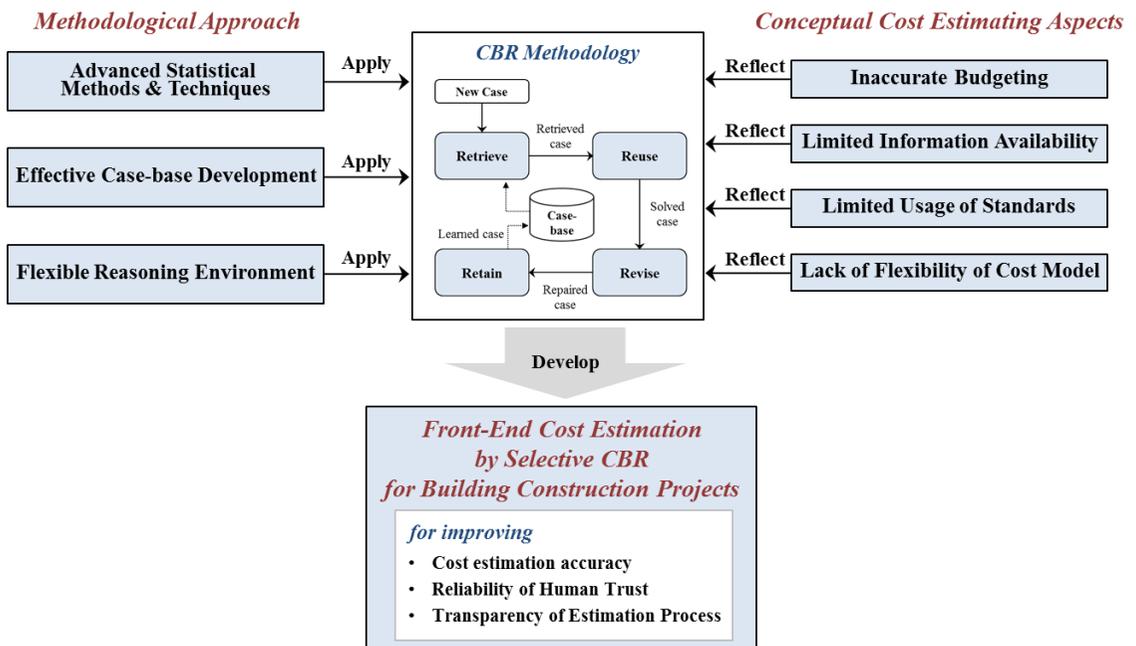


Figure 1-1. Methodological Approach and Cost Estimation Aspects

Importantly, since CBR methodology highly relies on past experience or data, a quality case-base needs to be prepared to yield quality cost estimation results (Kotsiantis et al. 2006). Essentially, historical data is the foundation of accurate cost estimation because historical data provides credibility, accuracy, and defensibility (GAO 2004; ISPA 2008). Moreover, Stiff and Mongeau (2002)

observed that a more familiar and reliable source will attract more comprehension and persuasion from a specific group of people. For this reason, even owners and cost estimators who do not have much experience and knowledge can build effective strategies and be more persuasive by using CBR cost estimation. Also, to improve explanatory power of CBR cost estimation, CBR model needs to be designed for flexible reasoning environment and applies advanced statistical methods and techniques.

To deal with the aforementioned problems descriptions, this dissertation aims to develop a front-end cost estimation methodology by selective case-based reasoning for building construction projects for improving 1) cost estimation accuracy, 2) reliability of human trust, and 3) transparency of cost estimation process.

More specifically, to propose the cost estimation methodology (in Chapter 5), this research has objectives of developing three modules that are 1) *Module 1: Case-Base Development*, 2) *Module 2: Method Selection*, and 3) *Module 3: CBR Cost Estimation*. The *Module 2* again comprises 1) *Sub-Module 1: Normalization Method Selection*, 2) *Sub-Module 2: Attribute Weighting Method Selection*, and 3) *Sub-Module 3: Similarity Measurement Method Selection*.

To obtain verification of the applied CBR model components, mechanisms, and validation procedures of the suggested cost estimation methodology, this research also aims to perform CBR model design experiments (in Chapter 4).

The proposed cost estimation methodology is targeted to be used in front-end or conceptual stage as defined in Figure 1-2. For the validation, case studies on multi-family housing (100 cases), military barrack (117 cases), and government office (52 cases) projects are carried out.

To logically conduct the overall research, a research process of the dissertation is outlined in Figure 1-3.

	Scheming	Planning	Preliminary Design	Detail Design	Bid/Contract	Construction	Repair/Maintenance
Ferry & Brandon (1991)	Single Price Method		Elemental			Operational	
Adrain (1993)	Feasibility	A/E Approximate		Detail		Operational	Final
Smith (1995)	Preliminary	Appraisal	Proposal		Approved	Pre-tender Post-contract	Achieved
AACE (1995)	Order of Magnitude		Budget	Definitive			
Gould (1997)	Conceptual	Schematic	Design develop				
Son et al. (2013)	Conceptual		Schematic	Detail	Definitive		
This Research	Front-End/Conceptual		Schematic	Detailed	Documentation		

Figure 1-2. Definition of Front-End/Conceptual Stage (revised from Son et al. 2013)

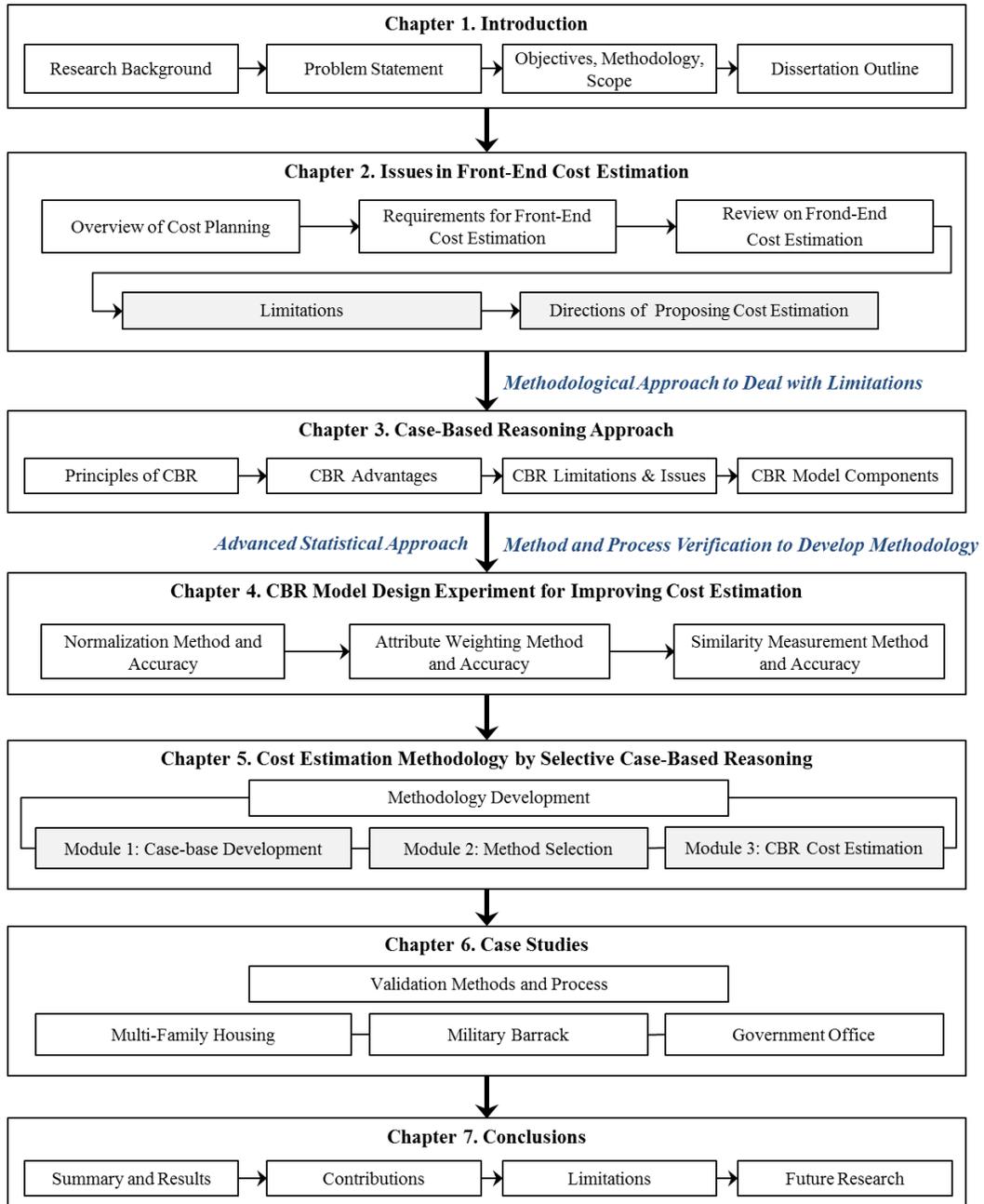


Figure 1-3. Dissertation Outline

Chapter 2. Issues in Front-End Cost Estimation

Front-end cost estimation refers to the cost estimation done in the conceptual stage before the start of a project. At the beginning of a project, the owner only has limited information concerning it and the design. However, the owner of the project is supposed to have knowledge of the approximate ways of evaluating the economic feasibility of proceeding with a project. This is where the need to approximate the likely costs of a project at the conceptual stage arises.

Front-end cost estimation is thus the initial efforts taken to evaluate the cost required for the development and adoption of a project. It is usually carried out by the estimators as part of the project feasibility analysis at the beginning of a project thus making the estimates be developed using limited information on the scope of the project and without a detailed design and engineering data. Front-end cost estimation is very important in the construction and industrial projects. This is because these estimates obtained are used as a basis for decision-making by both the clients and the project engineers. Accuracy and reliability are core requirements in the conceptual cost planning, and they should be considered.

2.1 Overview of Cost Planning

2.1.1 Role of Front-End Cost Estimation in Cost Planning

Front-end/Conceptual cost estimation plays a very important role in the decision-making during the initial stages cost planning of a project (McCarthy and McCarthy 2011). This is because the estimates made in the front-end stage are the ones used to make the rough estimate of the costs that will be undergone during the adoption of a project as well as the initializing and completion. Cost estimates in the early stages have a great influence on the decisions made by the expectations and the subsequent conduct to avoid uncertainties, inaccuracy and inefficient project cost planning and control.

The second role of front-end cost estimation in cost planning is that it usually support evaluations made on the project funding requirements during planning. This is because the cost estimates that were made in the conceptual stage were involved in the formulation of a budget that acted as the cost constraint for a project. The budget formed would show a financial analysis of how the funds accounted for in the estimates would be used. This helped in saving on costs, time and avoiding the wastage of resources in a company. The budget made helped to ensure full utilization of the resources in the cost planning such that everything planned in the project could be fully accounted.

The third role of the front-end cost estimates in cost planning is that it was used to ensure accuracy in the cost planning. This was done by evaluating the necessary funding that was required and comparing it with the bids and the tenders available for procurement of all the requirements for a project. Accuracy was ensured by comparing the estimates made with the actual cost that is likely to be incurred by choosing the most cost-effective and favorable tender. In construction contracts, the cost estimate was usually made to present a tender and ensure completion of the awarded contract. On the other hand in the activities such as maintenance and operations that revolved around the establishment of the source of funding as well as the budget allocation.

The other role of front-end cost estimation in cost planning is to help contingency element of any estimate made meet the required likelihood and impact of the uncertainties. The estimation plays a very great role in helping the planners to identify and measure the specific risks that are likely to be come across in the project, define the unmeasured uncertainties in the estimate and also have a clear knowledge of the unknown uncertainties that are not know or even easily understood in the estimates as per a given time. If the project is self-contained, contingency is most likely to be experienced in both the specifically measured uncertainties risks as well as the defined, but unmeasured uncertainties incurred around the estimates made. In large infrastructure projects, a large number of risks are encountered in the unknown uncertainties that are not known or even understood at a given specific time.

2.1.2 Importance of Front-End Cost Estimation in Cost Planning

One of the importance's of front-end cost estimation in cost planning is that it assists in the overall cost control program by acting as the immediate check against the budget. This is done by increasing the cost overruns in advance before the project team reviews the cost and design for the other possible alternatives. Since the early cost estimates are made before the completion the detailed design, the error margin is presupposed to be very high, and it thus helps the people involved in cost planning to take precautions to reduce the error and avoid duplicating it. This is done by assisting them to learn the importance of applying the contingency that varies with the amount of data available and the amount of information that can be easily obtained from similar alternative projects.

The second importance of the front-end cost estimation is that it is very useful in dealing with the numerous uncertainties encountered during the process of cost planning of the project (Kesavan et al. 2008). Cost estimators are usually required to have a lot of knowledge and expertise that is very necessary for reducing the risks resulting from uncertainties to an acceptable level in the conceptual estimates during the time planning is being done. Cost estimation helps to deal with specific, measurable uncertainties, defined immeasurable uncertainties as well as unknown uncertainties that take place at a time that cannot be understood. This helps in making the process of cost planning accurate and reliable.

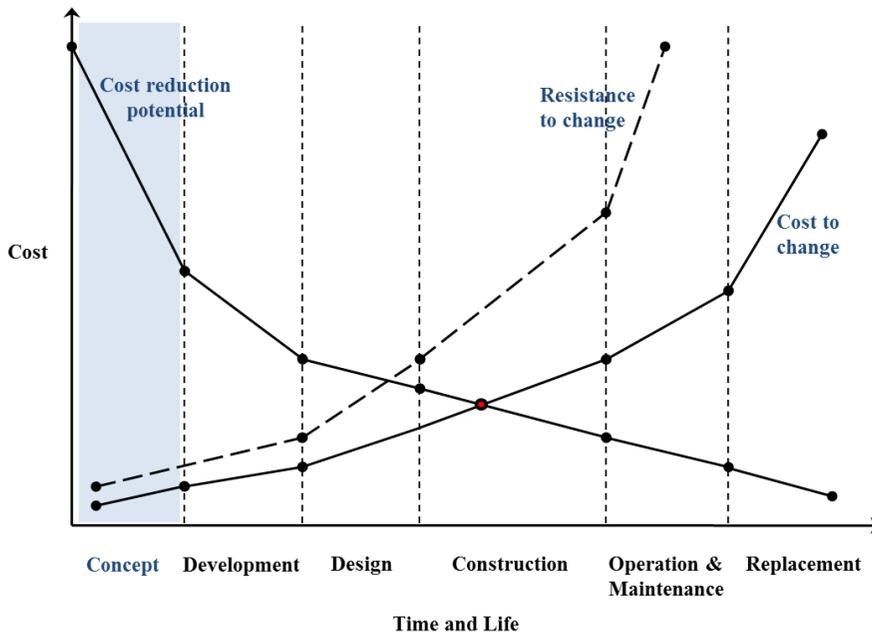


Figure 2-1. Cost Reduction Potential through the Project Life Cycle (Kirkham 2014)

The third importance of front-end cost estimation in the cost planning is that it helps to know the reliability of the estimates made before the cost planning decisions is made. In the construction projects, clients are also willing to know the conceptual cost estimates well as their extent of reliability. The clients thus use the information on the cost estimates to gauge the reliability of a project for them to consider which is best for them to undertake after comparing with the estimates of other alternative projects. The clients consider the results of a conceptual cost estimates reliable if the quality of the early cost estimates is presumed to be high. Quality is shown in the conceptual cost estimate if the estimated cost has the expected accuracy range.

The other importance of the front-end cost estimates in the cost planning is that it provides an overview of the credible cost that is likely to be incurred in the construction at the early stages of construction projects. This assists the people involved in the process of cost planning in knowing some resources that they are supposed to allocate for use during the period of initializing and development of the project. The indication of the necessarily necessary costs of production is also a very important factor in influencing the decision that the client take concerning the project. A very high-cost estimate may lead to a loss of opportunity or the scope of reconsideration while a very low estimate may lead to wasted energy that was directed towards the project development, efforts as well as dissatisfaction on the clients that may at times leads to litigation.

The fifth importance of front-end cost estimates in cost planning is that to determine the necessary amount or resources in monetary value that are required to undertake a project. The determinants include the raw materials, labor, machinery and equipment and other variables that affect the carrying out of all the required activities in a project. The estimates assist the people involved in the cost planning to come up with a reasonable budget of the quantity, cost and the price of the required resources to start the operations in a project. The estimates are also very important as they assist the planners to project into the future of the activity that they are willing to adopt and forecasts the costs that are likely to be incurred in the project or activity. This also tends to assists the clients when taking tenders on a certain project so that they can see if the project is worth the investment and they thus compare the estimates with other alternative suggestions to make their investment decisions based on the most profitable activity.

2.2 Requirements for Front-End Cost Estimation

2.2.1 Cost Estimate Accuracy

This is one of the requirements of front-end cost estimation. The accuracy of front-end cost estimation can be evaluated by the differences between the estimated cost and actual cost incurred and it depends on several factors. One of the factors taken into account has who prepared the estimates. This entails the experience of the person who prepared the estimates, engineer's level of experience, the intended use of the estimate, the level of involvement of the project manager as well as the resources used and the level of acceptance of the estimates the appropriate parties who are likely to use the estimates such as the clients.

The second factor that needs to be assessed is to determine the accuracy of the estimates is how the estimate was prepared. This involves checking for the completeness of the cost information used, the accuracy and reliability of the information, the time is given for the preparation of the estimates, how the information used to prepare the estimates is documented, as well as the method, applied to the development of the contingency (Trost and Oberlender 2003) The third factor considered is the knowledge of the knowledge the one who prepared the estimates had on the project. This entails the kind of knowledge in form of the capacity to perform relevant tasks ion the preparation of the estimates, the level of technology the person has, processes, the strategies laid down on the

project, the design criteria of the project, the knowledge on the environmental assessment, as well as the sources of utility and the supply conditions that are required. The last factor to consider in determining the accuracy of the estimates is any other factor that may be taken into account when preparing the project. This revolves around the costs of the owners, logistics for construction and engineering, the labor force available and its productivity, the taxes and insurance as well as the economic climate at the time of bidding.

Accuracy is very important in the engineering and construction projects as it is used by the sponsoring organization as well as the project team. The sponsoring organization uses the cost estimates to make vital business decisions such as laying down the strategies for asset development, screening potential projects available as well as setting aside of the necessary resources required to complete a project that they want to initialize. Accurate front-end estimates help to reduce the instances of wastage of resources, losing opportunities and also getting low returns from an investment project. On the other hand, the project team uses the cost estimates to compare the performance and the overall success of the actual project with the set early cost estimates. The initial cost estimates lay the basis at which all the other future estimates developed are compared. In most cases, the future estimates are expected to agree with the initial estimates by either being equal or slightly less than the initial estimates. However, in most cases, the final cost is usually more than the initial estimates due to errors arising from the uncertainties in the initial estimation.

2.2.2 Reliability of Human Trust

This is the second requirement of the front-end cost estimation, especially in construction projects. Reliability and accuracy are a major requirement in the conceptual cost estimates by the clients and the cost engineers. In establishing the quality and reliability of the cost estimates in the past, scoring methods and common rules were the most applied methods. Due to their various limitations, the methods have currently been substituted with the Conceptual Cost Estimate Reliability Index (CCERI) which is a simple tool, easy to use and understand and is capable of incorporating the weights of up to 20 factors thus influencing the quality of the conceptual cost estimates obtained. The CCERI score helps the clients to recognize the reliability of the conceptual cost estimates thus supporting and assisting them to make informed decisions using the estimates obtained.

Reliability is very essential in the front-end stages of a construction project as it helps to deal with the uncertainties in the projects and thus it's an essential for cost estimators to have enough expertise, knowledge as well as skills in order to reduce the risks associated with the various uncertainties in the conceptual cost estimates to an acceptable level. It also helps the clients in the construction projects to know the estimates made as well as their reliability when deciding which direction to take in their investments by comparing the estimates with other available alternative suggestions. Cost estimates in the cases of clients are said to be reliable if the quality of the conceptual cost

estimates is high meaning that the estimated cost is within the expected accuracy range.

Reliability is measured in a project by evaluating the quality of the early cost estimates. It is measured by the range of accuracy i.e. by checking if the expected range in accuracy measures the required accuracy range. There are several critical factors that affect the evaluation of the conceptual cost estimates reliability. These include; the time and level of data used to get the estimates, identification of the factors influencing and the degree of impact. These factors are classified in various categories. One of the categories is information or data which entails its availability, experience on similar construction projects and the level of the survey on the site. The second category is the definition of the project that entails the level of planning definition, the definition of quality, starting date of construction definition and the capacity of the team involved. The third category is the team involved in estimating cost i.e. its knowledge and skills, career experience and commitment. The fourth category is the procedure used. It revolves around the time allocated for estimation and the standard procedure used for estimation. The last category is uncertainty that entails the level of complexity in construction, the level of competition, contingency and the client's capacity level.

2.2.3 Transparency of Estimation Process

This is the third requirement for front-end cost estimations. It refers to the openness and accessibility to view the estimates obtained where the estimator can easily be able to defend his estimates with a good explanation for the actions taken during the process of preparation of the estimates. Transparency is very useful in construction as well in the industrial projects as it helps the owners of the estimates to gain tremendous of value from the data they have executed in their wide range of project estimates. Transparency in the projects is very useful since it increases the loyalty of the clients who purchase the products from the companies as it reduces their hustle in comparing the estimates from a project with other alternative suggestions in their process of decision making. It increases the reliability and accuracy of the information and estimates obtained from a company's estimates since everything is done in the limelight and the estimators can answer all the challenging questions from the clients to give a proof.

In front-end cost estimates, transparency of the cost is very essential for the decision makers who are in need of information on the efficiency of the available estimates and other alternative suggestions because effective decision making largely depends on the accuracy and the reliability of the estimates. Even if the estimates are obtained in accordance with the existing methods of cost estimations such as the conceptual cost estimation reliability index and other techniques of economic estimation and feasibility, without enough

explanation of the process of the conceptual cost estimates process, the transparency cannot be assured and the estimates may be termed as inaccurate and unreliable (Kirkham2014). The transparency of the estimates is measured based on certain criteria.

The estimating process is very essential in the construction projects, and it poses a lot of challenges to the clients and the contractors when transparency is an issue. This is due to undisclosed cost alterations due to unexpected circumstances such as changes requested by the owner during the building process. This may also be attributed to other deviations that significantly changes the amount of cost estimates made and already accepted in the bid. To ensure transparency in the cost estimates, all modifications, and deviations from the already made and accepted in the bid estimates should be recorded.

The estimators, as well as the people involved in the cost planning, should also include some amount of funds in the budget that should be used to cater for any unplanned for miscellaneous expenses. Inaccurate projections, as well as the delayed estimates revisions, tend to interfere with the transparency in the early cost estimates. This is because they have a philosophical impact on the relationship between the contractor and the owner of the project as they both tend to feel that the other person is not honest, and they are trying to hide something. To ensure transparency, contractors should thus come up with an estimate that explains feasibility costs and the labor costs with accuracy and efficiency.

2.3 Current Practice Reviews on Front-End Cost Estimation

2.3.1 Inaccurate Budgeting for Building Construction Projects

There are many methods used to estimate cost currently. These methods are classified into three main types that include preliminary estimates, intermediary estimates, and final estimates. Preliminary estimations are in most cases used in the early planning stages of cost estimation. They are purposed and based on the needs of the owner and are mainly used to help contractors lay down their bid price. Intermediary estimates are used at certain stages of the design of the project development and are purposed in keeping initial budget's accountability. Final estimates are used in the last stage when the design development is complete. Through it, cost databases are formulated which can be referenced as sources of information for future references (Stoll 1999).

Some of the many methods used to estimate cost before beginning a construction project among others include comparison project estimation, Area and Volume Estimation, Assembly and System estimation and unit price and schedule estimation. Although there are other methods, these four methods are the basic methods used and are therefore perfect representatives of the current status of the available means of estimating the cost of building and construction. Project Comparison Estimation to start with is an estimation method that is also known as parametric cost estimation and is in most cases used in the early stages of planning.

During these stages, there is little-known information concerning the projects whose cost is being estimated. This method therefore mainly relies on historical data of the total costs of similar projects or same building types and used that information to formulate the cost of the projects in hand. If an example of a road construction project is taken, one can estimate the total cost of that project from comparing the project with another road construction project of similar design. If the road being constructed was, for example, a tarmac road about 20 kilometers and the current project is similarly a tarmac road but maybe of fewer kilometers, use of this method can be applied where the current price of the new project can be estimated basing the facts from construction document of the already complete project of the 20 kilometers road (Stewart et al. 1995).

The next method, assembly, and system estimation is one which applies to the intermediary level stage. It is used when the design drawings range between ten to seventy percent. This method operates by grouping several trades' work and other work items into only a single unit purposed to estimating. Estimations done using this method in most cases have their accuracy ranging in around 10 percent. Like project comparison estimation method, this method acquires its data and any other information needed from historical information of previously worked on and completed projects. It also incorporates the use of estimation software in the process oof estimation, which enhances accuracy and makes it easier to acquire estimations.

Unit price schedule estimating is another method that operates by dividing the project into the smallest possible project portions. An estimated unit price is obtained for each portion and total estimation price calculated from the addition of all unit prices. Square and cubic foot estimation are the last basic and main estimation method used to estimate the cost of the building. It is used in developing both preliminary and intermediate budgets most of its information is based on historical data and the already designed project. It can make reliable estimates when the design of the project is developed fully to allow calculation of the areas and volumes of the proposed project content.

These methods are however faced with many problems and challenges. Some of these main problems include lack of information, lack of similarly designed buildings and projects from which one can make comparisons and the immense changes that progressively take place. Lack of information affects programs in a great way. It is through reliable sources of information such as historical data of previous projects that estimation methods are put into use. Without this information, it is almost impossible to estimate the cost of a certain project. A possible solution is the project owners hiring only highly experienced contractors and estimator who probably have access to the required information. The other problem, lack of similar projects that can be used for comparison purpose is a major problem and estimation barrier especially for certain estimation methods like the project comparison estimation method that relies on other buildings with similar designs for estimation. A solution to this problem projects cost estimators to have several other methods of estimation they can use for the project in case one of the methods lack information.

2.3.2 Limited Information Availability

Information is among the most important resource needed for cost estimation. Every cost estimation method requires certain information for it to be put into practice. Most of the estimation methods use historical data of already completed building projects. The documents by these projects are used by estimators in obtaining sufficient information that make it possible for the costs of proposed projects to be estimated. Information is however not limited to historical data. Firsthand information from the owner of the project is first put into recording, from where the process of designing the preference building begins. Through this information, contractors can make a graphical drawing of the contract owner's description concerning the preferred building. From the owner's description, main factors of the building such as size and design are obtained. Owner's project description is, therefore, the first information required for estimation. Without it, there is a large number of choices from where the contractors and construction managers can choose from (Anderson et al. 2007).

Limitation of this information makes it extremely difficult for the whole cost estimation process to take place. It becomes impossible to systematically develop the design of the project without which cost estimation cannot take place. The only solution available to avoid such an occurrence is for the project owners to describe clearly the preference building to design experts and estimators. With this information, preliminary stage estimation methods can be

used to make estimations. In most cases projects do not experience difficulties obtaining this sort of information since the owner gives descriptive information concerning how he or she would like the building to appear and further clarifications can still be made.

Together with information from the project sponsors, cost estimators and project contractors require information from past building projects of similar design, from which they obtain comparative information and can estimate the price of the current project. Unlike information obtained from project sponsors, however, this information is sometimes challenging to obtain, especially in cases where for example the project consists of constructing a unique building whose design is not common. For this sort of a project, if made unique enough, it may be a difficult task to obtain such an already completed project from which estimation data can be obtained. A solution to this barrier and problem is hiring highly experienced cost estimators who can relate building projects designed in a different way to the proposed building project and obtain a reliable estimation of the cost of the project.

Due to insufficient information, other problems such as cost burden assumptions arise. These are the assumptions that some certain cost of the project is included in a contractor's bid price while the contractor does not regard that cost as part of her bid price. This kind of confusion occurs due to insufficient provision of information and lack of clarity during the signing of a bid contract. This sort of a problem has great consequences of affecting the

whole estimation process. To start with such kind of mistakes causes overruns the already prepared budget and as a result, the budget can be termed as inaccurate. A solution to such a problem can be obtained by simply ensuring that for the case of both parties sealing their agreement with a contract, proper understanding is obtained concerning what is expected from each one of them (Potts and Ankrah 2014).

Insufficient information, therefore, is the main challenge and barrier to a project's cost estimation. Lack of information, necessary for estimating the cost of a project, results to the impossibility of the available cost estimation processes in obtaining an accurate estimation cost of the project. This may, however, be misleading thus resulting to major problems later in the process of working on the project. It is, however, possible to lack a certain portion of information however and still make accurate cost estimation by simply by changing the estimation method and using one that does not require the missing data portion. The best way to, however, avoid lack of information problem in cost estimation is ensuring to hire highly experienced people who have gained sufficiently experienced and worked on many projects thus they are more likely to have sufficient information that can be used for the current project.

2.3.3 Limited Usage of Unit Price of Actual Construction Cost

Unit price involves a cost estimation method that is used by cost estimators to estimate the total cost of a building project, through dividing the project into smaller unit portions, whose estimations are easy to obtain. This method if properly used can make accurate estimations since estimating a possible cost for a single unit is in most cases likely to be more accurate than making estimations for a mega building project. The unit price method is therefore of a greater advantage but also has its limitations. One using this method obtains the total cost estimation figure through the addition of unit costs. This summation can be calculated using the formulae:

$$\text{Summation of } P_i Q_i + T = K. \quad (\text{Eq. 2-1})$$

For the case of this formula, i represent the total number of the priced items, p represents the unit price of each item, Q represents the quantities of each item, T the amount of added tax and K represents the total estimation cost. The total number of the cost is obtained by summing the value representation of all values of i . for this method to be put in place or to be used, there is certain crucial information required without which it is not applicable. The components of the formula are simply the main requirements needed for estimation by this method to take place.

The number of priced items is obtained in the middle stage and requires information concerning all the items that are to be purchased for the project.

The P_i in the formula represents the total number and price of these items. Q_i represents the total quantity of all the items and T that represents the value added tax. All this information is not available for projects in the preliminary stages. A preliminary stage in most cases only contains historical data and information from project sponsors describing their preference for the building. This kind of information is insufficient to enable usage of unit price estimation while estimating cost. This method's limitation is that it cannot be used in certain stages of cost estimation, despite its high advantages and accuracy level.

This limitation is a major setback of the available methods and means of cost estimation. Its limitation from enabling estimators to estimate the cost using it at any stage has caused many problems and inconveniences. To start with, since it is among the most accurate methods of cost estimation, some people would prefer using it due to its accuracy. Their preference is however considered impossible in the early stages due to lack of the required information. If the case of building a school is considered, there are many ways through which this project can be estimated using the unit price method. It is however only possible when the design of the school is at a certain level. The most common unit area that can be considered for this case is dividing the school into classrooms. To do this, however, one needs certain information of the number of classes the school is proposed to build the classes. The information concerning the number of classes is not obtained from the preliminary stage but the middle stage of design development. In this stage, the number of classrooms, their design, their components and geographical locations are determined. With

this information, therefore, the method can efficiently estimate the total cost of building the school.

This estimate can be obtained by taking an individual case of a single class into account, calculating all the components of the class that requires being included while building the class. These may include certain infrastructures such as classroom boards, seats, a teacher's table and any other item which the project sponsor specifies should be in a classroom. Calculations of the total cost of the project are obtained from the multiplication of the total cost of a single class including that of all its components with the number of classrooms. This method is much better and manages to make simply estimations of even much more complex projects, since, after division into unit portions of priced items, it becomes easier to estimate. The unit price method can, therefore, be considered as a good estimation method that can be relied on. It's only, and major problem, however, is that it is limited to certain stages of production, and is not applicable to others such as preliminary estimation stages (Gerrard, 2000).

2.3.4 Lack of Flexibility for Diverse Building Projects

There exist many different types of building projects; these projects differ in design, building materials, geographical areas, surrounding environments, and contractors who worked on the projects among many other distinguishing factors. Building projects made of different designs differ from each other both regarding cost and the view of the buildings. A difference in design might result

in a requirement for adjustment of building and cost estimation methods. This, in turn, would cause difficulties while obtaining historical documents with an aim of acquiring information that might help in price estimation of the current document. A difference in geographical areas would also cause a significance difference in two building projects. If for example two same sizes with similar design construction projects are in progress, with the only difference being that one is situated in a remote area and the other in an urban area, then the cost estimation would be likewise different. This is because for the project being carried out in remote areas; additional transportation cost should be added. Other inconveniences such as lack of manual labor during the building process also are evident in rural area projects unlike those geographically based in most urban areas.

Different contractors working on similar projects would also result in a change in the cost and many other changes. This is because every contractor has different ways and methods in which he or she uses while working on a certain project. This difference in contractors results to a general difference in the total price of a certain building project. The diversity of building projects is however not advantageous concerning cost estimation. To start with, most cost estimation methods require information from similar projects that were already completed at a later date. When the available prior projects and data are not similar, therefore, cost estimation becomes an extremely challenging task. Sometimes, similar building projects may be obtained, but those which were situated in a different area from the current position. If such a case for example

occurred, then there would be high possibility of inaccurate cost estimations this is because if the project was for example situated in rural, remote areas and the current project is situated within an urban area, then basing the current projects estimation from the previous project without changing much would mean that the estimated price would be more than the actual price, and thus, inaccurate estimation would result (Smith 1995).

Some cost estimation methods are only applicable to certain building projects and are not applicable to others. These methods lack flexibility since although they may work with almost all building projects, it is only to some of the projects that they are fully put to use and can give an accurate estimate of the project's cost. A unit price method, for example, can be applied in most of the cost estimation stages; however this method is put into total use when estimating a mega structure that requires being divided into portions and estimation is made for these individual portions that are then summed up to get the total estimated cost of the mega project. This fact explains that cost estimation methods are rigid and lack flexibility thus making it hard to determine the best method to work on the variety of the building projects. There is also no estimation method yet, which can apply to any form of building project thus one is expected to keep choosing from the available options.

To solve these problems, there is need to develop a more flexible method of estimating cost, capable of being flexible enough to estimate accurately in all building projects. This kind of a method should be highly applicable to all

stages of design development and can be used any available information. With such a method, cost estimation process would be made easy, and contractors and other cost estimators would not have to put too much effort and spend a lot of time and resources trying to determine the best and most accurate method to use with a certain building project. Lack of this kind of a method, however, has greatly negatively impacted the estimation of cost in building projects. Due to this fact, cost estimators are forced to keep changing the methods they use to estimate cost now and then with how the design developments unfolds. This results in sometimes having to use two or more cost estimation methods in the same project.

2.4 Literature Reviews on Front-End Cost Estimation

Front-end cost estimation is a very important aspect of a construction project. It is through it that constructors are hired by certain project sponsors after they present a favorable budget for the project during the bid session. According to Pica (2015), it clearly describes the probable price of a complete construction project, which guides those sponsoring the project in determining the worthiness of proceeding on with the project, or deciding that the project is not worth the effort and attention it requires and thus withdrawing the intentions of working on it.

Kim (2005) and Kirkham (2014), on the other hand, explains that cost estimation has many other uses and reasons as to why it is of great importance to constructors and all other parties involved in a construction project. To start with, they explain that it guides constructors and project designers on the best design of the building in considering the budget, the owner's preferences and the cost it is likely to require. With all these considerations put into place, the best building design and the most economical is chosen, ensuring that it conforms to the preferences of the owners' of the project and the already formulated budget. Early cost estimation also guides the contractor to determine what construction method is best applicable to a certain project. This is done in certain estimation stages where certain construction method's cost is estimated and then compared. The method that gets the work done in the most economical way is thus chosen. Through cost estimation, therefore,

constructors, designers, cost estimators and project owners can successfully complete construction projects in the most economical way.

Stoll (1999) explains that most of the estimation methods used are not accurate. They face certain challenges and problems that cause their total estimated cost in most cases not be similar to the actual cost completing a project requires. Some of the main problems that cause inaccuracy in cost estimation methods involve lack of required information, use of wrong methods at certain stages and other errors such as assumptions and omissions made during estimation. Lack of the required accurate information is the main problem hindering these methods from obtaining accurate cost estimation. Without this information, it is impossible for cost estimators and contractors to estimate the cost of a building. The information includes documents with information from previously completed and similar projects, from where further analysis and assumptions are made to obtain an estimated cost of the current project. Potts and Ankrah (2014), however, argues that even in cases where all the required information is provided, it is difficult for cost estimation methods to obtain an accurate estimation cost or one matching the actual cost of a project. He bases his argument on an explanation that change occurs now and then thus even with information provision, change that occurs during the duration between when the project from where information is derived is complete and when the current project begins might still cause a difference that would reflect in the price thus causing it to differ from the actual one.

If for example a case where one is building a house acquiring information from the same size and similar design house build about two years ago, that information may be sufficient in every stage of estimation but the two year difference between completion of the previous project and starting from the current one would cause a great difference in price between the two projects thus if that difference is not taken into consideration, the estimated price would be far much different. Using wrong methods for estimation, which is the next problem, also contributes in a great way towards obtaining of inaccurate estimation costs. Some estimation methods are only accurate in certain construction stages and thus if used on different stages; inaccurate results are obtained. A unit price estimation method is limited in that it cannot be applicable in the preliminary estimation stage. This is because this stage lacks the required information for the method to be used thus any usage of the method in the stage would result in inaccurate estimation cost. It is due to these problems and challenges therefore that early cost estimation becomes a difficult task and estimation methods are unable to acquire accurate estimation cost. Although these methods are unable to provide accurate results due to the challenges they face, they however if properly done provide information that is essential in the construction projects.

2.5 Summary

Front-end cost estimation is very useful in industrial as well as construction projects. It is used in the cost planning to develop estimates that are used in the formulation of the budget for the construction projects. Accuracy, reliability and transparency are the basic requirements for the front-end cost estimation. The estimators should ensure the accuracy of the information used to develop the estimates to reduce the errors arising from overvalued or very low-cost estimates. In this case, the reliability of the cost estimates which is measured by checking the accuracy range should be taken into account by ensuring that the expected cost estimates obtained are close to the actual cost estimates required.

Reliability is very essential to the clients and the contractor engineers. This is because it acts as a basis for making their initial decisions where the clients check for the reliability by looking at the quality of the requirements in the project. They thus use those estimates as a basis for deciding whether to undertake the project or to compare the estimates with other alternative suggestions to choose the most profitable and applicable one to implement. In the past, the scoring method and the common rules method were the ones that were used to measure reliability but due to their complexity and being sophisticated, they have been replaced by the conceptual cost estimates reliability index that are simpler, easy to use and easy to understand as compared to the other two methods in the past.

Building construction projects require being budgeted for before they are worked on. The budget is only formulated when all the components and requirements of the project are well priced. Before the project begins, however, it is a difficult task to determine the materials and resources needed for the project, and their prices to formulate the budget. This, therefore, raises the need for having estimation experts assigned the task of conducting research concerning the requirements of the project, pricing those requirements and finally estimating the total cost that the project requires. These experts accomplish this by using certain estimation methods that include unit price estimation method, comparison project estimation, Area and Volume Estimation and Assembly and System estimation method. Through these methods, these experts can acquire an estimation of the total cost which would accrue to the project.

Despite the many efforts put into place by cost estimators, the figures they currently obtain are not accurate and differ from the actual cost of the projects. This difference and inaccuracy arise from the many challenges and problems facing their practice. Lack of sufficient information is one of the main problems they face. All methods used for estimation requires certain information. Insufficiency of this information results to the impossibility of using those methods to obtain estimation values or if they are used they provide inaccurate estimation figures. Lack of information also disqualifies the use of certain methods like the unit price method in some stages of design development. Other challenges involve the use of wrong methods for estimation and errors during

the process like omissions and assumptions. Cost estimation is of extreme importance; building projects would be extremely risky in case there are no means of estimating prices before the project begins. The available means should, however, be improved to increase their accuracy.

Chapter 3. Case-Based Reasoning Approach

Using the knowledge gained from past experiences to solve new problems, case-based reasoning (CBR) has become a popular problem-solving methodology and has been applied widely to improve the accuracy of construction cost estimation (Doğan et al. 2006; An et al. 2007; Sevgi et al. 2008; Kim and Kim 2010; Koo et al. 2011; Ji et al. 2012; Jin et al. 2012; Ahn et al. 2014). Ozorhon et al. (2006) stated that there is a tendency among most people to think about the problems they have encountered in the past when they are faced with new challenges. Apparently, the main reason why people flash back to their earlier experiences is to try and establish a relationship between a current problem and those they have faced in the past (Schank et al. 1994).

This chapter elaborates principles of CBR and discusses advantages of CBR as an effective methodology for conceptual cost estimating, limitations, and challenging issues that need to be examined to improve an explanatory power of CBR cost estimation. Furthermore, normalization for case representation, attribute weighting for case indexing, and similarity measurement for case retrieval are reviewed in depth to find out directions of enhancing reasoning environment of CBR cost estimation.

3.1 Principles of CBR

3.1.1 Overview of CBR

CBR stemmed from cognitive science, which deals with psychology and theories about how human memory works (Pal and Shiu 2004). Episodic memory coined by Tulving in 1972, reflects the memory of autobiographical events such as times, places, and other contextual knowledge and well demonstrates a method of human brain for storing and recollecting large amounts of information from an individual's past (Rubin 2005). Schank (1999) asserts humans utilize cases and the reminding process is the vital process in intelligence. People choose a matching case and require more experience when inappropriate matches occur. Based on this primary human memory work, CBR methodology was developed.

CBR process helps cost estimators to remember previous cases when confronted with present cost estimating challenges for new building projects, thus facilitating the process of redressing such problems (Golding 1995). CBR process solves problems through analogy of such problems against previously recorded cases to determine whether such problems are worth investigating (Golding 1995). The rationale of the approach is that previous circumstances can be analyzed to provide insight that may help in solving presently experienced problems. Therefore, when a present problem occurs, it is first compared to case-base to determine if it is similar to any of the previously solved cases. If it

is similar to any of the stored cases, then the same process applied to solve the previous one (Golding 1995). Not only does this approach simplify the process, but also facilitates a faster problem solving if the cases can be retrieved faster and comparison finished faster.

3.1.2 CBR Problem-Solving Process

In general, CBR comprises four phases: retrieve, reuse, revise, and retain (Figure 3-1). Singular or multiple similar cases are retrieved at a conceptual level of CBR, when a new problem is matched with cases in the database, (Mantaras et al. 2005). Then, the suggested solutions or cases are reused to treat the problem. If the suggested cases are not a close match, then the solutions need to be revised. The revised solutions are retained as new cases in the database (Watson 1997). Also, in each phase of the CBR life cycle, task and method can be decomposed as shown in Figure 3-2. Among each phase of CBR cycle, many research efforts have been made especially on retrieve phase as case retrieval is considered the most significant process within the case-based reasoning (Pal and Shiu 2004; Ji et al. 2011b; Ahn et al. 2014). Furthermore, similar experience can lead to future reasoning, problem solving, and learning (Smyth and Keane 1998).

There are three types of case-based reasoning problem solving approach, including conversational, structural, and textual case-based reasoning. In texted CBR, cases are represented as free texts. The approach is also characterized by a collection of many documents, easy acquisition of cases, keyword matching

and syntactic retrieval. In structural approach, cases following vocabulary arrangement. It is also characterized by partially filled description questions, and values are assigned to predetermined values. Conversational approach was the first problem solving approach to be used under the framework of CBR (Aha et al. 2001). It applies a different principle than the others in the sense that it does not assume the availability of similar cases to the current problem. Rather, the process entails establishing a dialogue with the user to elicit the main features of the problem (Aha et al. 2001). This approach aims at reducing the number of features of the problem that has to be compared against previously solved cases. Thus, the process is relatively more time saving compared to other approaches.

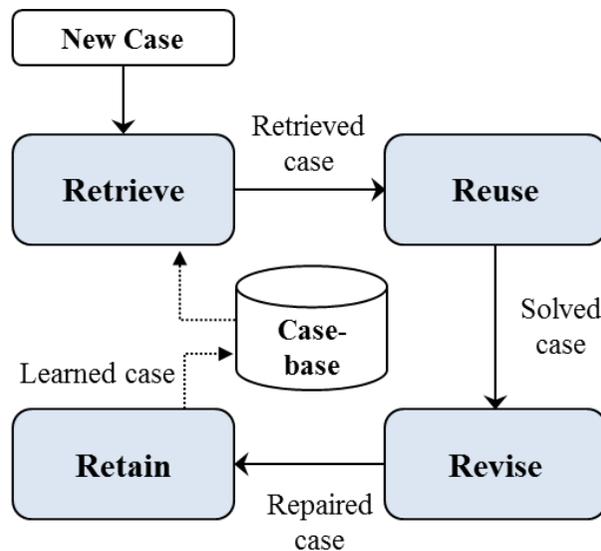


Figure 3-1. CBR Problem-Solving Process (Aamodt and Plaza 1994)

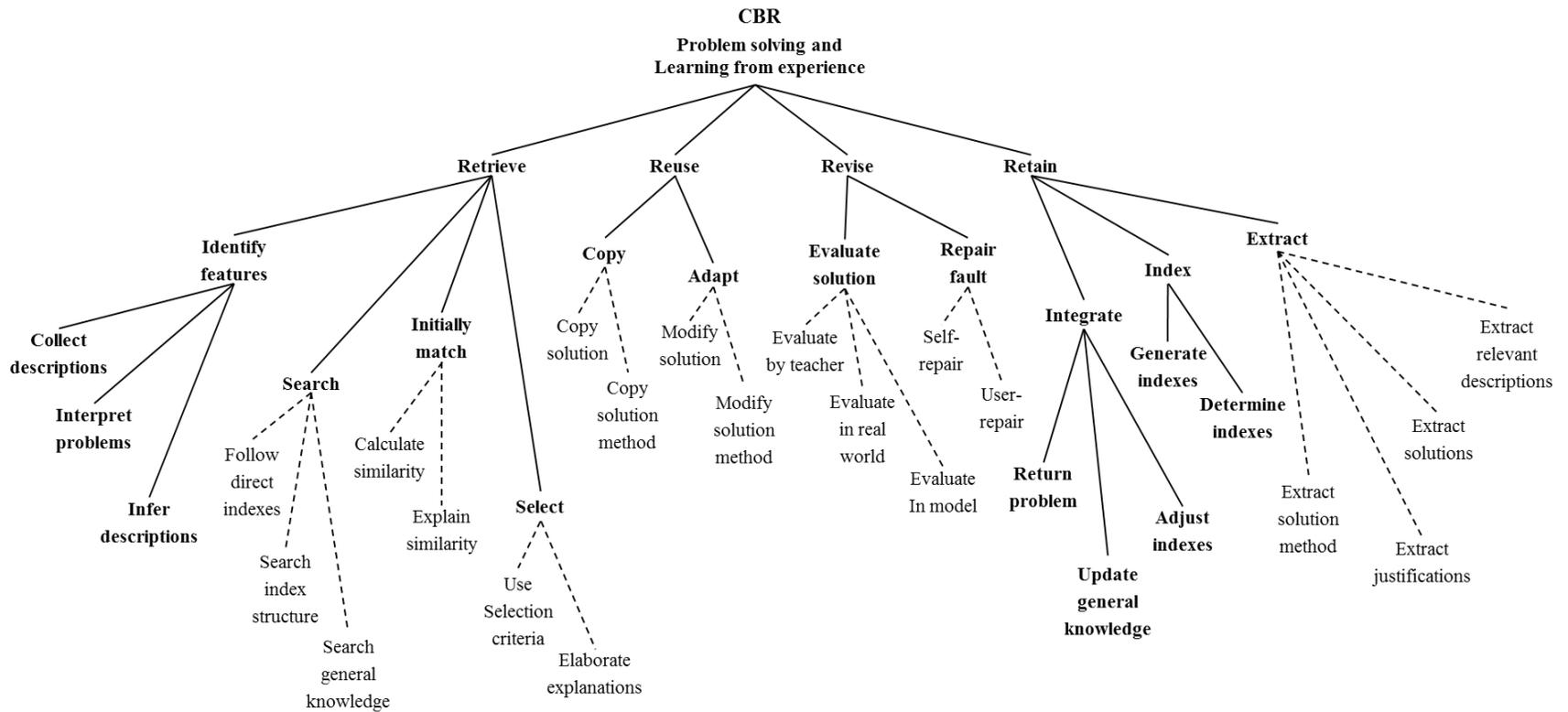


Figure 3-2. Task-Method Decomposition of CBR (Aamodt & Plaza 1994)

3.1.3 CBR and Rule-Based Reasoning

The main difference between case-based reasoning and rule-based reasoning (RBR) is that the latter requires an explicit model of a domain while the former does not harbor such a requirement (Lee 2008). RBR is carried out using induction rules to identify challenges and establish whether they are worth investigating. On the other hand, CBR process identifies problems and whether they should be investigated by matching the current one to previous records of the same.

In CBR, the process is much simpler than RBR because it mainly involves gathering cases that can be used identify important aspects and adding them to the case-base during the development process. CBR method is relatively easier compared to RBR because knowledge acquisition is unnecessary during the development of case-bases. CBR process is also simpler than RBR process because case-bases can be used by people with little computer expertise even when it is not complete. In the non-completed conditions, the CBR can facilitate addition of cases to the structure that was developed previously by a computer expert.

Conducting a maintenance process using RBR can be a daunting task especially if the rules were not carefully written. In most cases, it requires computer experts to debug the process to complete the maintenance process (Prentzas and Hatzilygeroudis 2007). Moreover, the addition or deletion of

cases does not qualify as debugging. However, it can have an effect on the outcome of the process. However, Bryant (2009) explained that RBR suits problem-solving processes where the system is single-purpose, specialized and rules are predetermined.

Nearest-Neighbor Retrieval and Inductive Retrieval

To retrieve similar cases, two methods, nearest-neighbor retrieval and inductive retrieval, are commonly utilized. Nearest-neighbor approach computes and identify similarity of cases by measuring the distance between a target case and source cases (Everitt et al. 2011). This method needs to define attributes and their values, which characterize the cases in the population (Hall and Park 2008). It requires computation of similarity (i.e., distance) among attributes and cases. This means that although this method is simple, the retrieval time is likely to be an issue as the size of database or number of attributes of cases increase (Pal and Shiu 2004). Inductive retrieval extracts rule or derive decision trees based on historical data. Hence, to extract rules or decision trees, the case-base should be analyzed to exactly classify the cases (Pal and Shiu 2004). However, a major disadvantage is that where case data is missing or unknown, the retrieval of cases cannot be carried out. Also, there is a difficulty to discover all the necessary rules beforehand (Kim and Hong 2012).

3.2 CBR Advantages, Limitations, and Issues

3.2.1 Advantages of CBR

As a problem solving methodology, CBR solves new problems by adjusting previously utilized solutions to solve old problems (Riesbeck and Schank 1989). Regarding knowledge level, CBR differentiates by using the specific knowledge of past concrete cases whereas other artificial intelligence methods such as neural networks and expert systems depend on general knowledge and relationships between problem information and conclusions (Aamodt and Plaza 1994; Arditi and Tokdemir 1999).

CBR entails accessing information stored in cases that are similar or close to the newly addressed problems. According to Leake (1996), information retrieved from this assessment is vital in predicting the value of the target features of the new problem. The methodology has proven effective in solving complex problems with many alternative solutions. In CBR systems, memory is seen to form the basis of a learning capability (Marir and Watson 1994). Furthermore, a CBR system has the potential to learn without looking in to particular formulas, cases, symbolic representations or rules (Globig et al. 1997). This type of demand driven approach has many advantages (Leake 1996; Aha et al. 2001), which include improved user acceptance, incremental learning, reduced problem solving and easier knowledge acquisition. Also, Leake (1996) asserts that CBR is more effective than the rule-based systems, which are useful where only one or a few solutions to a problem are possible.

According to Pal and Shiu (2004), advantages of using CBR can be summarized as below:

- Decreasing the knowledge gaining task
- Avoiding continuous mistakes happened in the past
- Supporting flexibility in knowledge modeling
- Providing reasoning the solution to the problem that have not been fully understood, defined, or modeled.
- Forecasting the possible success of a suggested solutions
- Acquiring continuous knowledge acquisition over time
- Possible to reason in an area with a small body of knowledge
- Possible to reason with inadequate or inaccurate data and concepts
- Preventing from repeating all the steps that requires to arrive at a solution
- Facilitating a way of explanation
- Applying to various different domain areas and objectives
- Stemmed from human reasoning

One of the artificial intelligence tools, neural networks, represent explored as an estimation method. Its usage can prove beneficial when involving intuitive judgment or when patterns of data become too irregular to identify with traditional techniques (Hegazy and Ayed 1998). Also, Smith and Mason (1996) found neural networks can accept more cost factors than regression and can more easily deal with multicollinearity. However, another study found it time-consuming to determine the network factors that best fit the application (Bode 2000). More importantly, Hegazy and Ayed (1998) and Smith and

Mason (1997) admit that the process of the artificial neural network lacks transparency, and they regard it as a “black box.”

Among the several estimation methods, parametric estimation serves as the most commonly used method during early design stages. This method does not require detailed information on a project and works relatively quickly at minimal expense for estimating approximate costs (AACE 1999). Also as Ashworth (1983) explains, it has greater accuracy in cost estimates compared to the square foot method. Considered the simplest method used to establish reasonable cost estimation, it identifies the cost of similar projects and compares the findings with the cost of the new one (Ellsworth 1998). Also, the greater the accumulation of historical data from similar projects, the higher the estimate accuracy.

3.2.2 Limitations of CBR

A case-based reasoning process is a form of scientific research based on the focusing of a particular aspect or a task within an institution as the target sample size (Mansar et al. 2003). It is adequate in achieving the best results of the intrinsic characteristics of the predetermined set sample size, and it is only applicable to that sample size. The rigorous application of this research and forecasting method is incomplete. It is because the validation and reliability of the method are limited by the criterion used for the selection of the sample size and the data used for the research. The analysis is only applicable to the particular scenario. However, to replicate the results of the research, it would entail designing the similar attributes that are considered in the study. The attributes used in the research are specific and categorical to the focused sample.

The case-based reasoning methodology is difficult to reproduce and thus not applicable to other situations. However, one of the most critical benefits of the method is its ability to give intrinsic and extrinsic characteristics/analysis of the focused scenario in a broad and in-depth analysis. It is mostly preferred scientific approach when dealing with a unique set of attributes that needs analysis. In short, it can be said that the CBR is case sensitive.

The CBR relies on the already existing data from which comparative projections, estimations, computations, and relationships are evaluated. If the validity of the existing data is questionable, and thus, the results obtained by

the use of the data are uncertain. It can only be valid if and only if it is employed in the case study. It is critical to focus on the methods that were used to obtain the data that is to be used for the study. Changes and allowances could be adjusted accordingly to improve the reliability of the research.

3.2.3 Challenging Issues in CBR

Challenging issues in CBR (Aamodt and Plaza 1994; Watson 1997; Pal and Shiu 2004; koo et al. 2010b; Ji et al. 2012) are summarized as below:

- Do all indexed features have the same weight?
- Is the similarity linearly proportional to the distance a case is from the new problem?
- What distance measure should be used?
- Uniformity of solution case
- How many cases are needed
- How to organize case-base
- How to remove overlapping cases
- What variables to use for indexing
- How to weight the attributes
- How to revise the case
- How to extend and retain the cases of CBR database

Do all indexed features have the same weight?

Using the previous knowledge and experience in order to solve new problem has been faced by a serious challenge where the weight of the indexed feature does not coincide with the previous one. This is very true especially in a case where the nature of the new problem is not exact like the one done before. Although the problem needs similar knowledge so as to arrive at the solution, weight of the materials which differ may lead to a difference in approach in solving problem solving technique (Montani and Jain 2013). A good example is an engineer who designed a simple building using and has a task of designing a larger similar building.

Is the similarity linearly proportional to the distance a case is from the new problem?

Similarity in solving a problem cannot be linearly proportional. This is due to many errors that are experienced during the process of problem solving. A good example of this comes out clearly when the new problem needs the same approach that the one before used although it has different parameters. This parameters may include measurement, quality, and quantity among others which may change in the process of working out for solution of the new problem (Richter & Weber, 2013). A good example comes out clearly when comparing two similar buildings which differ in dimensions and quality of the material used. This difference thus leads to nonlinear proportionality in comparison.

What distance measure should be used?

Engineering work may pose a challenge for use of CBR technique. This is due to the difference that may arise due to surfaces that has to be measured. For example, CBR may pose a challenge at a place where the surface to be measure completely differs from the previous one. For instance, measuring distance on a circular place where construction has to take place may be difficult especially if the previous one was done on a linear surface (Richter and Weber 2013).

Uniformity of solution case

Even though solution used may be uniform, the application of the solution may differ. The little leakage in the solution of the material used may cause an error hence very difficult to achieve accuracy. It thus becomes hard to use CBR to solve the present problem.

How many cases are needed?

The number of cases required for CBR model may vary according to ranges and variety of cases. Also, a small number of good quality of cases is more important than poor quality of data in large numbers.

How to remove the overlapping cases

Many construction field have had overlapping cases which have been solved differently depending on a myriad of factors. Changes in factors may lead to changes in overlapping in the new problem. At the end, treatment of overlapping cases may take a different course which may contradict the previous experience. In some cases, such cases may not exist hence carrying forward a little doubt about the new solution due to difference in expectation.

What variables to be used for indexing

When variables do not match or give similar results as that which was obtained in a previous problem, then indexing may be difficult. It results to unknown source of error which may give troubles in estimating the source of problem. In turn, variables to be used in indexing may be difficult to determine (Ram and Wiratunga 2011).

3.3 CBR Model Components

3.3.1 Normalization for Case Representation

The CBR method can be viewed as an effective and accurate method for cost estimation in construction by utilizing knowledge gained from past experiences (Doğan et al. 2006; Koo et al. 2011; Kim and Hong 2012). In CBR cost estimation, data preprocessing is a preliminary process that is often used for working out vital or meaningful relationships and patterns hidden within a large quantity of information (Pyle 1999; Liu and Metoda 2001; Ji et al. 2010). Thus, normalization is very crucial in data mining, which apparently comprises a variety of processing procedures that are aimed at preparing raw data for further processes, including the actual estimations (Han and Kamber 2006; Shalabi et al. 2006). Normalization in this context refers to the adjustment of values measured on different scales to a notionally common scale. Pal and Shiu (2004) stated that the process of normalization involves the conversion of all collected data into standardized values.

There are different methods of normalization applied and each method focuses on achieving a different goal. Ji et al (2011b) suggested a cost estimation model for building projects using CBR, and utilized z-score normalization to standardize errors when population attributes were known. According to Ji et al (2011), this is a normalization method that can be applied to normally distributed populations. In statistics, the standard score (also referred to as the z-score, normal score or z-value) is the number of standard

deviations an observation or datum is above the mean. According to Sevgi et al. (2008), the median (Med) method is almost similar to the total count (TC) method in which unit counts are divided by the total number of mapped reads associated with their lane and multiplied by the mean total count across all samples of the dataset. However, in Med, the total counts are replaced by the median counts that are different from '0' in the computation of the normalization factors.

The Sevgi et al. (2008) method has proven instrumental in instances where decision trees are used to determine attribute weights in a case-based model of early cost prediction. On the other hand, feature scaling is a normalization method used to standardize the range of independent attributes of data. Koo et al. (2011) developed a construction cost prediction model with improved prediction capacity using the advanced CBR approach. The method is utilized to deal with the challenges posed by the wide range of raw data values. This data normalization procedure is carried out during the data preprocessing step and is intended to standardize features that have a broad range of values so that each of the values contribute proportionately to the investigated final attribute.

Shalabi et al. (2006) emphasized the different methods of normalization against the induction decision tree (ID3) using the Hue Saturation Value (HSV) data testing procedures. Three normalization methods, the z-score method, min-max, and decimal point normalization, were tested. The results showed that min-max was the best method for the training data set with regard to the

accuracy and efficiency of the whole HSV data set. However, the test on normalization methods was limited to the use of training data and the data structure and types of attributes were not clearly stated. Therefore, tests on normalization methods needs further examination based on real construction project data using CBR methods.

Furthermore, as can be summarized in Table 3-1, very few research on the area of CBR cost estimation have emphasized the importance of applying the normalization method. Most previous studies are limited in stating if normalization was or was not performed. Also, some studies do not indicate which normalization methods were utilized. Overall, in-depth examination in which normalization methods are more reliable in terms of accuracy and stability when they are applied to CBR cost estimation have rarely been conducted. Moreover, the appropriate selection method of normalization needs to be further discussed. Therefore, there are significant needs to conduct comparative research that examines and determines which normalization methods lead to higher accuracy and stability for CBR cost estimation.

Table 3-1. Literature Reviews on Normalization in CBR Cost Estimation (revised from Ji et al. 2011b)

Researcher	Objective	Data Profile				Attribute			Normalization		
		Project type	Year	# of cases for model construction	# of cases for validation	# of Attribute	Scale type	Weighting method	Applied	Method	Reason for selection
Kim and Hong (2012)	Cost estimation for railroad-bridge construction project in the planning phase	railroad-bridge	1998-2009	134	5	8	nominal, ratio	GA	X	N/A	N/A
Ji et al. (2012)	Develop a case adaptation method for construction cost estimation	military barrack	2004-2008	129	13	18	nominal, ratio	GA	O	STANDARDIZE & NORMDIST function	N/A
Jin et al. (2012)	MRA-based revised CBR cost prediction model	business facility, multi-family housing	-	59 (multi-family), 31 (business)	40 (multi-family), 10 (business)	10	nominal, ratio	MRA	X	N/A	N/A

Ji et al. (2011b)	Develop military facility cost estimation system	military barrack	2004- 2009	422	10	9-18	nominal, ratio	GA	O	STANDARDIZE & NORMDIST function	N/A
Koo et al. (2010a)	CBR based hybrid model for cost and duration estimation	Multi- family housing	2000- 2005	101	-	20	nominal, ratio	GA	O	N/A	N/A
Kim and Kim (2010)	Preliminary cost estimation	bridge	2000- 2005	585	30	5	nominal, ratio	GA	X	N/A	N/A
An et al. (2007)	CBR cost estimation model using AHP	residential building	1997- 2002	540	40	9	nominal, ratio	AHP, Feature counting, gradient descent	X	N/A	N/A
Doğan et al. (2006)	Cost of structural system estimation	residential building	-	24	5	8	nominal, ratio	Decision tree	X	N/A	N/A
Yau and Yang (1998)	Cost and duration estimation for a building project	Office building	-	60 (hypothetical)	3 (hypothetical)	10	nominal, ratio	Subjectively assigned by authors	X	N/A	N/A

3.3.2 Attribute Weighting for Case Indexing

In the literature, Jrade and Alkass (2007) illustrated a methodology for an integrated parametric cost estimate and life-cycle costing of building projects during the initial phase. Their study showed attributes selected for the parametric cost estimate; however, it did not address how they chose attributes. Soutos and Lowe (2005) developed a parametric cost model using a linear regression based on building cost data and introduced an early stage cost-estimating package. This research adopted a methodology of identifying cost-significant variables by literature review and via a series of meetings with industrial collaborators. Cheng et al. (2009) described the process of developing a web-based conceptual cost estimator using a genetic algorithm, fuzzy logic, and neural networks. The authors selected key factors affecting construction costs through a literature review, brainstorming, and the drafting of an influence diagram and a hierarchy of objective techniques.

Seong et al. (2008) presented cost estimating using a parametric method for apartment building projects and conducted multiple regression analysis to derive a cost-estimate relationship formula. As a means of extracting influential attributes, they performed correlation analysis between cost and influential factors. Sonmez (2008) introduced the parametric range estimating of building costs using regression models and bootstraps. To select attributes that have significant effects on the cost item, they chose and included in the cost models attributes with a regression coefficient significance at a 0.2 significance level.

An et al. (2007) proposed a case-based reasoning cost-estimating model and attempted to include experience using an analytic hierarchy process. They conducted a questionnaire survey applying an AHP approach to elicit domain knowledge from experts and to determine the weights of attributes. Dogan et al. (2006) introduced a spreadsheet-based CBR prediction model of structural systems and assessed its performance by testing the impact of attribute weights generated by three different techniques, namely, feature counting, gradient descent, and genetic algorithm.

Regarding cost estimates using various methods, the question of how many and which attributes to use for estimation purposes is very critical, and research on these topics continues. This is because such an issue can affect the accuracy and efficiency of cost estimation in the computation process. Also, if all of the attributes are inputted and stored in cost databases, an enormous amount of time will be required for database construction and the size of the database will be significantly increased. Therefore, extracting several critical attributes is crucial in parametric or CBR cost estimations. Furthermore, as the weights of attributes reflect relative importance among attributes in relation to a project cost estimation, an appropriate weight-assigning method should be utilized to retrieve not simply similar but suitably similar past cases (Kim 2013).

Some researchers have adopted the intuitive approach, which relies on the experience of construction industry practitioners. However, this method, based on subjective aspects of experience and knowledge, involves a potentially less

reliable selection process of dominant attributes compared to other methods capable of quantifying attributes (Duverlie and Castelain 1999). Other research has utilized correlation coefficients, genetic algorithms, feature counting, regression coefficient significance, or analytic hierarchy processes to determine attribute weights and to extract influential attributes.

However, historical cost databases serve as the bases for cost-estimating models. The accuracy and diversity level of cases of the cost database, attribute weights, cost-estimating equations, and cost models can affect results to a certain level. The level of uncertainty associated with cost-estimating equations is also affected by the location of attributes within or outside of the range of attributes of the historical cost databases (Book 2012). If similar historical projects do not exist, the estimation results could be highly fluctuated over or under compared to the actual cost (Kim et al. 2012). Therefore, an attribute weight-assigning method that can consider the relationships between attributes and the variety level of the database apart from mere relationships between attributes and costs is required..

3.3.3 Similarity Measurement for Case Retrieval

To figure out which cases are nearest-neighbors to a target case, various similarity or distance measures have been adopted in a CBR system. Continuous studies on similarity measures applied to CBR cost models have been carried out as they have great influence on retrieval performance in a CBR system (Yau and Yang 1998; Doğan et al. 2006; An et al. 2007; Chou 2009; Kim and Kim 2010; Koo et al. 2010a; Ji et al. 2012; Ahn et al. 2014).

The Euclidean distance is the most commonly used measure, which is the ordinary distance of the line segment connecting between two points in Euclidean space. The distance is calculated as the square root of the sum of the squared differences between the values of each attribute (Pal and Shiu 2004; Ji et al. 2011b). Ji et al. (2011a) proactively employed a weighted Euclidean distance as the similarity-measuring method in developing a CBR-based military facility cost estimation system. Ji et al. (2011b) also proposed a CBR cost estimate model using Euclidean distance concept as similarity measure and genetic algorithms as weight assignment method. This research further compared the proposed model with combinations of other similarity measures and weight assignment methods, showing that the proposed model was more accurate than others using both one-NN (Nearest Neighbor) and Ten-NN. This paper also insisted the other several similarity measures such as arithmetic summation; fractional function (Burkhard 2001; Ozorhon et al. 2006; An et al. 2007; Ryu et al. 2007; Qian et al. 2009; Chou 2009) had limitations in lacking

of an explanation and in calculation when the target case does not exist inside the case-base range.

Kim and Kim (2010) suggested a CBR-based preliminary cost estimation model using genetic algorithms. Their research mainly focused on determining the significant weights of attributes to measure similarity and retrieve similar cases in the case-base with minimum prediction errors. However, to accurately measure similarity of attributes, not only do weight assigning methods need to be further examine but also similarity measurement methods. In their research, if attributes in character format were used, they assigned similarity score of 100 for a match and 0 for an unmatched. If attributes are represented numerically, where variation with the target case was less than a particular level of percentages, scores were given simply as 100, otherwise, 0. Jin et al. (2012) utilized multiple regression analysis technique in the revision phase of the CBR cost model for cost prediction in the early stage of business facility and multi-family housing projects. In calculation of similarities for the numerical attributes, similarity type of a minimum value of maximum value, which were adopted from Kim and Kang (2004), was utilized. Kim and Hong (2012) proposed a MRA-based (Multiple Regression Analysis) revised CBR cost estimation model for railroad-bridge construction projects. This research utilized relative differences of attributes for similarity measurement and allocated attribute similarity scores by granting 100 if the error rate is within 10%, 90 within 20%, 80 within 30%, and 0 over 31%.

The existing similarity measures such as Euclidean distance, arithmetic summation, and fractional function have been most commonly and widely adopted in CBR cost estimating to retrieve the most similar singular or plural cases from a case-base. However, these existing similarity measures have limitations in taking the correlations among attributes into consideration and reflecting the effects of covariance in computation of distances among attributes. In case-base or database, which is comprised of the attributes, there are very high possibilities of the existence of the covariance among attributes. As the CBR cost estimating is performed heavily based on the case-base, an inaccurate similarity measure among attributes might lead to the less similar retrievals of cases, and eventually mislead the results of cost estimates. Hence, under such circumstances, establishing an appropriate similarity function is necessary to deal with the hidden relationships among the attributes associated with cases (Burkhard 2001; Ji et al. 2011b). Therefore, to handle this challenging issues, a similarity measure that can reflect the covariance effects in distance calculation among attributes is required. Furthermore, as one of the most important issues of CBR is what kinds of similarity measurement would be utilized to accurately retrieve similar cases, the comparative examinations on existing distance measurements need to be further carried out in depth.

3.4 Summary

Human beings often face different cases, and when to do, they often relate before experiences that they or people whom they know have dealt with. The reason case-based reasoning is prevalent today is that people want to establish the comparison between two cases and weigh on several aspects. Using the CBR approach ensures that the estimations measure up to the current project, and the range is sensible.

Using the CBR approach helps in establishing individual aspects like cost, period, and materials. CBR approach helps in solving the problems that may exist by providing alternative remedies to the problems. The CBR strategy is efficient and allows the learner to probe into different situations that took place and the interventions that took place. Relating to the past to solve the current problem is an indication that the learner can open up their minds and reasoning. CBR strategy provides easier acquisition of knowledge compared to rule-based system learning.

One outstanding feature of using the CBR approach is that the previous mistakes made will not re-occur, and the learner will be able to predict the outcomes. CBR provides an upper hand on the matters related to past issues as people can decide on steps to take and the expected outcome. The experience gained with the CBR approach is gradual, and the learner is eventually sharpened to deliberate on minor and major issues.

The ability to reason some solutions or issues that have not yet emerged is a feature of CBR approach. Learners can open up their minds to possible outcomes that may affect the outcome positively or negatively. It is clear that case-based reasoning is an effective approach to empower learners with the right problem-solving capacity in different areas.

As much as the CBR approach is termed as useful in reasoning, there are setbacks associated with the case-based reasoning methodology. This method denies the user the ability to be accurate in all dimensions. The ideas borrowed from other situations may be a product of wrong perceptions, and this may eventually not give the desired result. The speculation on the comparisons on weight, quality and other aspects is a dilemma faced with CBR methodology. The dilemma in the ways of organizing the case-base and overlapping the cases is a major challenge. Weighing the attributes, measuring the similarities, representing the case-base, and maintaining the characteristics in solutions is another setback of CBR methodology. Based on the information provided on the benefits and setbacks of the CBR approach, learners can design the most suitable reasoning environment to solving problems and achieving the given objectives.

Chapter 4. CBR Model Design Experiment for Improving Cost Estimation

Since the applied methods and process for developing CBR cost models are different across researchers (Yau and Yang 1998; Schirmer 2000; An et al. 2007; Chou 2009; Ji et al. 2011b; Ahn et al. 2014), details of the CBR cost model components such as normalization, similarity measurement, attribute weight assignment methods need to be elaborated to particularly satisfy the research aims. This chapter conducts comparative experiment for 1) normalization method, 2) attribute weighting method, and 3) similarity measurement methods. Furthermore, various validation methods and processes are performed to analyze and establish an effective validation procedures. The results and learnings from the CBR model design experiments in Chapter 4 are sourced as examination and verification of CBR model components and validation procedures for developing a front-end cost estimation methodology by selective CBR in Chapter 5.

4.1 Normalization Method and Accuracy

4.1.1 Normalization Issue

A quality case-base is accomplished through data preprocessing, which is a preliminary process that prepares secondary data to identify the relationships and available patterns hidden by a large quantity of data (Kotsiantis et al. 2006).

Furthermore, data preprocessing is a standard practice of data mining used for normalization, denoising internal errors or abnormal values (Shalabi et al. 2006). Among the data preprocessing, it is always important first to ensure normalization or standardization in CBR cost estimation in order to achieve the desired high levels of cost estimate accuracy (Koo et al. 2010a). Normalization refers to the adjustment of all collected data into standardized values assigned to specific ranges such as 0.0 to 1.0 in order to allow for comparison of corresponding standardized values for different datasets in a way that eliminates the effect of certain gross influences (Han and Kamber, 2006). Apparently, this is essential because CBR cost estimation is more accurate when all attributes are analyzed under identical standards; and is the same case with the evaluation of attribute and case similarity (Watson 1997; An 2007).

4.1.2 Comparative Experimental Design

To carry out in-depth research on the effects of normalization methods applied to CBR cost models, this research initiates by formulating two hypotheses: 1) “CBR cost estimation accuracy and stability can be improved by employing statistically accurate normalization methods.” 2) “A CBR cost model has appropriate normalization methods.” To examine and confirm the above two hypotheses, this paper conducts comparative research (Figure 4-1) on CBR cost estimation models based on five different normalization methods that are interval standardization, Gaussian distribution-based normalization, z-score normalization, logistic function-based normalization, and ratio

standardization. Also, Euclidean distance as a similarity measurement and genetic algorithms as an attribute weight assigning method are utilized in these models. The cost models are used in the early design stages (i.e., conceptual) and the cost data is constructed using public multi-family housings.

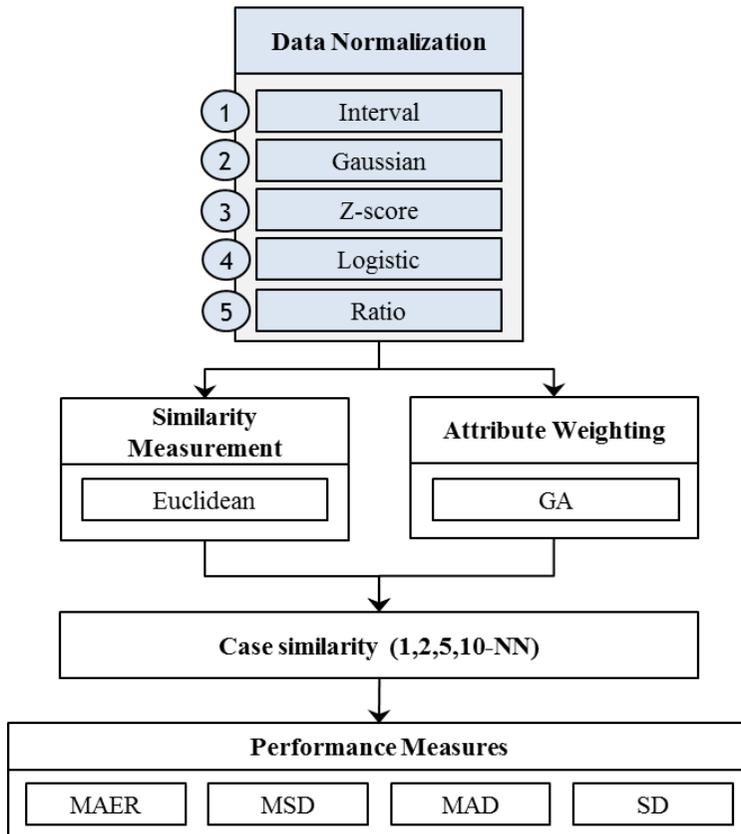


Figure 4-1. Experimental Design for Normalization Method

4.1.3 Results and Discussions

As shown in Figure 4-2 and Table 4-1, MAER displayed some distinctive characteristics that arose from the different normalization methods. MAER is calculated using the z-score method denoted negative values, whereas the other normalization methods use positive values. The ratio normalization method had the lowest MAER values when $k=1, 2, 5,$ and 10 as compared to other methods. As lower values of MAER indicate a higher accuracy of the CBR model, the ratio standardization based CBR cost model appeared to be more accurate than other methods.

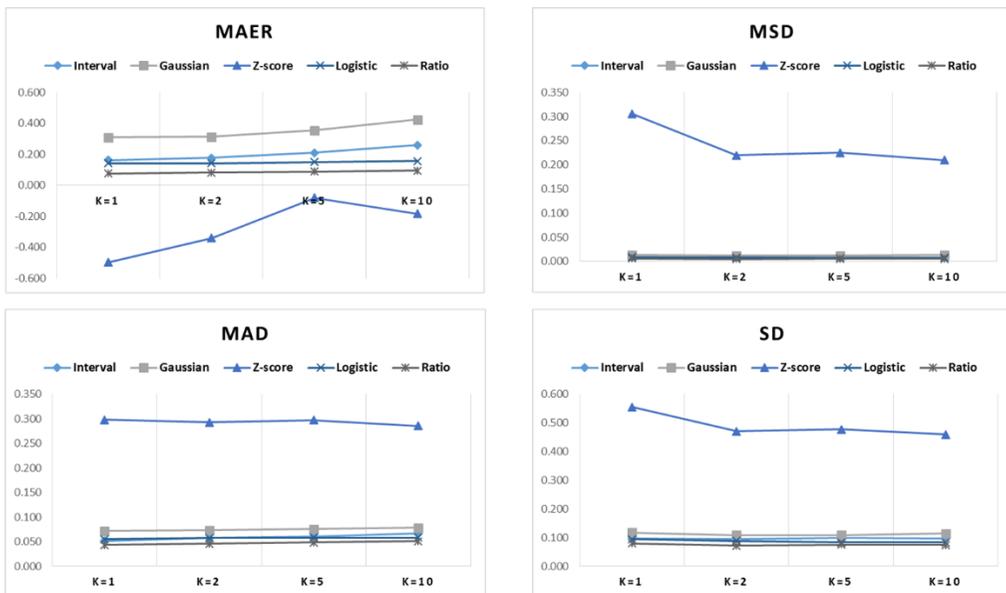


Figure 4-2. Results of MAER, MSD, MAD, and SD for Normalization

Table 4-1. Results of MAER for Normalization Method

MAER	k=1	k=2	k=5	k=10
Interval	0.163	0.177	0.211	0.259
Gaussian	0.308	0.313	0.355	0.425
Z-score	-0.497	-0.340	-0.080	-0.183
Logistic	0.141	0.142	0.150	0.157
Ratio	0.076	0.082	0.089	0.095

Other patterns of results were obtained from the MSD (Figure 4-2 and Table 4-2). The results using the z-score method were plotted on the positive plane while the other values were plotted immediately above the ‘0’ point. Interval, Gaussian, logistic, and ratio normalization based CBR cost models had very low MSD, which means that these methods had high preciseness.

Table 4-2. Results of MSD for Normalization Method

MSD	k=1	k=2	k=5	k=10
Interval	0.009	0.009	0.010	0.009
Gaussian	0.014	0.012	0.012	0.013
Z-score	0.306	0.220	0.225	0.210
Logistic	0.009	0.008	0.007	0.007
Ratio	0.006	0.005	0.006	0.005

A similar trend was obtained when the MAD was computed from the results (Figure 4-2 and Table 4-3). In data analysis, the MAD value is significant and denotes the average distance for every element in a set of data from the entire mean of the whole set. As compared to the other normalization methods, the values established using the z-score model were higher, which ranged approximately from 0.286 to 0.298 for all values of k. The MAD values resulting from the other normalization methods ranged from 0.043 to 0.078 for 1, 2, 5, and 10-NN. The least MAD values were obtained using the ratio normalization method. Overall, except for the z-score method, other normalization methods had relatively low MAD values, which mean that these methods had very narrow ranges of estimate errors.

Table 4-3. Results of MAD for Normalization Method

MAD	k=1	k=2	k=5	k=10
Interval	0.051	0.057	0.061	0.067
Gaussian	0.072	0.073	0.076	0.078
Z-score	0.298	0.293	0.297	0.286
Logistic	0.055	0.057	0.058	0.058
Ratio	0.043	0.046	0.049	0.051

The results obtained from the computation of the SD were relatively stable using the Gaussian, interval, logistic, and the ratio normalization-based CBR models, which ranged from 0.071 to 0.100 for all values of k (Figure 4-2 and

Table 4-4). However, the z-score model yielded relatively high values of SD. The results established from the z-score model suggested that the dataset values were dispersed over a wider range. On the contrary, the other models denoted that the data sets were closer to the mean, which represented more stable CBR-based cost estimate models.

Table 4-4. Results of SD for Normalization Method

SD	k=1	k=2	k=5	k=10
Interval	0.097	0.095	0.100	0.097
Gaussian	0.117	0.109	0.109	0.114
Z-score	0.556	0.471	0.477	0.459
Logistic	0.094	0.089	0.083	0.083
Ratio	0.080	0.071	0.075	0.074

To summarize, in terms of the estimate accuracy, ratio normalization yielded a relatively lower MAER, MSD, and MAD in comparison to the interval normalization, Gaussian distribution-based normalization, z-score normalization, and logistic function-based normalization. Since the lower MAER, MSD, and MAD indicate that the more accurate and precise cost estimate results are obtained by the selected normalization methods, the ratio normalization method is considered to be relatively more accurate in retrieving similar cases. Regarding estimate stability, a relatively lower SD for the ratio standardization compared to the other normalization methods was obtained.

Consequently, the results verified the first hypothesis that CBR cost estimation accuracy and stability can be improved by employing statistically accurate normalization methods. Also, in terms of the appropriate selection of normalization methods, the kernel density estimation results showed that interval and ratio normalization can be appropriate methods to be used. Thus, this experiment results satisfied the second hypothesis that a CBR cost model has appropriate normalization methods.

4.2 Attribute Weighting Method and Accuracy

4.2.1 Attribute Weighting Issue

As an effort to improve estimate accuracy during initial stages, researchers continuously develop various estimation methods. Among these, the most commonly used method, parametric estimation, utilizes attributes developed from historical cost databases and construction practitioners (Meyer and Burns 1999). Using this parametric method, many cost estimation models have resulted, in many cases by adopting the regression analysis (Trost and Oberlender 2003; Soutos and Lowe 2005; Seong et al. 2008; Ji et al. 2010). As well as the parametric method, researchers have explored case-based reasoning (CBR), which utilizes knowledge obtained from past experience, and the construction industry has applied it as an estimation method for construction costs (Yau and Yang 1998; Karshenas and Tse 2002; An et al. 2007; Chou 2009; Ji et al. 2010b).

When using the parametric or CBR methods to estimate cost, extracting influential attributes on cost and measuring them in a quantitative manner plays a vital role. However, numerous attributes affect cost, and using these to estimate decreases the accuracy of early cost estimation and complicates the efficiency of the estimate process. Therefore, it becomes necessary to measure and prioritize the impact of attributes in the order of high cost implications. To prioritize attributes, those involved commonly use various methods to calculate the weights of attributes, including analytic hierarchy process (AHP), genetic algorithm (GA), principal component analysis (PCA), feature counting (FC), and correlation analysis (CA).

However, a review of the literature reveals that the calculation process of determining weights by these methods occurs without considering the possible range of each attribute. In general, parametric or CBR-based cost estimation models are based on historical databases. Depending on the accuracy and diversity level of the cases of the database, attribute weights, cost-estimating equations, and a cost model itself can change to a certain degree. Therefore, the weights of attributes should be derived considering not only the relationships between attributes and costs, but also the database comprising a range of various quality buildings. To deal with this challenging issue, this research introduces Attribute Impact (AI), which can quantitatively measure the weights of attributes.

4.2.2 Concept of Attribute Impact

Impulse-momentum theorem

How does the impact affect the motion of each vehicle when two automobiles collide? What mechanisms can be used to overcome the impact to prevent serious injury? To answer such questions, physics introduces momentum and impulse. The momentum \vec{p} of an object of mass m moving with velocity \vec{v} is defined by the product of an object's mass and velocity. Therefore momentum will depend on an object's mass and velocity.

$$\vec{p} = m\vec{v} \quad (\text{Eq. 4-1})$$

The magnitude of the momentum p of an object of mass m can be related to its kinetic energy KE:

$$KE = \frac{1}{2}mv^2 = \frac{m^2v^2}{2m} = \frac{p^2}{2m} \quad (\text{Eq. 4-2})$$

The application of a force is required to explain the momentum changing. This is originally stated in Newton's second law of motion—the conservation of energy:

$$\vec{F} = m\vec{a} = m \frac{\Delta\vec{v}}{\Delta t} = \frac{\Delta(m\vec{v})}{\Delta t} \quad (\text{Eq. 4-3})$$

Based on this relationship the impulse is defined as written below in order to measure an object's change in momentum (Serway and Vuille 2012).

$$\vec{I} = \Delta\mathbf{p} = \vec{F}\Delta t = \Delta m\vec{v} = m(\mathbf{v}_i - \mathbf{v}_f) = m\left(\frac{\Delta\mathbf{r}_i}{\Delta t} - \frac{\Delta\mathbf{r}_f}{\Delta t}\right) \quad (\text{Eq. 4-4})$$

Where \mathbf{v}_i and \mathbf{v}_f denote the initial and final velocities of an object; and $\Delta\mathbf{r}_i$ and $\Delta\mathbf{r}_f$ denote the initial and final displacements of an object in the time interval Δt .

Attribute Impact Concept

Based on above-discussed impulse-momentum theorem of physics, the concept of attribute impact was developed. As discussed above, impulse can be used to estimate the force exerted during the impact. If all of the force of an object is consumed and delivered to another place ($\mathbf{v}_f = 0$), the impulse is in proportion to the mass and initial velocity. If the time interval Δt is assumed to be constant, the initial velocity is proportional to the displacement. By adopting this relationship, this research substitutes the variables of impulse-momentum theorem with equivalents: the mass of an object is substituted by the weight of the attribute, and the displacement by the range which is the variation between maximum and minimum values. Hereafter, Attribute Impact (AI_i) is defined as equal to Attribute Weight (AW_i) multiplied by Attribute Range (AR_i).

$$\text{Attribute Impact } (AI_i) = AW_i \times AR_i \quad (\text{Eq. 4-5})$$

where AI denotes attribute impact, AW denotes attribute weight, AR denotes attribute range, and i is the index of the i^{th} case project.

This proposed AI can be emphasized by applying the architectural interpretation. With momentum directly proportional to the mass and displacement of an object, it can be also assumed that AI is proportional to the weights and ranges of attributes. In general, historical cost databases containing information of various attributes serve as the basis for cost-estimation models. However, significantly, the numbers of cost data accumulated to build the database can vary across each cost model, and because cost data was manually inputted into databases, the accuracy of the databases can differ. If the accuracy and diversity level of the database vary, the range of each attribute generally changes. Ultimately, these can affect the weights of attributes, the cost-estimating equations, and the cost model itself, to a certain degree. Therefore, if a computing process to quantify attribute weights can consider not only the relationships between attributes and costs but also the database comprising a range of various-quality buildings built for research, then weights of attributes can be reflected the more accurately; and eventually estimate results can be improved as well.

Computing Process

To calculate the values of AI, weights and ranges of attributes need to be figured out, as defined in Eq. 4-5. To obtain the weights of attributes, this research adopts correlation coefficients due to its relatively simple computation process and easiness to interpret compared to other attribute weight-assigning methods such as regression coefficient, GA, and AHP. The quantification of a correlation is usually executed by specifying the correlation coefficient. In statistics, a correlation refers to the departure of two variables from independence, while the correlation coefficient indicates the strength and direction of a linear relationship between two random variables (Fellows and Liu 2003). Measuring the degree of correlation involves several coefficients. The best-known coefficient, the Pearson product-moment correlation coefficient, by dividing the covariance of two variables by the product of their standard deviations is obtained. The correlation coefficient R is defined as

$$R = \frac{1}{n-1} \sum \frac{x_i - \mu_x}{\sigma_x} \frac{y_i - \mu_y}{\sigma_y} \quad (\text{Eq. 4-6})$$

where μ_x and σ_x denote the sample mean and the sample standard deviation, respectively, for the variable x; and μ_y and σ_y denote the sample mean and the sample standard deviation, respectively, for the variable y. Generally, a correlation coefficient of under -0.5 or over 0.5 indicates that the two variables have a strong correlation. Table 4-5 shows the results of the correlation analysis between total cost and attributes for each unit type.

Table 4-5. Correlation Analysis (Pearson Correlation Coefficient)

Type	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
Type 49	0.92	0.95	0.40	-0.22	0.47	0.44	0.42	-0.53	-0.34	0.37	0.46
Type 59	0.96	0.98	0.76	0.11	0.53	0.33	0.67	-0.19	-0.02	0.50	-0.31
Type 84	0.97	0.98	0.90	0.41	0.79	0.56	0.70	-0.71	-0.03	0.62	-0.68
Type 114	0.97	0.97	0.93	0.86	0.85	0.58	0.26	-0.49	0.24	0.28	-0.14

Note: (X1) Number of households, (X2) Gross floor area, (X3) Number of unit floor households, (X4) Number of elevators, (X5) Number of floors, (X6) Number of piloti with household scale, (X7) Number of households of unit floor per elevator, (X8) Height between stories, (X9) Depth of pit, (X10) Roof type, (X11) Hallway type

The range of an attribute by subtracting the minimum value of each attribute from the database from the maximum value is obtained. Consequently, the Attribute Impact (AI_i) is derived by multiplying a correlation coefficient (CC_i) by an attribute's range of data (AR_i), as below (Eq. 4-7). Table 4-6 summarizes derived values of AIs, correlation coefficients, and ranges. Furthermore, the eleven attributes are ranked under the obtained values of AIs, and correlation coefficients, respectively, to compare how the order or ranking of the attributes are changed (Table 4-7).

$$AI_i = CC_i \times AR_i \quad (\text{Eq. 4-7})$$

where AI denotes attribute impact, CC denotes correlation coefficients, AR denotes attribute range, and i is the index of the ith case project.

Table 4-6. Derived Value of Attribute Impact & Correlation Coefficient

	Type	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
49	CC	0.92	0.95	0.40	0.22	0.47	0.44	0.42	0.53	0.34	0.37	0.46
	Range	42.00	2,865.47	4.00	1.00	8.00	6.00	6.00	0.10	4.16	1.00	1.00
	PI	38.78	2,720.98	1.61	0.22	3.78	2.66	2.51	0.05	1.43	0.37	0.46
59	CC	0.96	0.98	0.76	0.11	0.53	0.33	0.67	0.19	0.02	0.50	0.31
	Range	58.00	4,797.18	4.00	2.00	12.00	8.00	3.00	0.10	4.16	1.00	1.00
	PI	55.55	4,680.90	3.03	0.22	6.40	2.64	2.01	0.02	0.09	0.50	0.31
84	CC	0.97	0.98	0.90	0.41	0.79	0.56	0.70	0.71	0.03	0.62	0.68
	Range	51.00	5,633.99	3.00	1.50	11.00	6.00	2.00	0.10	4.36	1.00	1.00
	PI	49.57	5,521.55	2.71	0.62	8.67	3.36	1.40	0.07	0.13	0.62	0.68
114	CC	0.97	0.97	0.93	0.86	0.85	0.58	0.26	0.49	0.24	0.28	0.14
	Range	51.00	7,231.94	3.00	1.50	10.00	8.00	1.00	0.10	3.60	1.00	1.00
	PI	49.38	7,006.80	2.78	1.28	8.53	4.64	0.26	0.05	0.86	0.28	0.14

Note: CC: Correlation Coefficient, Range = Maximum value – Minimum value, PI: Parameter Impact, (X1) Number of households, (X2) Gross floor area, (X3) Number of unit floor households, (X4) Number of elevators, (X5) Number of floors, (X6) Number of piloti with household scale, (X7) Number of households of unit floor per elevator, (X8) Height between stories, (X9) Depth of pit, (X10) Roof type, (X11) Hallway type

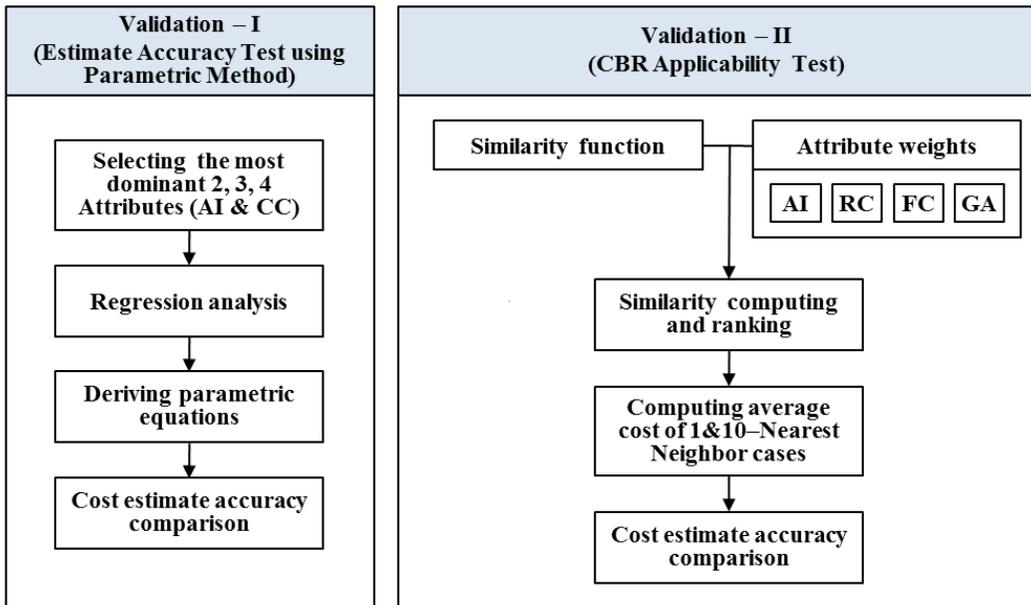
Table 4-7. Ranking Comparison between AI & CC

Ranking	Type 49		Type 59		Type 84		Type 114	
	PI	CC	PI	CC	PI	CC	PI	CC
1	X2	X2	X2	X2	X2	X2	X2	X2
2	X1	X1	X1	X1	X1	X1	X1	X1
3	X5	X8	X5	X3	X5	X3	X5	X3
4	X6	X5	X3	X7	X6	X5	X6	X4
5	X7	X11	X6	X5	X3	X8	X3	X5
6	X3	X6	X7	X10	X7	X7	X4	X6
7	X9	X7	X10	X6	X11	X11	X9	X8
8	X11	X3	X11	X11	X10	X10	X10	X10
9	X10	X10	X4	X8	X4	X6	X7	X7
10	X4	X9	X9	X4	X9	X4	X11	X9
11	X8	X4	X8	X9	X8	X9	X8	X11

Note: CC: Correlation Coefficient, PI: Parameter Impact, (X1) Number of households, (X2) Gross floor area, (X3) Number of unit floor households, (X4) Number of elevators, (X5) Number of floors, (X6) Number of piloti with household scale, (X7) Number of households of unit floor per elevator, (X8) Height between stories, (X9) Depth of pit, (X10) Roof type, (X11) Hallway type

4.2.3 Comparative Experimental Design

Two types of validation were designed to examine the reliability of the AI (Fig. 4-3). Validation-I, an estimate accuracy test, uses the parametric method. Parametric cost estimation utilizes one or several attributes that have high-cost implications to provide reliable and accurate cost predictions. Multiple regression analysis was conducted to derive cost-estimation relationships between dependent attributes and to develop parametric equations. Depending on the selected attributes for multiple regression analysis, parametric equations with different levels of estimate accuracy can result.



Note – AI: Attribute Impact, RC: regression coefficient, FC: feature counting, GA: genetic algorithm

Figure 4-3. Experimental Design for Attribute Weighting Methods

Therefore, the design of validation-I allows for the comparison of the estimate accuracy based on the parametric equations using the most dominant three attributes obtained by AI and correlation analysis, respectively.

66 building-cost data from type 84m² households between the fifth and fifteenth stories was utilized for multiple regression analysis; and validation-I procedures as follows: 1) obtain two, three, four high-ranked attributes (gross floor area, number of unit floor households, number of floors, height between stories) by correlation analysis for multiple regression analysis to derive cost-relationship equations; 2) obtain two, three, four high-ranked attributes (gross floor area, number of floors, number of pilotis with household scale, number of unit floor households) from AI for multiple regression analysis to derive cost-relationship equations; and 3) select the costs of 10 test cases of type 84m² (Table 4-8), by random sampling estimated using each derived cost-relationship equation (Table 4-9), and compare to actual costs.

Additionally, the estimate results was compared to traditional unit cost methods: cost per gross floor area (\$/GFA). The estimate results are represented and compared with their absolute error ratios (AER) utilized by Ji et al. (2011b), which can be defined as below (Eq. 4-8). C_A and C_E denote actual cost and estimated cost, respectively.

$$AER(\%) = \begin{cases} \text{if } C_A - C_E > 1, & \text{then } [(C_A - C_E) - 1] \times 100 \\ \text{otherwise,} & [1 - (C_A - C_E)] \times 100 \end{cases} \quad (\text{Eq. 4-8})$$

Note that Table 4-10 identifies the existence of multicollinearity from the result of correlation analysis. We must treat multicollinearity, the state caused by strong intercorrelation among the attributes that disturbs the reliability of the statistical inference. Table 4-10 reveals the most dominant three attributes as gross floor area, number of households, and number of unit floor households. However, we detect that the number of households has a stronger correlation of 0.99 with gross floor area than cost (0.97). Thus, if we used the three attributes including number of households for multiple regression analysis, multicollinearity is highly likely to exist and the estimate results will be considered statistically unreliable. To deal with the multicollinearity issue, we eliminate number of households and include number of floors in the most dominant three attributes. Additionally, we select gross floor area, number of floors, and number of pilotis with household scale as the most dominant three attributes of AI by applying the same method mentioned above.

Table 4-8. Profile of Test Cases

Type 84	Attribute											Total Cost
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	
A	13.00	1,416	2.00	1.00	7.00	1.00	2.00	2.90	7.55	-	1.00	970,004,637
B	13.00	1,423	2.00	1.00	7.00	1.00	2.00	2.90	7.55	-	1.00	942,560,605
C	22.00	2,420	2.00	1.00	11.00	-	2.00	2.80	9.36	-	1.00	1,495,622,168
D	44.00	3,368	4.00	2.00	12.00	4.00	2.00	2.80	5.00	-	1.00	2,094,424,957
E	40.00	4,448	4.00	1.00	11.00	4.00	4.00	2.80	5.85	1.00	-	2,894,608,265
F	44.00	4,890	4.00	2.00	12.00	2.00	2.00	2.80	9.36	-	1.00	3,038,402,053
G	50.00	5,500	4.00	1.00	13.00	2.00	4.00	2.80	5.85	1.00	-	3,220,935,698
H	52.00	5,750	4.00	1.00	15.00	4.00	4.00	2.80	5.85	1.00	-	3,694,260,409
I	56.00	6,146	4.00	1.00	15.00	4.00	4.00	2.80	5.85	1.00	-	3,881,454,958
J	56.00	6,189	4.00	1.00	15.00	4.00	4.00	2.80	8.60	1.00	-	3,654,859,216

Note: (X1) Number of households, (X2) Gross floor area, (X3) Number of unit floor households, (X4) Number of elevators, (X5) Number of floors, (X6) Number of pilotis with household scale, (X7) Number of households of unit floor per elevator, (X8) Height between stories, (X9) Depth of pit, (X10) Roof type, (X11) Hallway type.

Table 4-9. Parametric Equations Using Multiple Regression Analysis

Category	Regression Equations	Adjusted R²
2-attribute	AI =663108.42*X2 – 22905916.04*X5+148724617.74	0.96
	CC =561525.08*X2+126672260.76*X3 – 115231144.51	0.96
3-attribute	AI =646499.98*X2 - 20959123.90*X5+27855689.39*X6+133778413.54	0.96
	CC =559762.58*X2+128301527.44*X3+626751.83*X5 - 120729721.01	0.96
4-attribute	AI =568642.63*X2 – 2789602.94*X5+21164805.20*X6+101610796.01*X3 – 76030943.14	0.96
	CC =577516.26*X2+88483267.56*X3 – 8032506.65*X5 – 899770471.10*X8+2579816336.95	0.96

Note: CC: Correlation Coefficient, AI: Attribute Impact, (X2) Gross floor area, (X3) Number of unit floor households, (X5) Number of floors, (X6) Number of pilotis with household scale, (X8) Height between stories

Table 4-10. Correlation Analysis (Type 84)

		Type 84									
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
X1	1.00										
X2	0.99	1.00									
X3	0.92	0.90	1.00								
X4	0.44	0.40	0.58	1.00							
X5	0.82	0.82	0.58	0.23	1.00						
X6	0.55	0.54	0.59	0.14	0.41	1.00					
X7	0.70	0.71	0.66	-0.22	0.49	0.59	1.00				
X8	-0.72	-0.71	-0.72	-0.39	-0.62	-0.48	-0.50	1.00			
X9	-0.06	-0.04	-0.11	-0.01	0.09	-0.25	-0.13	0.21	1.00		
X10	0.64	0.65	0.58	-0.24	0.51	0.51	0.92	-0.38	0.00	1.00	
X11	-0.68	-0.69	-0.64	0.22	-0.47	-0.59	-0.97	0.49	0.11	-0.89	1.00
cost	0.97	0.98	0.90	0.41	0.79	0.56	0.70	-0.71	-0.03	0.62	-0.68

Note: (X1) Number of households, (X2) Gross floor area, (X3) Number of unit floor households, (X4) Number of elevators, (X5) Number of floors, (X6) Number of pilotis with household scale, (X7) Number of households of unit floor per elevator, (X8) Height between stories, (X9) Depth of pit, (X10) Roof type, (X11) Hallway type.

Validation-II represents an AI-applicability test to the CBR estimation method. Measuring and scoring the similarity of the retrieved cases in CBR-based cost estimation requires the weights of attributes and a similarity function. It is crucial to have accurate and reliable attribute weights and similarity function, as the accuracy of similarity measurement depends highly on them. After a closer examination of the similarity function, we find that Euclidean distance—the square root of the sum of the square of the arithmetical

differences between two corresponding objects—is the most commonly used distance-measuring method (Pal and Shiu 2004). Hence, this research adopts the similarity-measuring function based on the Euclidean distance concept employed by Ji et al. (2011b) as below (Eq. 4-9). $SIM(x_i, x_j)$ represents the degree of similarity between x_i and x_j , and $DIS(x_i, x_j)$ signifies the weighted distance between the two cases x_i and x_j . $a_r(x)$ indicates the value of the r^{th} attributes of case x , and w_r denotes the weight of the case's attributes.

$$SIM(x_i, x_j) = 1 - DIS(x_i, x_j) = 1 - \sqrt{\sum_{r=1}^n w_r^2 (a_r(x_i) - a_r(x_j))^2} \quad (\text{Eq. 4-9})$$

The similarity measurement also requires attribute weights apart from the similarity function mentioned above. The specially designed validation-II examines the validity of the suggested AI as a means of assigning the weights of attributes. We compare CBR-based cost estimation using the weights of AI to that of standardized regression coefficient (RC), feature counting (FC), and genetic algorithm (GA). Table 4-11 summarizes the weights of the attributes obtained using AI, RC, FC, and GA. Consequently, we compare absolute error ratios (AER) of 1-NN (K-Nearest Neighbors) and 10-NN using each attribute weight.

Table 4-11. Weights of the Attributes Obtained by AI, RC, FC, and GA

Method	Attribute										
	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
AI	67.52	7,433.60	4.87	0.97	9.48	3.88	2.73	-0.05	0.07	0.43	-0.41
RC	0.00654	0.93701	-0.1615	0.13886	-0.0482	0.04609	0.11363	-0.0929	0.01528	0.07564	0.05241
FC	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
GA	0.019	0.361	0.176	0.004	0.292	0.007	0.041	0.002	0.006	0.007	0.024

Note: AI: Attribute Impact, RC: Standardized Regression Coefficient, FC: Feature Counting, GA: Genetic Algorithm, (X1) Number of households, (X2) Gross floor area, (X3) Number of unit floor households, (X4) Number of elevators, (X5) Number of floors, (X6) Number of pilotis with household scale, (X7) Number of households of unit floor per elevator, (X8) Height between stories, (X9) Depth of pit, (X10) Roof type, (X11) Hallway type.

4.2.4 Results and Discussions

As shown in Table 4-12, in terms of the estimate accuracy using the parametric method, relatively lower overall absolute error ratios (AER) resulted from using the suggested AI. For the average AER of 10 test cases using two, three, and four attributes obtained by AI are 2.75%, 2.88% and 3.29%; and by correlation coefficients are 3.22%, 3.24% and 4.04%, respectively. For the traditional unit cost method of cost per gross floor area (\$/GFA), 4.04% is the result. In terms of the estimate stability, we achieve stable estimates using the attributes derived by AI since we reached low standard deviations compared to that of correlation coefficients.

As described earlier in Table 4-7, we can measure the weights of the attributes differently when considering the ranges of the attributes. Furthermore, the validation results support that AI can provide alternatives in selecting attributes in quantitative measure when the parametric cost estimate method is used. However, we found no statistically significant difference between means after carrying out one-way ANOVA. Additionally noteworthy, the cost estimation using \$/GFA also resulted in a relatively low AER. First, we consider this to be due to the distinct characteristic of the standardized apartment buildings. Second, among the sample data used in the research, quite a large number of sample data have a gross floor area similar to that of the test cases.

Table 4-12. Comparison of Absolute Error Ratio (AER)

Case	<u>2-Attribute</u>		<u>3-Attribute</u>		<u>4-Attribute</u>		\$/GFA
	AI	CC	AI	CC	AI	CC	
A	4.62%	3.95%	4.28%	3.99%	3.87%	6.73%	10.86%
B	1.15%	0.59%	0.84%	0.63%	0.50%	3.25%	7.20%
C	0.39%	0.09%	1.90%	0.12%	1.56%	3.30%	0.00%
D	0.61%	8.25%	3.53%	8.35%	8.81%	7.45%	0.63%
E	1.70%	0.19%	0.15%	0.18%	0.66%	0.01%	5.30%
F	2.50%	3.15%	1.97%	3.15%	2.61%	3.30%	0.54%
G	7.93%	7.45%	7.26%	7.43%	7.02%	7.62%	5.24%
H	2.11%	2.04%	1.26%	2.04%	1.41%	2.20%	3.96%
I	0.01%	1.01%	0.59%	1.02%	0.34%	0.99%	2.18%
J	6.51%	5.48%	7.05%	5.47%	6.11%	5.52%	4.44%
Mean	2.75%	3.22%	2.88%	3.24%	3.29%	4.04%	4.04%
S.D	2.71%	3.00%	2.59%	3.01%	3.04%	2.68%	3.39%

Note: CC: Correlation Coefficient, AI: Attribute Impact.

Regarding the CBR-applicability issue of AI, as summarized in Table 4-13, CBR-cost estimation using the weight of AI resulted in relatively higher estimate accuracy for both 1-NN and 10-NN. An overall AER of 1-NN and 10-NN for AI was 4.01% and 4.34%; for RC, 3.41% and 5.28%; for FC 8.69% and 6.46%; and for GA, 5.08% and 6.93%; respectively. Significantly, the lowest and highest AER of the average cost based on the 10 most similar cases (10-NN) was 0.01% and 9.95% for AI, 0.08% and 10.15% for RC, 1.41% and 24.54% for FC, and 0.25% and 19.27% for GA; respectively. Regarding estimate stability, we obtained a relatively low standard deviation for both 1-NN and 10-NN using AI. To summarize, the CBR-applicability test verified that the weights of attributes derived from AI can apply to CBR cost estimation as we achieve an overall lower AER than when using other widely studied weighting methods such as regression coefficient, FC, and GA.

The advantages and disadvantages of the suggested AI method can be summarized as follows: Firstly, its computing process is relatively simpler and more recognizable than other methods that require statistical transformation. Next, the AI method is flexible as it can reflect the ranges of attributes according to the accumulated database. However, when additional data are added to cost databases, the weights of attributes need to be updated. Other weight assigning methods should deal with this matter as well; hence, a method to automatically update the weights of attributes is required for the future research. Also, the logics of the AI should be further validated for the generalization.

Table 4-13. CBR Applicability Test

Case	Parameter Impact		Regression Coefficient		Feature Counting		Genetic Algorithm	
	1-NN	10-NN	1-NN	10-NN	1-NN	10-NN	1-NN	10-NN
A	2.91%	7.09%	2.91%	10.15%	2.91%	5.52%	2.91%	19.27%
B	2.83%	9.95%	2.83%	6.71%	2.83%	8.44%	2.83%	15.51%
C	11.31%	1.13%	2.93%	8.10%	49.33%	1.93%	11.31%	9.35%
D	2.48%	5.59%	6.42%	1.58%	9.88%	24.54%	14.72%	2.19%
E	4.26%	7.94%	4.26%	7.06%	4.26%	7.43%	4.26%	6.46%
F	3.90%	0.96%	2.31%	0.08%	5.19%	1.54%	2.31%	3.31%
G	9.45%	0.01%	9.45%	8.28%	9.45%	3.16%	9.45%	6.15%
H	0.06%	3.17%	0.06%	4.09%	0.06%	2.66%	0.06%	0.25%
I	2.20%	7.24%	2.20%	5.54%	2.20%	8.01%	2.20%	5.86%
J	0.75%	0.35%	0.75%	1.23%	0.75%	1.41%	0.75%	0.93%
Average	4.01%	4.34%	3.41%	5.28%	8.69%	6.46%	5.08%	6.93%
S.D	3.61%	3.65%	2.75%	3.41%	14.65%	6.92%	4.96%	6.24%

Note: (X1) Number of households, (X2) Gross floor area, (X3) Number of unit floor households, (X4) Number of elevators, (X5) Number of floors, (X6) Number of piloti with household scale, (X7) Number of households of unit floor per elevator, (X8) Height between stories, (X9) Depth of pit, (X10) Roof type, (X11) Hallway type

4.3 Similarity Measurement Method and Accuracy

4.3.1 Covariance Effect Issue

Although CBR provides a simple and plain methodology frame, there are many challenging issues such as distance measure, attributes selection, weights assignment, threshold of reuse cases, etc. (Aamodt and Plaza 1994; Watson 1997; Arditi and Tokdemir 1999; Pal and Shiu 2004; koo et al. 2010b; Ji et al. 2012). Specifically, distance measurement method is still a crucial problem that must be proven logically. In this regards, a large number of the studies on distance measurement methods have been carried out and Euclidean distance (Mitchell 1997; Pal and Shiu 2004; Ji et al. 2012), arithmetic summation (Ahn et al. 2006; Ryu et al. 2007), fractional function (Burkhard 2001; Qian et al. 2009) based similarity measurement have been widely utilized and applied to CBR retrieval.

However, past researches have rarely considered the impact of correlation of attributes when they calculate the similarity. Truly, all attributes are correlated, thus undesired influence provoked by attributes covariance would affect the reliability of cost estimation results. Therefore, the influence of covariance among attributes should be considered and calculated when measuring the similarity (Mahalanobis 1936; Farrar and Glauber 1967; Du and Bormann 2012). In this context, this research aims to examine the weighted Mahalanobis distance concept, which accounts for covariance among attributes, when it is applied to CBR cost estimation. Therefore, this paper carries out a

comparative research on various similarity measurement methods applied to CBR cost models and their cost estimation results in terms of estimate accuracy and stability.

4.3.2 Comparative Experimental Design

To verify the degree of covariance effects of similarity measures, this research further proposes the weighted Mahalanobis distance which can reflect covariance effects existing among the attributes. In order to examine how the suggested weighted Mahalanobis distance based similarity measurement is different from other previously adopted distance measures in terms of accuracy, stability, and propriety, validation methods and process for comparative analysis is designed as illustrated in Figure 4-4.

The framework of the experiment can be divided into two sections. The first part is a theoretical examination on similarity measurement methods based on simulation data tests. Three different simulation data conditions are created by setting different mean vectors and variance-covariance matrix for attributes of multi-variate normal distribution. Simulation Data Construction. The second part is an applicability test by performing a case study of 99 multi-family housings.

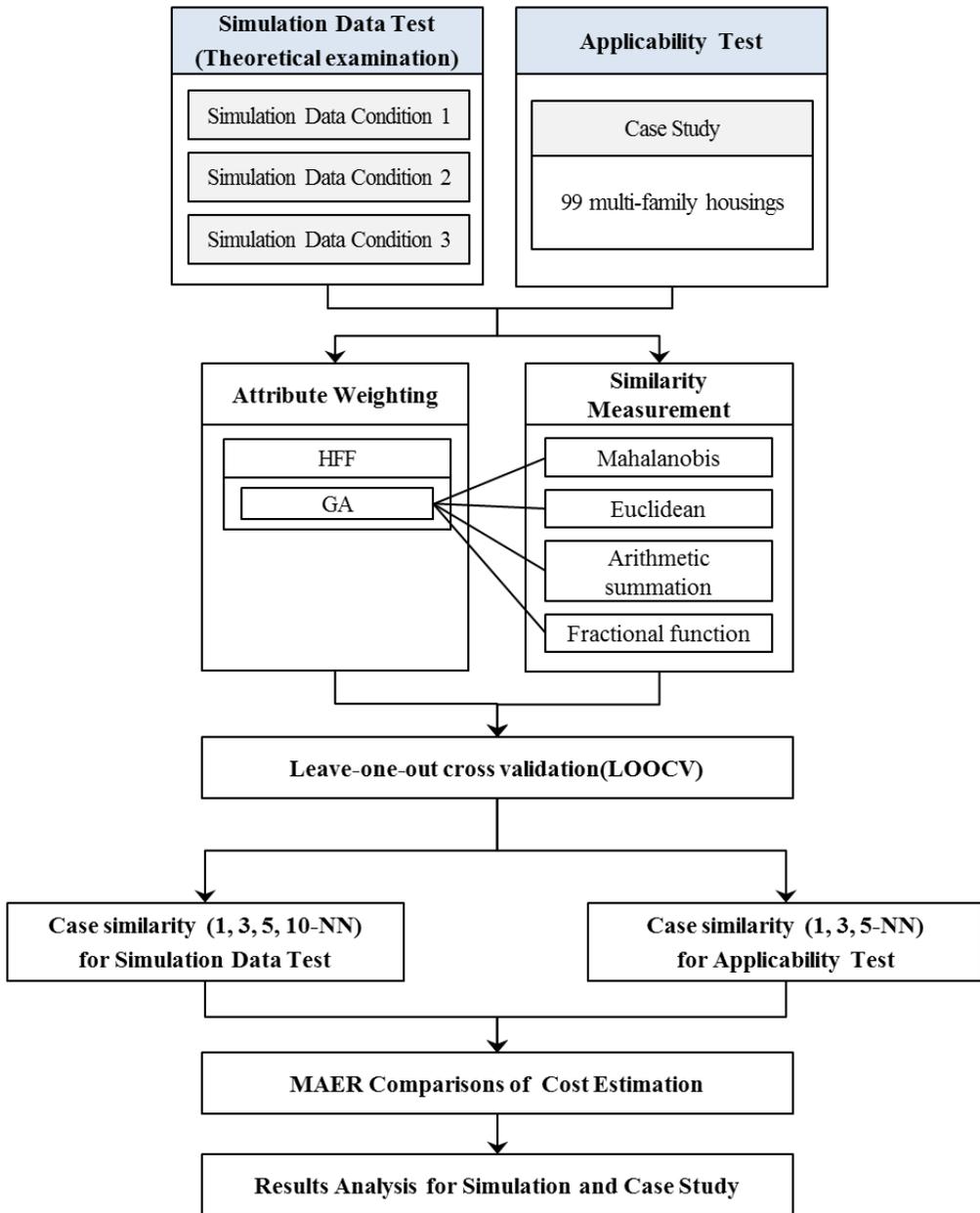


Figure 4-4. Experimental Design for Similarity Measurement Method

Based on these two different types of data sets, the CBR cost models which are distinguished by their four different similarity measurement methods are performed. Simultaneously, the weight value of each attribute is computed using the genetic algorithm optimization process which satisfy the hypothesis fitness function. To perform validation of the CBR cost models, leave-one-out cross validation (LOOCV) method is utilized. LOOCV is a type of k-fold cross validation with k equal to the number of data points in a given set (N).

To estimate building costs, the retrieved cases are used using the k-nearest neighbor principle. The k-nearest neighbor is a concept which searches for the k nearest cases to the target case using a distance measure and then selects the majority of these k cases as the retrieved cases. 1, 3, 5, and 10 nearest neighbors for simulation data test and 1, 3, and 5 nearest neighbors for applicability test are utilized respectively. Next, the comparisons of their mean absolute error rate (MAER) of cost estimation results, which can be defined as below (Eq. 4-10) are carried out. Finally, the comparative analysis on results of the theoretical examination based on simulation data and the applicability test based on case studies of 99 multi-family housings are conducted.

$$MAER = \sum_{i=1}^n \frac{1}{n} \left| \frac{c_i - \hat{c}_i}{c_i} \right| \times 100, \hat{c}_i: \text{estimated or hypothetical cost} \quad (\text{Eq. 4-10})$$

4.3.3 Simulation Data Test

Simulation Data Construction

The tests primarily encompass the simulation of data. In Simulation data condition 1, the mean vector μ is taken to be; $\mu = [52.0 \ 100.1 \ 151.3]^T$. After obtaining the variance-covariance matrix, we can generate 100 samples from the multivariate distribution. We will then find the relationship of the attributes in μ , denoted as w , and the costs, denoted as c . Simulation data condition 2 and 3 take a similar course. We will obtain the appropriate covariance matrices and μ for each case. Similar relations are employed to find the relationship of costs and the attributes.

1) Simulation Data Condition 1:

Parameters of multi-variate normal distribution:

mean vector: $\mu = [52.0 \ 100.1 \ 151.3]^T$

$$\text{variance-covariance matrix } V = \begin{pmatrix} 13.0321 & 2.6544 & 0.0899 \\ 2.6544 & 6.5883 & 1.4438 \\ 0.0899 & 1.4438 & 12.2219 \end{pmatrix} \quad (\text{Eq. 4-11})$$

100 random samples are created from multivariate normal distribution $MN(\mu, V)$. The relationships between cost c and attributes x_1, x_2, x_3 are $c_w = 0.2*x_1 + 0.3*x_2 + 0.5*x_3$. And for normalized data, the relationships between normalized cost r_w and normalized attributes r_1, r_2, r_3 are $r_w = 0.2*r_1 + 0.3*r_2 + 0.5*r_3$.

2) Simulation Data Condition 2:

Parameters of multi-variate normal distribution:

mean vector: $\boldsymbol{\mu} = [132.1 \ 259.5 \ 77.5 \ 346.6 \ 542.4]^T$

$$\text{variance-covariance matrix } \mathbf{V} = \begin{pmatrix} 996.3 & 544.6 & 266.9 & 504 & 480.4 \\ 544.6 & 484.5 & 234.4 & 133.1 & 169.8 \\ 266.9 & 234.4 & 153.2 & 48.8 & 86.2 \\ 504 & 133.1 & 48.8 & 815.4 & 693.1 \\ 480.4 & 169.8 & 86.2 & 693.1 & 622 \end{pmatrix}$$

(Eq. 4-12)

100 random samples are created from multivariate normal distribution $MN(\boldsymbol{\mu}, \mathbf{V})$. The relationships between cost c and attributes x_1, x_2, x_3, x_4, x_5 are $c_w = 0.1*x_1 + 0.2*x_2 + 0.3*x_3 + 0.1*x_4 + 0.3*x_5$ And for normalized data, the relationships between normalized cost r_w and normalized attributes r_1, r_2, r_3, r_4, r_5 are $r_w = 0.1*r_1 + 0.2*r_2 + 0.3*r_3 + 0.1*r_4 + 0.3*r_5$.

3) Simulation Data Condition 3:

Parameters of multi-variate normal distribution:

mean vector: $\boldsymbol{\mu} = [132.1 \ 259.5 \ 77.5 \ 346.6 \ 542.4]^T$

$$\text{variance-covariance matrix } \mathbf{V} = \begin{pmatrix} 996.3 & 0 & 0 & 0 & 0 \\ 0 & 484.5 & 0 & 0 & 0 \\ 0 & 0 & 153.2 & 0 & 0 \\ 0 & 0 & 0 & 815.4 & 0 \\ 0 & 0 & 0 & 0 & 622 \end{pmatrix}$$

(Eq. 4-13)

100 random samples are created from multivariate normal distribution $MN(\mu, V)$. Parameters $x_1 \sim x_5$ follows multivariate normal distribution, however, they are not correlated each other. The relationships between cost c and attributes x_1, x_2, x_3, x_4, x_5 are $c_w = 0.1*x_1 + 0.2*x_2 + 0.3*x_3 + 0.1*x_4 + 0.3*x_5$. And for normalized data, the relationships between normalized cost r_w and normalized attributes r_1, r_2, r_3, r_4, r_5 are $r_w = 0.1*r_1 + 0.2*r_2 + 0.3*r_3 + 0.1*r_4 + 0.3*r_5$.

Results and Discussions

The Mean Absolute Error Rate (MAER) is a performance measure that is used to compute how close a prediction is to the final recorded outcome and is the mean average of the absolute errors. When the first simulation test was performed in Table 4-14 (where the correlated three attributes and not normalized), the following results were obtained. The MAER for 1-NN, 3-NN, 5-NN, and 10-NN were 0.0042, 0.0042, 0.0045, and 0.0046 respectively. The computations were based on Mahalanobis distances. For the Euclidean distances, the results of MAER obtained were 0.0032, 0.0033, 0.0035, and 0.0036. When we performed an arithmetic summation, the results obtained for MAER were 0.0037, 0.0036, 0.0038, and 0.0044 respectively. Finally, an analysis of fractional function yielded 0.0040, 0.0037, 0.0041, and 0.0049.

Table 4-14. MAER Comparison for Simulation Test 1

Similarity Measurement Method	Test 1-1 (3 attributes, correlated)				Test 1-2 (3 attributes, correlated, normalized)			
	1-NN	3-NN	5-NN	10-NN	1-NN	3-NN	5-NN	10-NN
Mahalanobis	0.0042	0.0042	0.0045	0.0046	0.0746	0.0743	0.0764	0.0800
Euclidean	0.0032	0.0033	0.0035	0.0036	0.0593	0.0578	0.0577	0.0658
Arithmetic	0.0037	0.0036	0.0038	0.0044	0.0597	0.0594	0.0619	0.0729
Fractional	0.0040	0.0037	0.0041	0.0049	0.0676	0.0612	0.0653	0.0813

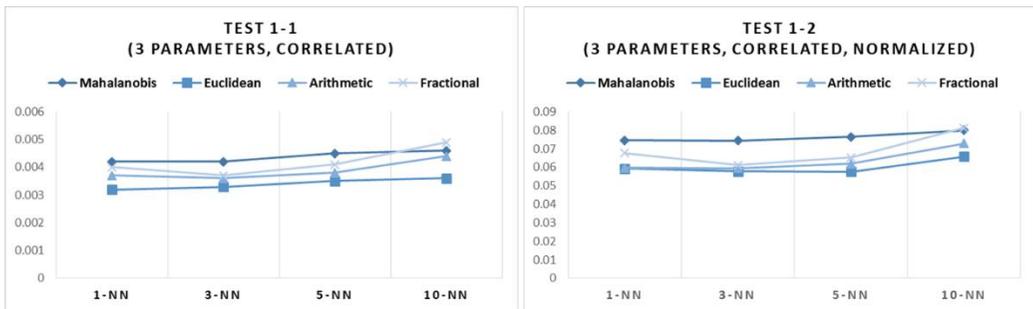


Figure 4-5. MAER Comparison for Simulation Test 1

The test results using simulation data condition 2 (Table 4-15) where the correlated five attributes and not normalized are as follows: The results herein are also varied just as the first case. For Mahalanobis distances, the MAER results for 1-NN, 3-NN, 5-NN, and 10-NN were computed. They were found to be 0.0631, 0.0574, 0.0564, and 0.0549, respectively. 0.0101, 0.0078, 0.0082, and 0.0091 were obtained when computations were based on Euclidean distances. For arithmetic summation, we had 0.0096, 0.0082, 0.0085, and

0.0099. Further analysis by fractional function were 0.0098, 0.0088, 0.0088, and 0.0101 correspondingly as the error rate values.

Table 4-15. MAER Comparison for Simulation Test 2

Similarity Measurement Method	Test 2-1 (5 attributes, correlated)				Test 2-2 (5 attributes, correlated, normalized)			
	1-NN	3-NN	5-NN	10-NN	1-NN	3-NN	5-NN	10-NN
Mahalanobis	0.0631	0.0574	0.0564	0.0549	0.3649	0.3213	0.3200	0.3077
Euclidean	0.0101	0.0078	0.0082	0.0091	0.0638	0.0567	0.0573	0.0685
Arithmetic	0.0096	0.0082	0.0085	0.0099	0.0632	0.0569	0.0587	0.0678
Fractional	0.0098	0.0088	0.0088	0.0101	0.0640	0.0604	0.0607	0.0695

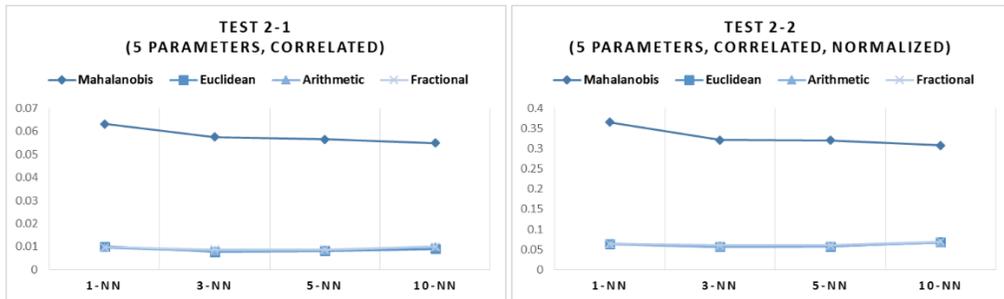


Figure 4-6. MAER Comparison for Simulation Test 2

In the third and final test, we performed experiments based on a simulation data condition 3. Table 4-16 shows results where five uncorrelated and not normalized attributes were utilized. The values obtained for MAER for Mahalanobis distances were 0.0157, 0.0116, 0.0119, and 0.0128 respectively.

Quite varying results of 0.0156, 0.0108, 0.0111, and 0.0121 were obtained for Euclidean distances. In the case of arithmetic summation, we had 0.0161, 0.0127, 0.0131, and 0.0135. Lastly, for fractional function, we obtained the MAER values as and 0.0166, 0.0144, 0.0138, and 0.0149 respectively.

Table 4-16. MAER Comparison for Simulation Test 3

Similarity Measurement Method	Test 3-1 (5 attributes, uncorrelated)				Test 3-2 (5 attributes, uncorrelated, normalized)			
	1-NN	3-NN	5-NN	10-NN	1-NN	3-NN	5-NN	10-NN
	Mahalanobis	0.0157	0.0116	0.0119	0.0128	0.0799	0.0609	0.0640
Euclidean	0.0156	0.0108	0.0111	0.0121	0.0724	0.0582	0.0612	0.0652
Arithmetic	0.0161	0.0127	0.0131	0.0135	0.0726	0.0606	0.0631	0.0669
Fractional	0.0166	0.0144	0.0138	0.0149	0.0763	0.0658	0.0672	0.0721

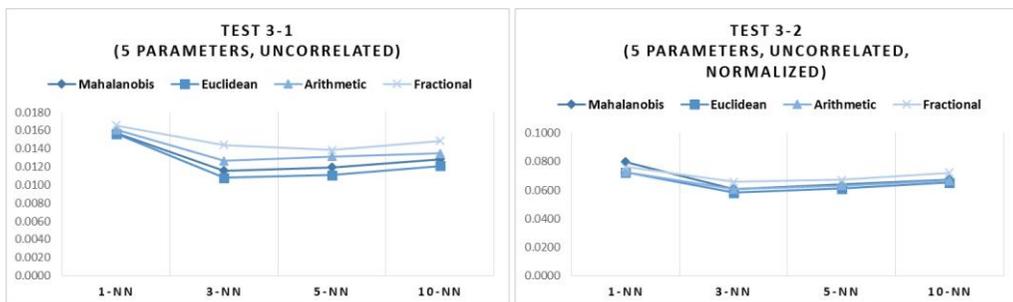


Figure 4-7. MAER Comparison for Simulation Test 3

To summarize, the MAER values computed by Mahalanobis distances were similar or slightly higher than those of other methods. In other words, the CBR cost model using Mahalanobis distance-based similarity measurement

yielded similar or relatively less accurate cost estimation results. When normalized data were used, we obtained higher values of MAER for all similarity measurements. Truly, when the attributes were correlated and not normalized, every similarity measurement methods had achieved overall higher accuracies compared to the conditions where normalized data were utilized. Especially, the Mahalanobis distance based similarity measurement performed much lower MAER with un-normalized data condition. In the condition of attributes uncorrelated, every similarity measurement methods have achieved overall lower MAERs when un-normalized data are utilized compared to normalize one.

4.3.4 Applicability Test

Results and Discussions

As shown in Figure 4-8 and Table 4-17, in terms of the estimate accuracy, the Mahalanobis distance based similarity measure yielded the relatively higher MAER compared to those of Euclidean, arithmetic summation, and fractional function based similarity measures. MAER of 1-NN, 3-NN, and 5-NN for the Mahalanobis distance were 0.475, 0.347, and 0.319, respectively, whereas 0.083, 0.096, and 0.090 for Euclidean distance, 0.088, 0.086, and 0.092 for arithmetic summation, and 0.095, 0.082, and 0.088 for fractional function, correspondingly. As the lower MAERs indicate that the more accurate cost estimate results were obtained by the selected similarity measures, the MAER results of the Mahalanobis similarity measurement method was considered to be relatively less accurate in retrieving similar cases compared to other methods.

Table 4-17. MAER and SD Comparison for Case Study (MFH)

Case	Mahalanobis			Euclidean			Arithmetic Summation			Fractional Function		
	1-NN	3-NN	5-NN	1-NN	3-NN	5-NN	1-NN	3-NN	5-NN	1-NN	3-NN	5-NN
Min.	-0.688	-0.481	-0.424	-0.411	-0.21	-0.239	-0.411	-0.198	-0.189	-0.411	-0.198	-0.195
Max.	0.549	0.383	0.367	0.411	0.457	0.333	0.411	0.457	0.417	0.411	0.457	0.417
MAER	0.475	0.347	0.319	0.083	0.096	0.090	0.088	0.086	0.092	0.095	0.082	0.088
SD_d	0.234	0.162	0.155	0.080	0.081	0.075	0.085	0.077	0.074	0.088	0.074	0.074
m_d	-0.137	-0.102	-0.088	-0.002	-0.004	-0.007	0.000	-0.002	-0.005	-0.002	-0.001	-0.004

Regarding the estimate stability, standard deviation of 1-NN, 3-NN, and 5-NN for the Mahalanobis distance were 0.234, 0.162, and 0.155, respectively, whereas 0.080, 0.081, and 0.075 for Euclidean distance, 0.085, 0.077, and 0.074 for arithmetic summation, and 0.088, 0.074, and 0.074 for fractional function, correspondingly. Regarding estimate stability, we obtained relatively lower standard deviations for Euclidean distance, arithmetic summation, and fractional function compared to Mahalanobis distance.

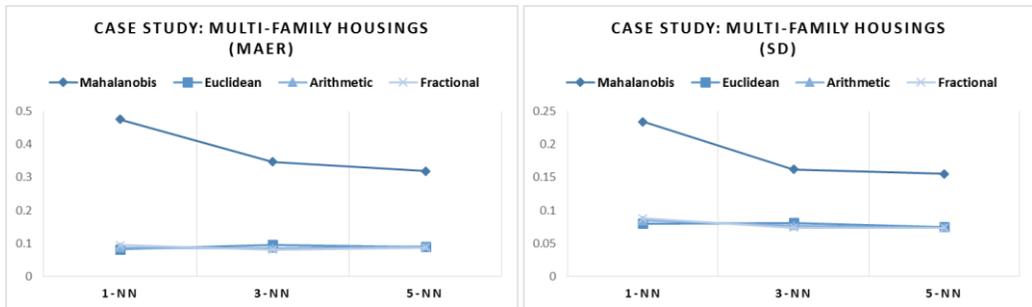


Figure 4-8. MAER and SD Comparison for Case Study (MFH)

Additionally, it was clear from the results of the experiment that lower accuracy (higher MAER) occurred on the side of Mahalanobis distance-based similarity measurement compared to other methods. It was the reverse of expectation because the research was using k-nearest neighbor principle. In k-NN, the final results are the value for the object that is the average of the properties of its k closest neighbors (Everitt et al. 2011). K-nearest principle is the kind of instance-based learning where data is locally approximated, and all calculations done after classification. One of the disadvantages of this principle

is that it is sensitive to the nature of data within its locality. Another demerit occurs when the class distribution is skewed (Beyer et al. 1999). Due to a large number of neighbors, similarity measurement using Mahalanobis distance was profoundly influenced. Mahalanobis distance in our case suffered all these drawbacks.

In overall, both cost estimation results of the theoretical examination based on simulation data conditions and the applicability test by the case study supported that Euclidean distance, arithmetic summation, and fractional function can yield relatively high level of cost estimate accuracy in retrieving similar cases whereas the Mahalanobis distance based similarity measurement achieved an overall higher MAERs and standard deviations. Furthermore, it is important to note that lower MAERs of the Mahalanobis distance method were resulted when the simulation data based experiment was executed compared to MAERs from the case studies using multi-family housings. This was mainly because limitations existed where a large number of attributes were used to compute variance-covariance matrix. The twelve attributes were used for the case study of multi-family housings whereas only three or five attributes were used for simulation test. Therefore, they contained highly correlated information and yielded less accurate results.

4.4 Summary

Normalization

- In terms of the estimate accuracy, ratio normalization yielded a relatively higher accuracy from MAER, MSD, and MAD in comparison to the interval normalization, Gaussian distribution-based normalization, z-score normalization, and logistic function-based normalization.
- In terms of estimate stability, ratio normalization also obtained relatively higher stability compared to the other normalization methods.
- The experiment results confirmed that the MAER, MSD, MAD, and SD can vary according to normalization methods. Thus, we verified the first hypothesis that CBR cost estimation accuracy and stability can be improved by employing statistically accurate normalization methods.
- In terms of the appropriate selection of normalization methods, the kernel density estimation results demonstrated that interval and ratio normalization can be appropriate methods to be applied. Thus, the validation results confirmed the second hypothesis that a CBR cost model has appropriate normalization methods.
- The ratio normalization-based CBR cost model was superior to its model counterparts.

Attribute Weight Assignment

The validation results support the use of the proposed AI in measuring the weights of attributes quantitatively, and it can yield acceptable estimate accuracy when performing parametric or CBR estimations. This research

contributes knowledge of where accurately assigning the weights of attributes is required, and it remains a challenging issue especially in machine learning areas such as attribute weight assignment in CBR and weights of connection strength of neural network method. However, this experiment applied only to public apartment buildings in Korea, and further research should validate the use of AI with other building types for greater generalization. Moreover, instead of correlation coefficients as the weights of the attributes to calculate values of AI, we need to use other weight-assigning methods such as standardized regression coefficients or genetic algorithm to verify the robustness of AI. Also, a sensitivity analysis in terms of how many attributes should be used and their accuracy result comparisons using various methods such as CBR with neural networks or others needs to be further examined. More importantly, the AI concept itself needs further examination and development.

Similarity Measurement

- Euclidean distance, arithmetic summation, and fractional function can achieve high level of cost estimate accuracy and stability when they are applied to CBR cost models.
- CBR cost model using Mahalanobis distance-based similarity measurement yielded relatively less accurate cost estimation results compared to other methods.
- Higher cost estimate accuracy were yielded when less number of attributes were utilized for Mahalanobis distance-based similarity measurement.
- When the attributes were correlated and not normalized, every similarity

measurement methods achieved overall higher accuracy than that of normalized attributes were utilized.

Ultimately, this research demonstrated that the Mahalanobis distance measurement may not be an effective method for CBR based cost estimating during the initial project stages in practical unlike theory. However, even though simulation data tests were conducted, this research only performed case study using multi-family buildings in Korea. Therefore, further research should validate the use of the Mahalanobis distance-based similarity measurement method with other building types for greater generalization.

Chapter 5. Cost Estimation Methodology by Selective Case-Based Reasoning

5.1 Overview of Methodology Development

To yield high levels of accuracy of cost estimation in the front-end stage for diverse building construction projects, a fixed or single cost model has limitation to cope with each building project type. In other words, most of existing CBR cost models have difficulties in term of flexibility in dealing with various characteristics of building construction projects. Rather than providing a single cost model, a systematic cost estimation framework that can provide more accurate CBR cost models responding to building types are required.

Indeed, developing a logical and systematic framework is very critical. Generally, a CBR cost model comprises a set of processes and methods regarding database construction, attribute selection, normalization, attribute weight assignment, similarity measurement, reuse of retrieved cases, retain of learned cases, and so on. However, the level of details, methods utilized in CBR cost models, and processes vary depending on the perspectives of the researchers (Watson 1997; Karshenas and Tse 2002; Ahn et al. 2006; Koo et al. 2010b). Therefore, designing an optimal reasoning environment, especially to deal with various building types is crucial to improve the level of the CBR-

based cost estimation accuracy as each process of the CBR cost model is closely interactory.

This research develops a front-end cost estimation methodology by selective case-based reasoning and its structure can be conceptualized as in Figure 5-1. Basically, the methodology comprises 1) *case-base development module*, 2) *method selection module*, and 3) *CBR cost model module*. The case-base development module includes acquiring building cost data, extracting attributes, designing case-base structure, storing data, analysis of matrix plot, and preprocessing data. This module aims to prepare well-preprocessed and organized data for the accurate and efficient use of CBR-based cost estimation. After a case-base is prepared from the case-base development module, methods of main CBR components are selected in the method selection module.

This research utilize normalization method, attribute weighting method, and similarity measurement method as CBR main components; and how each method can affect cost estimation accuracy was examined in Chapter 4. The method selection module is composed of three sub-modules. In sub-module 1, the most accurate and appropriate normalization method is selected among interval normalization, Gaussian distribution-based normalization, z-score normalization, logistic function-based normalization, and ratio normalization. In sub-module 2, the most accurate attribute weighting method is selected among attribute impact suggested by Ahn et al. (2014), entropy, feature counting, and genetic algorithms. In sub-module 3, the most accurate similarity

measurement method is selected among Mahalanobis distance-based similarity measurement, Euclidean distance-based similarity measurement, arithmetic summation-based similarity measurement, and fractional function-based similarity measurement. The three sub-modules can make cross references of a selected method from each sub-module to another sub-module. Next, a selected method from each sub-module is reflected to CBR cost model module; and based on the selected method for normalization, attribute weighting, and similarity measurement, building project costs are estimated and reported to owners or cost estimators to support decision-makings. The details of methods and processes for Module 1, and 2 will be explained in section 5.2, and 5.3, respectively.

The case base development module can also transfer information to the CBR cost model module directly without using the method selection module when main CBR components are pre-determined. In this case, the case-base is utilized straightly to retrieve and reuse similar cases to estimate costs. Fundamentally, it collects the information from the previous cases and maps it on the target problem. The concept involves bringing forth the solution and putting it to use in the given problem. If the problem needs a more advanced approach, then other CBR models can be applied through the method selection module to increase cost estimation accuracy.

After cost estimation have been made from the CBR cost module, the learned cases are retained in the case-base of the case base development module.

The newly acquired knowledge was previously nonexistent; therefore, these knowledge need to be recorded and kept in the case-base. Lastly, a new case is permanently retained in the memory. This is after the suggested solutions get approved and become the newly adapted notations that can be used to fix the problem.

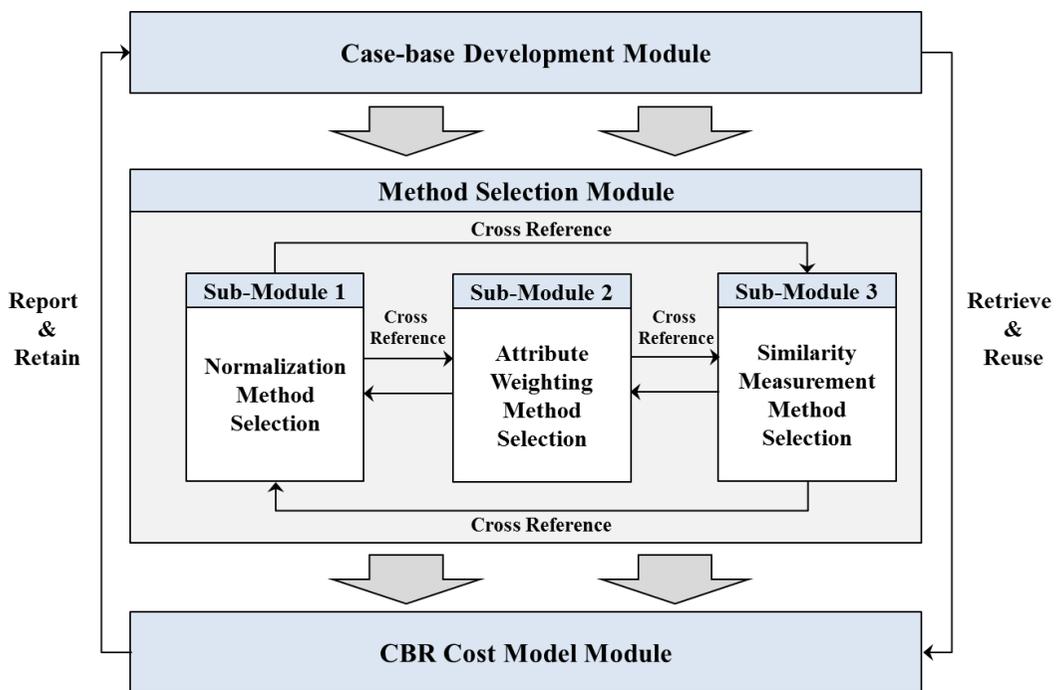


Figure 5-1. Framework of Front-End Cost Estimation by Selective CBR

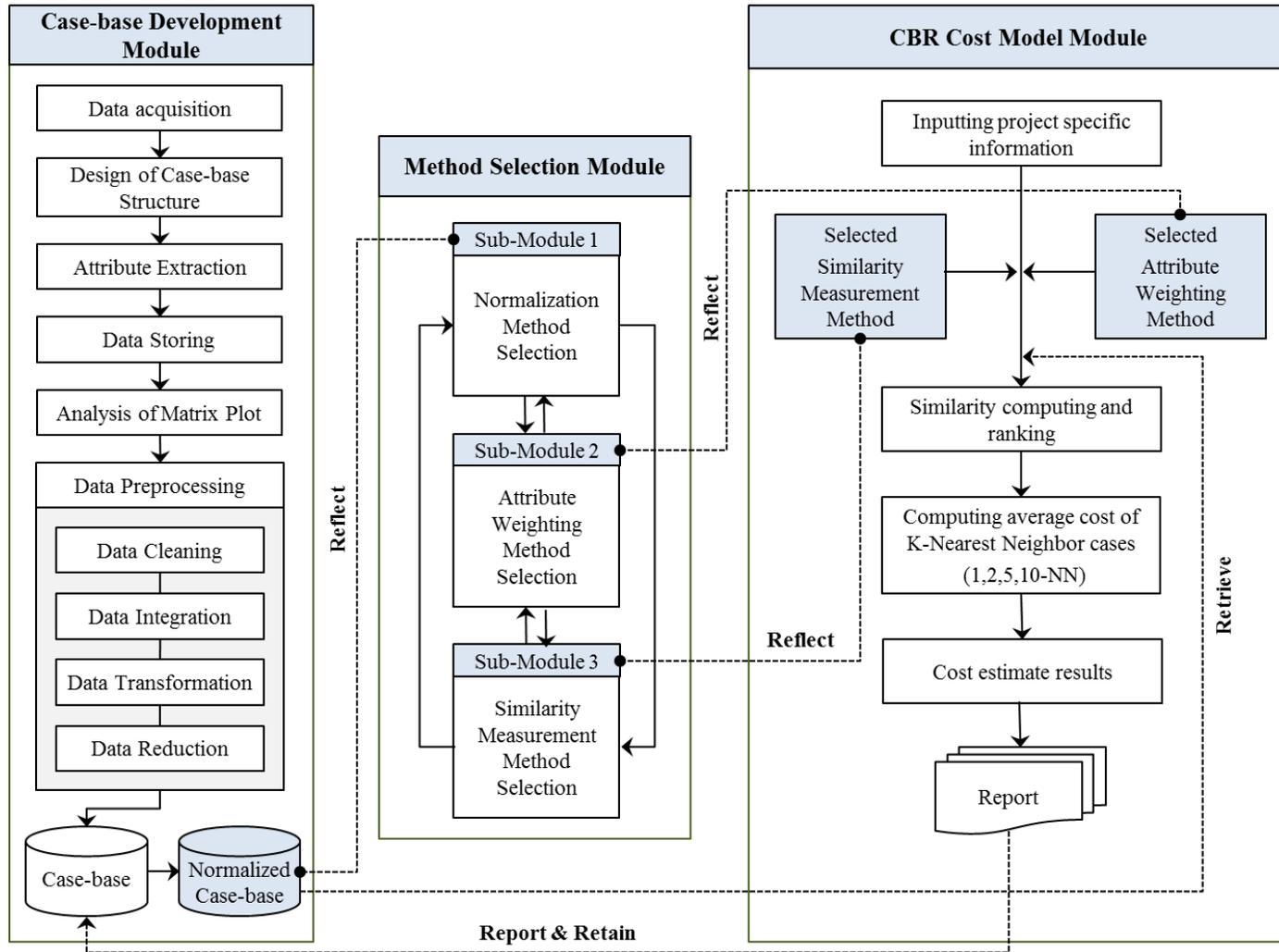


Figure 5-2. Mechanism of Front-End Cost Estimation by Selective CBR

5.2 Case-Base Development

The case-based reasoning methodology highly relies on the data presented in the computation, analysis, and recommendations (Pal and Shiu 2004). Therefore, the validity and the reliability of the CBR-based cost model will largely depend on the quality of the data. Since the data obtained for the research is source data whose reliability and validity might be questionable, it is critical to reprocess the data before using it for the research. When the data used is of quality, the recommendations and the decisions of high reliability from the CBR cost estimation results can be achieved. To accomplish such purposes, this research proposes a *Module 1: Case-Base Development* as illustrated in Figure 5-2. The *Module 1* comprises data acquisition, design of case-base structure, attribute extraction, data storing, and data preprocessing.

Data Acquisition and Design of Case-Base Structure

The CBR method is an effective approach when dealing with a unique set of attributes that needs analysis. In short, the CBR method is case sensitive. If the validity of the existing data is trustworthy, then the results obtained by the use of the data are reliable. Therefore, reliable data acquisition and analysis is a fundamental procedure in obtaining quality estimation results from a quality case-base.

Most of all, raw data must be collected and arranged in an exquisite manner to eradicate ambiguity and inconsistency (Melnyk and Morrison-Beedy

2012). Table 5-1 is a data structure for the case base. It comprises of many cases say from 1 to n. Attributes are numbered from one to an indefinite value say m. The relationships between the attributes and the cases give the estimated values of the project cost. Based on the regularized data structure, a case-base can be constructed and prepared for the accurate and efficient cost estimation.

Table 5-1. Design of Case-Base Structure

	Attribute						Project Cost	
	1	2	...	j	...	m		
Case	1	x_{11}	x_{12}	...	x_{1j}	...	x_{1m}	C_1
	2	x_{21}	x_{22}	...	x_{2j}	...	x_{2m}	C_2
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{im}	C_i
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	n	x_{n1}	x_{n2}	...	x_{nj}	...	x_{nm}	C_n
Project	0	x_{01}	x_{02}	...	x_{0j}	...	x_{0m}	C_0
Weight	w_1	w_2	...	w_j	...	w_m		

Attribute Extraction

Extracting attributes affecting construction cost and calculating their weights are very important in CBR cost estimation. This is especially for the reason that the utilization of too many attributes in CBR cost estimation decreases the accuracy of early estimates and the efficiency of the estimate process (Ahn et al. 2014). Therefore, attributes of high cost implication need to be extracted and measured in quantitative manner.

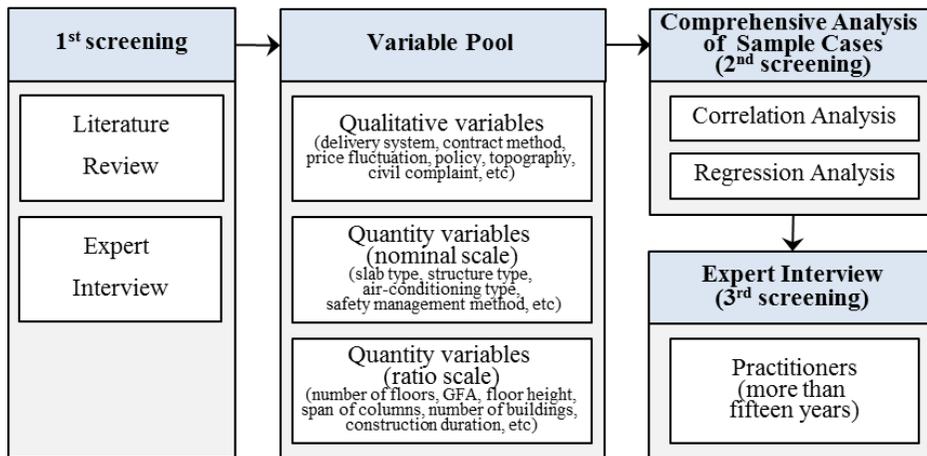


Figure 5-3. Attribute Extraction Process

Attributes can be selected through the attribute extraction process described in Figure 5-3. First, based on the literature reviews and expert interviews, a pool of variables comprised of qualitative, quantitative (nominal scale), and quantitative (ratio scale) is constructed. Next a comprehensive analysis of the sample cases is performed to reduce the number of variables by

conducting correlation and regression analysis which are relatively simple statistical methods to compute and interpret the weights of attributes. Finally, the attributes are selected and confirmed by the construction experts who had participated in the construction industry.

Analysis of Matrix Plot

After extracting attributes, attributes and cost information are stored in a case-base. As a next process, analysis of matrix plot is performed. A matrix plot is a diagram that can be used to evaluate the relationship and pattern among various pairs of attributes at the same time. Ideally, the matrix is a combination of different scatter plots. A scatter plot drawn to illustrate the relationship between two attributes and is helpful in providing analysis of correlation coefficient (Cleveland and McGill 1988; Cleveland 1993). A positive association between attributes is indicated by an upward trend (positive slope). On the other hand, a negative relationship is indicated by opposite effect (negative slope) whereas a non-associated attributes does not have any trend.

There are two types of matrix plots namely model plots and each y versus each X. Matrix plots indicate plots for any pairing attributes. A matrix plot is useful in instances when dealing with many attributes, especially when examining the relationships that exist between the pairs of attributes (Cleveland and McGill 1988; Cleveland 1993). Notably, an each Y versus each X matrix plot indicates a plot for every possible y and x combination but only when y and x attributes are specified. The type of matrix is suitable when examining

the relationships between estimators and responses when their figures are entered differently.

Data Preprocessing

Data preprocessing refers to any process performed on a fresh data in a way that would make subsequent operations easier (Famili et al. 1997). New data can also be called as raw data or source data because it is yet to be subjected to any processing. The importance of preprocessing is that it transforms data into a form that would be easy to manipulate in the subsequent processes (García et al. 2015).

Data processing is employed as the prioritizing process and is aimed to identify the useful trends and the interrelationships that could be hidden by the large quantity of the data. Many types of data processing methods are included in this phase. However, their purpose is to prepare the raw data for the further processing procedures (Kotsiantis et al. 2006). The principal aim of this phase is that if the raw data were used for the analysis prior to screening for discontinuity and discrepancies, the result produced by the analysis would be misleading. In the light of this, the careful screening of the raw data is the foremost critical step before proceeding with the analysis. It prevents losing the validity and the reliability of the research.

The presence of many redundant and irrelevant information or unreliable and noisy data would make the knowledge identification in the training phase more complicated. The raw data-filtering and preparation steps are a tedious and time-consuming process that cannot be overlooked by the researchers. It is one of the principal determiners of the observation of the rigorous processes in the scientific studies. In case the researchers may overlook this critical step, the validity and the reliability of their research would be lost.

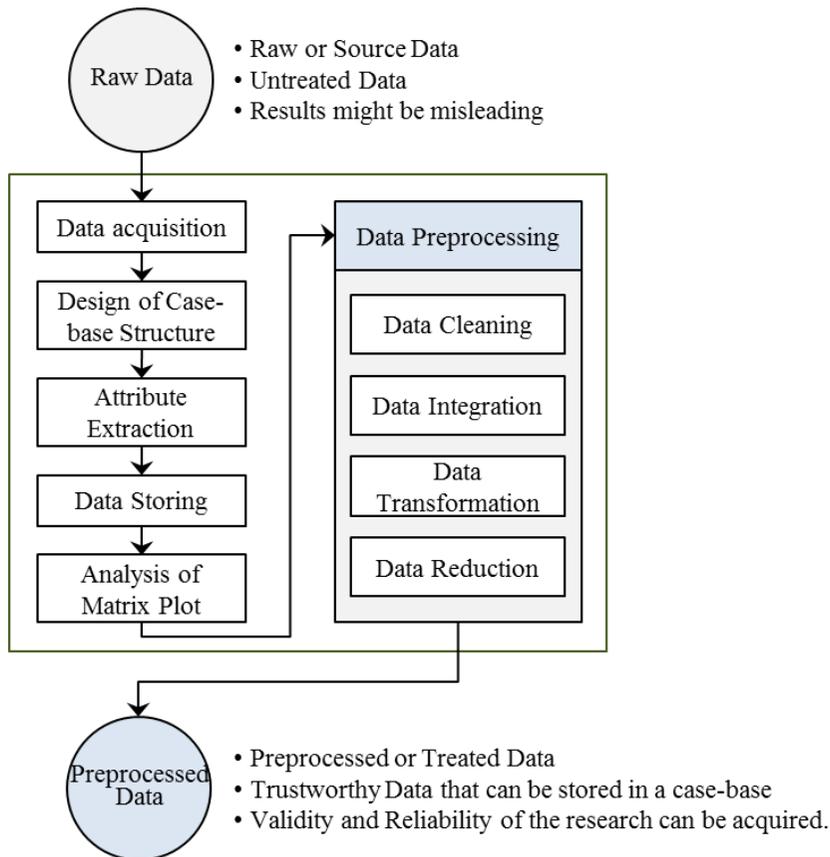


Figure 5-4. Procedures of Acquiring Preprocessed Data

In the raw data preparation and the filtering phase of the CBR research, many scientific methods and procedures are employed. The selection of the method to adopt will be determined by the degree of validity of the raw data or the percentage of the present redundant, irrelevant and the noisy data in the raw data. It would depend on the amount of the data present, the resources available in terms of time, money and labor. It is critical for the researchers to document the rationale of the chosen method and employ measures of mitigating the propagated discrepancies.

1) Data Cleaning

Data cleaning is sometimes referred to as data cleansing or data scrubbing. Data cleaning is a process that aims at detecting and correcting a database (Rahm and Do 2000). The correction process involves removing data that is considered erroneous. In other words, erroneous data is dirty data and is caused by typing input mistakes, contradictions, disparities, lack of validation of the changes made, and missing bits. Dirty data is also considered as unwanted especially if they are out of date, lack appropriate format, redundant or incomplete. Correspondingly, data cleaning process entails filling in the missing values after it has been put into a database. It also involves smoothing noises and removing outliers. In general, data cleansing is a process that ensures that every piece of data in the database is consistent and conforms to the standards set by the researchers.

The major issues in data cleaning are the existence of data that miss values or particular attributes and outliers (Van den Broeck et al. 2005). They also include data that are inconsistent with nature when compared to previous ones. The process of cleaning data that contain data that miss some values includes using attribute mean to fill in the missing values. Missing values can also be predicted using learning algorithms. The process of removing outliers includes binning, clustering and regression processes. Data cleaning is relevant because it ensures the reliability of the final information.

2) Data Integration

Data integration is the process of combining data from various sources to form a holistic view of the information being constructed (Nemati et al. 2004). In most cases, such pieces of data are stored in different devices and forms. Data integration changes them into a universal form before combining them. In most cases, data integration process comes into play when two different datasets are merging. The process is also applicable in situations where a company decides to consolidate its data applications and various data storage places to provide a holistic view of the data asset of the company and to create a data warehouse (Nemati et al., 2004). Combining various data storage places also enhances data management process. It also makes it easier to trace how information is used within and outside the organization. Lastly, data integration enhances access of information because there would be only one place to look at or few people to ask for a piece of information.

Data integration process covers areas such as data warehousing, migration, enterprise application, and master data management (Peltier et al. 2013). Testing is paramount to ensure that the consolidated data is realistic regarding accuracy and relevance. Data integration can be performed using various techniques, including manual integration, integration using application, and middleware data integration. Others include virtual and physical data integration processes.

3) Data Transformation

Data transformation refers to the conversion of data from one format to another (McDonald 2009). Data transformation includes smoothing, aggregation, generalization, normalization, and new attribute construction. In most cases, data is usually transformed from source to status to facilitate the execution of subsequent processes. Usually, source data always arrive in formats that cannot be supported by applications in the following operations. If at all they are supported, then they are still not easy to use or manipulate because of their bulkiness. The conversion process usually takes place in documents. However, a transformation process may also include the conversion of a computer program from one version to another, especially if it has to run on a different platform that does not support the original version. Either way, data transformation exists for one important reason, and that is to convert data into a form that would readily be used in the next step of data mining or analysis process.

4) Data Reduction

As the name suggests, data reduction process entails minimizing the amount of data to be stored in a particular environment (Ehrenberg 2000). Data reduction is usually performed to enhance the efficiency of a storage environment and reduce costs resulting from getting several storage environments. Data reduction can also be defined as the process of converting huge chunks of data into summarized version with the aim of using as less space within the storage environment as possible. Correspondingly, it can also be thought of as the process of transforming data into a form that is in order, correct and simple. Data reduction process can be facilitated using various methods such as data deduplication. Data deduplication eliminates redundant data that would otherwise occupy space unnecessarily. Data Compression removes redundant information from a database such that the remaining data is lean, concise and accurate.

Data reduction process may also comprise of editing, scaling, coding, sorting and collating, especially in cases where data was already converted into a digital format. Therefore, data reduction processes may be applied to data to remove erroneous cases such as duplication, noises, and irrelevant aspects. It may also be applied to cleaned data intentionally to compress it into a suitable size that can be accommodated by the available storage space without infringing on the existing information.

5.3 CBR Method Selection and Cost Estimation

The *Module 2: Method Selection* has three sub-modules. They are the *Sub-Module 1: Normalization Method Selection*, *Sub-Module 2: Attribute Weighting Method Selection* and finally *Sub-Module 3: Similarity Measurement Method Selection*. The process and methods of the three sub-modules are explained in sub-sections 5.3.1~5.3.3.

5.3.1 Sub-Module 1: Normalization Method Selection

The *Sub-Module 1: Normalization Method Selection* is 1) to determine which normalization method when applied to CBR can yield the most accurate and stable cost estimation results and 2) to select appropriate normalization method. The relationship between normalization method and CBR cost estimation accuracy was examined and verified through comparative experiment in Chapter 4.1.

The normalization method selection processes are as follows:

- 1) A case-base of a specific building project is normalized using interval normalization, Gaussian distribution-based normalization, z-score normalization, logistic function-based normalization, and ratio normalization, respectively. Then, total five different normalized case-bases are prepared. Depending on the research objectives, researchers can add other advanced normalization methods or delete the existing normalization methods.

- 2) Case similarity is computed applying a similarity measurement method and an attribute weighting method. At the initial run of this module, an existing validated similarity measurement method and attribute weighting method can be set. Another way is that a selected attribute weighting method from sub-module 2 and a selected similarity measurement method from sub-module 3 can be applied.
- 3) Leave-one-out cross validation (LOOCV) is used and similar cases of 1, 2, 5, 10-NN are retrieved to estimate the average costs.
- 4) As performance measures of CBR cost models based on five different normalization methods, MAER, MSD, MAD for estimate accuracy, and SD for estimate stability are utilized.
- 5) Kernel density estimation (KDE) is performed to find out appropriate normalization methods.
- 6) Based on the results of MAER, MSD, MAD, SD and KDE, the most accurate, stable, and appropriate normalization method is selected.

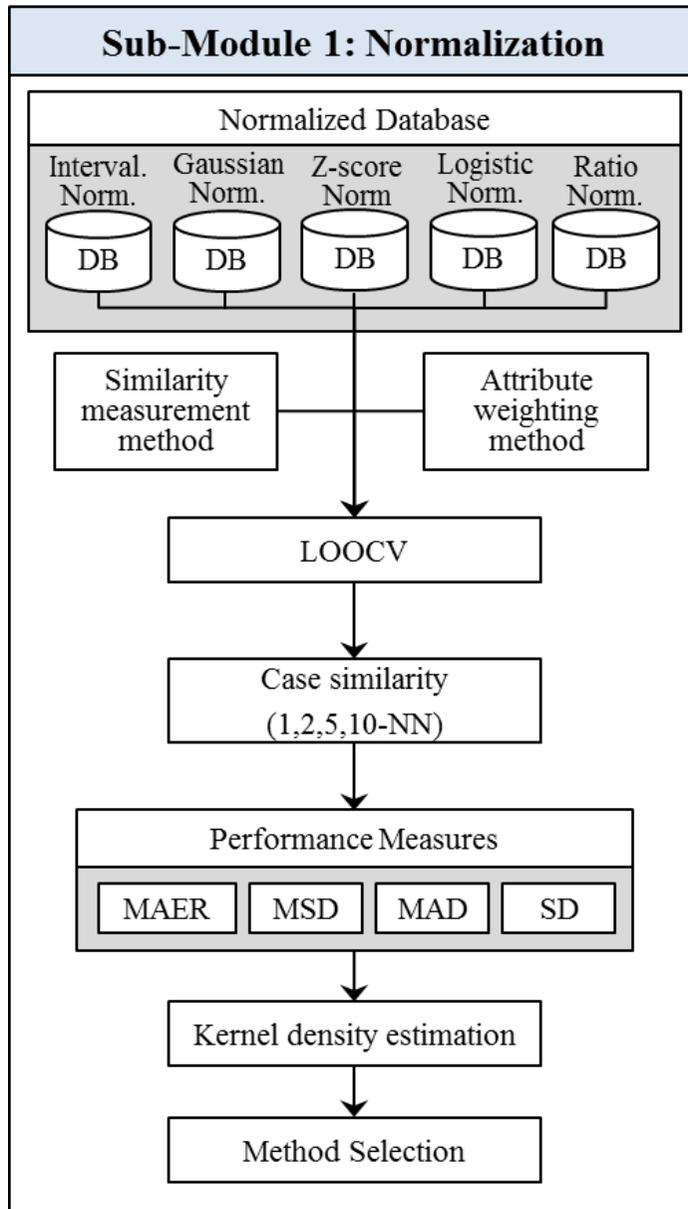


Figure 5-5. Sub-Module 1: Normalization Method Selection

Normalization Methods

This section explains five normalization methods: interval norm., Gaussian distribution-based norm., z-score norm., logistic function-based norm., and ratio norm. Characteristics, pros and cons of each method are described. These methods are applied to the case-base; and normalized data are utilized in the CBR cost models.

1) Interval Normalization

The interval normalization method (also called score range transformation) can cover attributes of ordinal, interval, and ratio scale types, and transform the score range exactly between 0 and 1. This method applies the theory and practice of rigorously working on a computer with certain and uncertain real numbers, which is represented as intervals. The method involves the appropriate use of standard floating-point calculations, direct floating-point calculations and interval arithmetic (Neumaier, 2008). These three computations are combined in a way that gives reasonable enclosures of the results with an acceptable cost. However, this type of normalization method also has the disadvantage in which transformed scores are not proportional to the original data. Thus, relative distances among the original data are not preserved. Also, the method requires too much attention, especially with the use of interval arithmetic. Hence, the results might be invalid (Neumaier, 2008).

$$r_{ij} = \begin{cases} \frac{x_{ij} - x_j^{\min}}{x_j^{\max} - x_j^{\min}}, & \text{for benefit criterion} \\ \frac{x_j^{\max} - x_{ij}}{x_j^{\max} - x_j^{\min}}, & \text{for cost criterion} \\ \frac{|x_{ij} - T|}{\max\{x_j^{\max} - T, T - x_j^{\min}\}}, & \text{for desired value of } T \end{cases} \quad (\text{Eq. 5-1})$$

2) Gaussian Distribution-based Normalization

Gaussian distribution normalization can be utilized to describe physical events where the number of events is very large. The Gaussian distribution is a continuous probability distribution that approximates the exact binomial distribution of events (Havil 2003). Gaussian distribution normalization has acquired other names such as normal distribution. Moreover, because of its curved flaring shape, the Gaussian distribution is referred as the "bell-shaped curve." According to Havil (2003), the biggest advantage of the Gaussian distribution-based normalization is the many convenient properties such as normal sum distribution and normal difference distribution. However, there is always an unfortunate tendency to invoke normal distributions in situation where they may not be applicable (Patel and Read 1996).

$$\text{Standard Score } z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}, \quad (\text{Eq. 5-2})$$

\bar{x}_j : mean of attribute j , s_j : standard deviation of attribute j

$$r_{ij} = \varphi(z_{ij}),$$

$\varphi()$: Cumulative distribution function of standard normal distribution

3) *Z-score Normalization*

Z-score is used to standardize errors when population parameters are known. A z-score or standard score is obtained when the population mean is subtracted from an individual raw score. The difference is then divided by the population standard deviation to give a dimensionless quantity which is the standard score (Kreyszig 1979). This whole conversion process is referred to as normalizing or standardizing. According to Carroll and Carroll (2002), z-score normalization's strong point is that it can work out prediction intervals. Also, as z-score normalization is performed based on standard deviation and not by its range, normalized attributes are less affected by outliers. However, the weakness of z-score normalization is seen in its ineffectiveness in cases where population's parameters are unknown or estimated. Each attribute needs to have normal or at least symmetric distribution to properly utilize the z-score method. The value v of attribute A is normalized to V' by computing:

$$V' = \frac{v - \mu_A}{\sigma_A}, \mu: \text{mean}, \sigma: \text{standard deviation} \quad (\text{Eq. 5-3})$$

4) *Logistic Function-based Normalization*

The application of the logistic function is mainly in logistic regression where it is used to model how the probability of an event may be affected by one or more explanatory attributes. The logistic function is instrumental in the Rasch model, which is used in item response theory (Bod et al., 2003). After computing the standard score of each attribute, normalization is performed by the values of the logistic function, which corresponds to the standard scores.

However, the coefficient in a logistic function changes according to the logit. Thus, the interpretation of normalized values using a logistic function is more complicated (Gershenfeld 1999).

$$\text{Normalized Score } r_{ij} = L(z_{ij}) = \frac{1}{1 + e^{-z_{ij}}}, L(): \text{Logistic Function} \quad (\text{Eq.5-4})$$

5) *Ratio Normalization*

Ratio normalization or maximum score transformation is the most simple linear transformation method. In the context of score transformation, ratio standardization is the procedure of converting scores from raw scores into transformed scores. The main advantage is that proportional properties still remain as the original data properties. In other words, the normalized values preserve relative distances of the original data. According to Bod et al. (2003), the procedure serves crucial purposes such as giving meaning to the scores, and hence allowing for easy interpretation of the scores. Moreover, direct comparison of two scores is possible. The two main types of transformation include percentile ranks and linear transformation. However, the disadvantages are that the minimum value of the transformed data is not '0', which causes difficulties for interpretations where both positive and negative values coexist in the original data.

$$r_{ij} = \begin{cases} \frac{x_{ij}}{x_j^{\max}}, & \text{benefit criteria} \\ \frac{x_j^{\min}}{x_{ij}}, & \text{cost criteria} \end{cases} \quad (\text{Eq. 5-5})$$

5.3.2 Sub-Module 2: Attribute Weighting Method Selection

The *Sub-Module 2: Attribute Weighting Method Selection* is to decide which attribute weighting method can lead to higher cost estimate accuracy and stability. How different attribute weighting method can affect CBR cost estimate accuracy was examined in Chapter 4.2.

The attribute weighting method selection processes can be described as shown in Figure 5-6.

- 1) A case-base of a specific building project is normalized. An existing normalization method or the proposed normalization method from sub-module 1 can be utilized.
- 2) Similarity of cases are calculated using a similarity measurement method and four different attribute weighting methods (AI, entropy, FC, and GA). As a similarity measurement method, either a widely utilized method in academia or a suggested similarity measurement method from sub-module 3 can be adopted.
- 3) LOOCV is used as a validation method and average costs of 1, 2, 5, 10-NN are computed.
- 4) To measure the performance of different CBR models by four different attribute weighting methods, MAER, MSD, MAD, and SD are used.

- 5) Using the results of MAER, MSD, MAD, and SD, an attribute weighting method that can achieve the highest accuracy and stability is suggested.

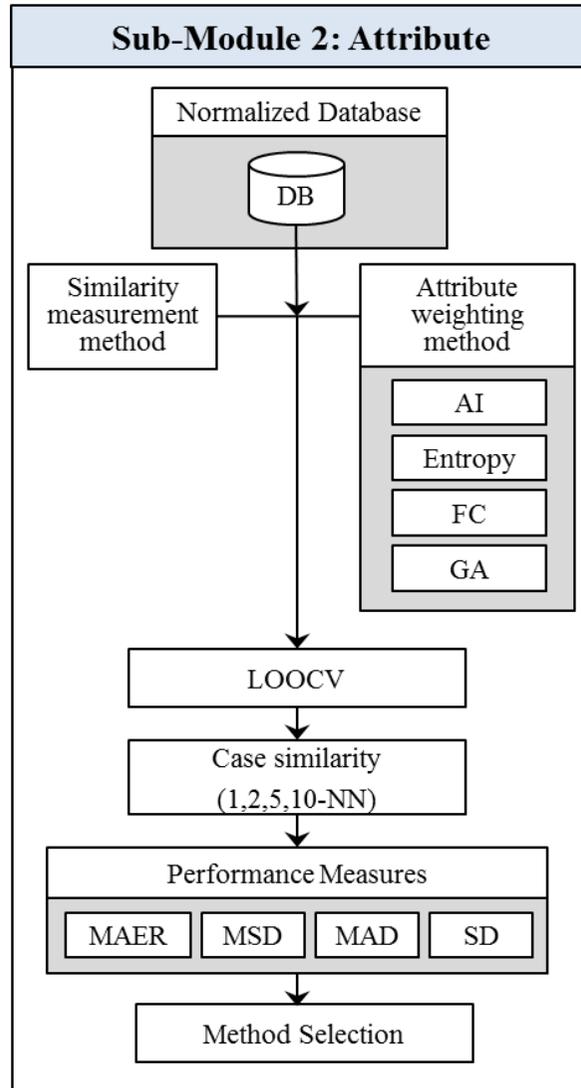


Figure 5-6. Sub-Module 2: Attribute Weighting Method Selection

Attribute Weighting Methods

1) Attribute Impact (AI) (Ahn et al. 2014)

$$\text{Attribute Impact (AI}_i) = \text{AW}_i \times \text{AR}_i \quad (\text{Eq. 5-6})$$

Since all AR can be regarded “1” for normalized data, AI can be equal to AW. In this module, coefficient of determinants among normalized cost and attributes are utilized as AW. Also, as total weights need to be “1”, each coefficient of determinants is divided by the sum of coefficient of determinants and then utilized as attribute weights.

2) Entropy (Ahn 2007)

The entropy procedure estimates weight vectors from data matrix. From the viewpoint of entropy, data matrix in itself contains information based on which the weights of attributes can be estimated. That is, an attribute showing a large variance among different cases is an important attribute, while an attribute showing a small variance among different cases is a less important criterion.

First of all, let the parity value of the j^{th} attribute for the i^{th} case, p_{ij} , be defined as the ratio of the value of the j^{th} attribute for the i^{th} case to the sum of the values of the j^{th} attribute for all cases.

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^m x_{ij}} \quad (\text{Eq. 5-7})$$

Thus $p_j = (p_{1j} p_{2j} \cdots p_{mj})^t$ is the composition ratio vector of the j^{th} attribute. Then the entropy of the j^{th} attribute, e_j , is defined as follows:

$$e_j = -\sum_{i=1}^m p_{ij} \log p_{ij} \quad (\text{Eq. 5-8})$$

Here, the possible range of e_j is given as follows:

$$0 \leq e_j \leq \log m \quad (\text{Eq. 5-9})$$

Therefore normalized entropy, $u_j = \frac{e_j}{\log m}$, can function as the degree of uniformity:

$$0 \leq u_j \leq 1 \quad (\text{Eq. 5-10})$$

Based on this, the degree of diversity of the j^{th} attribute, d_j , is defined as follows:

$$d_j = 1 - u_j \quad (\text{Eq. 5-11})$$

And then the weight of the j^{th} attribute, w_j , is defined as the following normalized diversity:

$$w_j = \frac{d_j}{\sum_{i=1}^n d_i} \quad (\text{Eq. 5-12})$$

If one has been placing subjective weights, s_j , on the importance of individual attributes, then corrected weights, w_j^* , can be calculated as follows:

$$w_j^* = \frac{s_j w_j}{\sum_{i=1}^n s_i w_i} \quad (\text{Eq. 5-13})$$

Weights in this research were obtained using Equation (5-12).

3) *Feature Counting (FC)*

Equal weights or importance are allocated for every attribute (Esteem 1996; Doğan et al. 2006; Ji et al. 2011b).

4) *Hypothesis Fitness Function (HFF) and Genetic Algorithms (GA)*

To carry out weight assessment based on the extracted twelve attributes, a hypothesis can be described as:

Hypothesis:

$$C_i = \sum_{j=1}^m w_j w_{ij}, \sum_{j=1}^m w_j = 1, w_j \geq 0, j = 1, \dots, m \quad (\text{Eq. 5-14})$$

Where distance be expressed as: Distance $d_i = C_i - \sum_{j=1}^m w_j x_{ij}$
(Eq. 5-15)

We formulate a Hypothesis Fitness Function (HFF), an optimization model that describes which attributes can best explain the cost and find w , minimizing the HFF model.

Hypothesis Fitness Function (HFF):

$$F(w) = \sum_{i=1}^n d_i^2 = \sum_{i=1}^n (c_i - \sum_{j=1}^m w_j x_{ij})^2 \quad (\text{Eq. 5-16})$$

There can be many different kinds of algorithms used if they can satisfy the HFF model. In this research, we adopt GA to assign the weights of attributes which satisfies the HFF model.

Genetic algorithms offer an approach for learning methods, which utilizes the concept of generation of successor hypotheses through processes such as

iterative mutation and crossover (Shin and Han 1999). The biological evolution processes inspire this generic procedure. Over the years, genetic algorithms have been considered a valid tactic to solve problems that require competent and efficient searching (Jarmulak et al. 2000).

Genetic algorithms are applied mostly in businesses, science, and engineering. In fact, genetic algorithms are efficient for improvement as they make computations simple. They provide a ranking criterion for potential hypothesis that are referred to as the hypothesis fitness function, and thus useful in weighing all the members of a given population (Goldberg 2006). Genetic algorithms are fundamental in CBR cost estimation. Courtesy of the general algorithms, it is possible to optimize the attribute weight values. Most importantly, genetic algorithms are used to examine and identify a space of a candidate solution and isolate the best solution among the other solutions (Mitchell 1997). In this research, we employ the genetic algorithms equation (Eq. 5-17) and attribute weights applied by Ji et al (2011b) and Ahn et al. (2014).

$$\min \sum_{n=1}^j \sqrt{D_n^2}, \quad \text{s. t.} \quad \begin{pmatrix} C_1 \\ \vdots \\ C_j \end{pmatrix} - \begin{pmatrix} X_{11} & \cdots & X_{1i} \\ \vdots & \ddots & \vdots \\ X_{j1} & \cdots & X_{ji} \end{pmatrix} \begin{pmatrix} \omega_1 \\ \vdots \\ \omega_i \end{pmatrix} = \begin{pmatrix} D_1 \\ \vdots \\ D_j \end{pmatrix} \quad (\text{Eq. 5-17})$$

5.3.3 Sub-Module 3: Similarity Measurement Method Selection

The *Sub-Module 3: Similarity Measurement Method Selection* is developed to propose a similarity measurement method that can lead to the highest estimate accuracy and stability.

Accuracy of cost estimate results according to different similarity measurement methods was tested in Chapter 4.3.

The similarity measurement method selection processes can be illustrated in Figure 5-7.

- 1) A case-base is normalized for the efficient and effective reasoning of datasets. Either an existing normalization method or a suggested normalization method from sub-module 1 can be adopted.
- 2) Similarity is computed using similarity measurement methods by Mahalanobis, Euclidean, arithmetic summation, and fractional function methods respectively and an attribute weighting methods. As an attribute weighting method, either a widely used method or a proposed attribute weighting method from sub-module 2 can be utilized.
- 3) LOOCV is used as a validation method and retrieved cases of 1, 2, 5, 10-NN are computed.
- 4) As performance measures of CBR models by four different similarity measurement methods, MAER, MSD, MAD, and SD are utilized.

- 5) Based on results of MAER, MSD, MAD, and SD, the most accurate and stable similarity measurement method is selected.

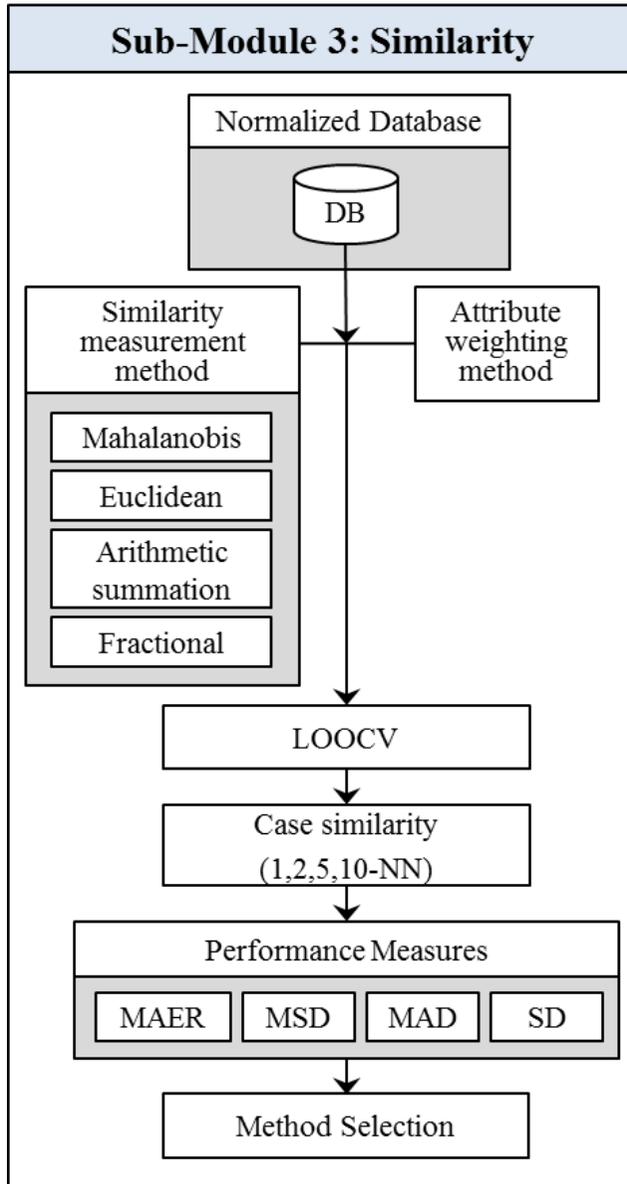


Figure 5-7. Sub-Module 3: Similarity Measurement Method Selection

Similarity Measurement Methods

Similarity is defined as a quantity that denotes the strength of a relationship between two features or two objects. In other words, a similarity measurement quantifies the resemblance between two objects. The purpose of the similarity analysis is to compare two lists of components and compute a single number that implies their evaluation (Ragnemalm 1993). In regards to CBR, a similarity measurement is a fundamental component that helps in problem-solving since the basic idea of the cost-based argument in cost estimation is the hypothesis that similar problems have similar solutions (Burkhard 2001). Similarity measurements are used in the retrieval of similar cases from the case-base approach. There are two primary retrieval approaches in CBR. The first approach deals with the measure of case similarity by computation of distances between the two cases whereas the other approach operates mostly with a representation or indexing method of the structures. The latter method is more suitable for text-based case applications (Aamodt and Plaza 1994).

Mahalanobis Distance

The Mahalanobis distance is a distance measure between a point and a distribution, introduced by P. C. Mahalanobis in 1936 and is widely used in cluster analysis and classification techniques (McLachlan 1992). This distance measure takes into account correlations between attributes by which different patterns can be identified and analyzed as it is computed using the inverse of variance-covariance matrix of the data set (De Maesschalck 2000). It is considered to be an appropriate measure as it eliminates the unnecessary influence of covariance between attributes (Mahalanobis 1936; Du 2014). Also,

it is useful to determine similarity of an unknown sample set to a known one.

The Mahalanobis distance of a multivariate vector $x = (x_1, x_2, x_3, \dots, x_N)^T$ from a group values with mean $\mu = (\mu_1, \mu_2, \mu_3, \dots, \mu_N)^T$ and covariance matrix S is defined as:

$$D_M(x) = \sqrt{(x - \mu)^T S^{-1} (x - \mu)} \quad (\text{Eq. 5-18})$$

A crucial difference from Euclidean distance is that it considers the correlations of the data set. Mahalanobis distance is affected by both variance and correlation. If the covariance matrix is the identity matrix, then it is the same as Euclidean distance. If covariance matrix is diagonal, then it is called normalized Euclidean distance. However, calculating the variance-covariance matrix using a large number of attributes still remains as limitation because they can contain a great deal of redundant or correlated information (De Maesschalck 2000). Also, the Mahalanobis distance concept can be used where the data set is multivariate normally distributed; in other words, each attribute should be normally distributed (Johnson and Wichern 1988).

Figure 5-8 compares Mahalanobis and Euclidean distance methods. The characteristics of the Mahalanobis distance can be summarized as: 1) reflecting that the variances in each direction are different, 2) accounting for the covariance between attributes, and 3) reducing to the familiar Euclidean distance for uncorrelated attributes with unit variance.

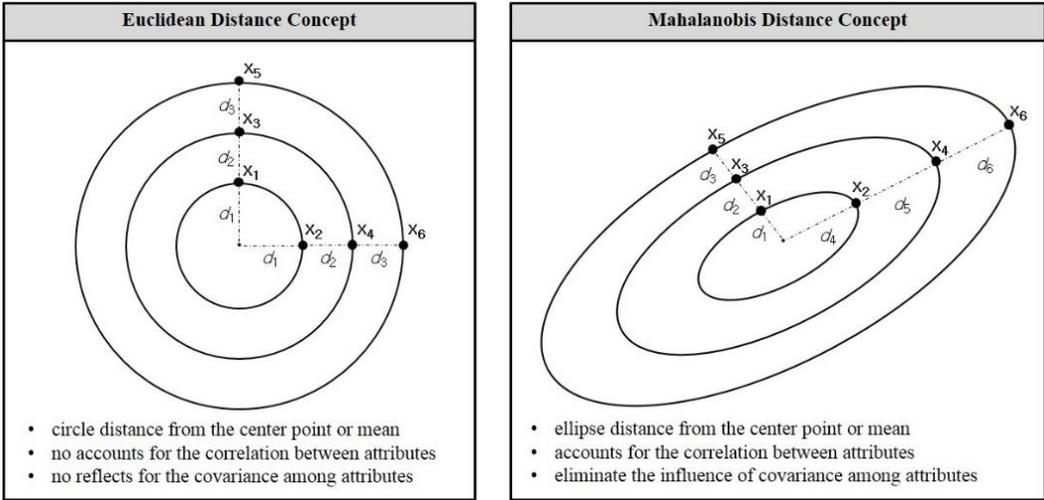


Figure 5-8. Comparison of Mahalanobis Distance and Euclidean Distance

Euclidean Distance

Euclidean space is the most used type of distance measure methods and is based on the location of objects. In Euclidean space, a distance is computed as the square root of the summation of squares of the numerical differences between two analogous objects (Deza and Deza 2009). In accordance with this method, the most fundamental procedure describing the relationship between the two cases, which constitutes the neighboring figures of an arbitrary case, are referred to as the standard Euclidean distance (Yau and Yang 1998). In CBR, Euclidean space is used to represent more complex, symbolic representations such as assumptions (Ji et al. 2011a). Most importantly, it is possible to define the weighted Euclidean distance in form of an equation. Euclidean space provides a good similarity measure since the concept of invariance is considered (Ragnemalm 1993).

1) *Weighted Mahalanobis Distance-based Similarity Measure*

$$WMDIS_{i0} = \sqrt{(x_i - x_0)^t W S^{-1} (x_i - x_0)} \quad (\text{Eq. 5-19})$$

$$= \sqrt{\sum_{j=1}^m w_j s_{(jj)}^{-1} (x_{ij} - x_{0j})^2 + 2 \sum_j \sum_{<k} s_{(jk)}^{-1} (x_{ij} - x_{0j})(x_{ik} - x_{0k})} \quad (\text{Eq. 5-20})$$

Where S is the {sample} covariance matrix, and $S_{(jk)}^{-1}$ denotes the (j, k) element of S^{-1} . If W is I , then $WMDIS_{i0}$ reduces to the Mahalanobis distance. If S is I , the unit matrix, then Mahalanobis distance reduces to the Euclidean distance. The term 'weighted Mahalanobis distance' is used sometimes for the meaning of the weighted sum of two Mahalanobis distances, for example, as in Wang (2007). We are using the term following the usage of Younis (1998). Then similarity measure based on the weighted Mahalanobis distance can be defined as:

$$SIM_{i0}^M = 1 - WMDIS_{i0} = \sqrt{\sum_j w_j s_{(jj)}^{-1} (x_{ij} - x_{0j})^2 + 2 \sum_j \sum_{<k} s_{(jk)}^{-1} (x_{ij} - x_{0j})(x_{ik} - x_{0k})} \quad (\text{Eq. 5-21})$$

2) *Weighted Euclidean Distance-based Similarity Measure*

The weighted Euclidean distance is defined as:

$$WDIS_{i0} = \sqrt{\sum_{j=1}^m w_j (x_{ij} - x_{0j})^2} = \sqrt{(\mathbf{x}_i - \mathbf{x}_0)^T W (\mathbf{x}_i - \mathbf{x}_0)}, \quad (\text{Eq. 5-22})$$

where W is the diagonal matrix whose (j, j) element is w_j . If W is I , then $WMDIS_{i0}$ reduces to the Euclidean distance. Then similarity measure based on the weighted Euclidean distance can be defined as:

$$SIM_{i0}^E = 1 - WDIS_{i0} = 1 - \sqrt{\sum_{j=1}^m w_j (x_{ij} - x_{0j})^2} \quad (\text{Eq. 5-23})$$

This similarity measure may apply to the interval/ratio data, and the 0-1 data, and the ordinal data also, if they can be treated as the interval data.

3) *Arithmetic Summation-based Similarity Measure*

The arithmetic summation-based similarity measure is defined as:

$$SIM_{i0}^A = \sum_{j=1}^m w_j \left(1 - \frac{|x_{ij} - x_{0j}|}{x_j^{\max} - x_j^{\min}}\right) \times 100 \quad (\text{Eq. 5-24})$$

$$= \sum_{j=1}^m w_j \left(\frac{x_j^{\max} - x_j^{\min} - |x_{ij} - x_{0j}|}{x_j^{\max} - x_j^{\min}}\right) \times 100 \quad (\text{Eq. 5-25})$$

The expression in parentheses () measures the similarity between the new case (case 0) and the retrieved case (case i) for the j-th attribute. This similarity

measure may apply to the interval/ratio data, and the 0-1 data, and the ordinal data also, if they can be treated as the interval data.

4) *Fractional Function-based Similarity Measure*

The fractional function-based similarity measure is defined as:

$$SIM_{i0}^F = \sum_{j=1}^m w_j \left(1 + \frac{|x_{ij} - x_{oj}|}{x_j^{max} - x_j^{min}}\right)^{-1} \times 100 \quad (\text{Eq. 5-26})$$

$$= \sum_{j=1}^m w_j \left(\frac{x_j^{max} - x_j^{min}}{x_i^{max} - x_i^{min} + |x_{ij} - x_{oi}|}\right) \times 100 \quad (\text{Eq. 5-27})$$

The expression in parentheses () of the second equation also measures the similarity between the new case (case 0) and the retrieved case (case i) for the j-th attribute. This similarity measure may apply to the interval/ratio data, and the 0-1 data, and the ordinal data also, if they can be treated as the interval data.

Note the $SIM_{i0}^F \geq SIM_{i0}^A$.

Chapter 6. Case Studies

In this chapter, the proposed front-end cost estimation methodology by selective case-based reasoning is validated for multi-family housing, military barrack, and government office projects. The case studies are conducted to examine the accuracy and stability of cost estimation results by the suggested CBR model. More importantly, the level of flexibility of the selective CBR model which provides the most accurate and stable normalization methods (from *sub-module 1*), attribute weighting method (from *sub-module 2*), and similarity measurement method (from *sub-module 3*) according to different types of building projects (or different characteristics of databases) is tested.

6.1 Validation Methods and Process

Table 6-1 shows the validation methods and process to examine the proposed front-end cost estimation methodology by selective CBR. For each normalization method selection (*sub-module 1*), attribute weighting method selection (*sub-module 2*), and similarity measurement method selection (*sub-module 3*), experiment conditions are set. As an example of understanding Table 6-1, validation procedures of normalization method selection (*sub-module 1*) for multi-family housings are explained as below.

Table 6-1 Validation Methods and Process for Case Studies

	Normalization Method Selection (Sub-Module 1)	Attribute Weighting Method Selection (Sub-Module 2)	Similarity Measurement Method Selection (Sub-Module 3)
Case-Base	<ol style="list-style-type: none"> 1. Multi-family housings (100 cases) 2. Military barrack (117 cases) 3. Government office (52 cases) 	<ol style="list-style-type: none"> 1. 100 Multi-family housings (100 cases) 2. Military barrack (117 cases) 3. Government office (52 cases) 	<ol style="list-style-type: none"> 1. 100 Multi-family housings (100 cases) 2. Military barrack (117 cases) 3. Government office (52 cases)
Normalization Method	<ul style="list-style-type: none"> • Interval Normalization • Gaussian Distribution Norm. • Z-score Normalization • Logistic Function Norm. • Ratio Normalization 	<ul style="list-style-type: none"> • Ratio Normalization 	<ul style="list-style-type: none"> • Ratio Normalization
Attribute Weight Assignment Method	<ul style="list-style-type: none"> • GA 	<ul style="list-style-type: none"> • Attribute Impact (AI) • Entropy • Feature Counting (FC) • Genetic Algorithms (GA) 	<ul style="list-style-type: none"> • GA
Similarity Measurement	<ul style="list-style-type: none"> • Euclidean Distance 	<ul style="list-style-type: none"> • Euclidean Distance 	<ul style="list-style-type: none"> • Mahalanobis Distance • Euclidean Distance • Arithmetic Summation • Fractional Function
Validation Method	<ul style="list-style-type: none"> • LOOCV 	<ul style="list-style-type: none"> • LOOCV 	<ul style="list-style-type: none"> • LOOCV
K-NN	<ul style="list-style-type: none"> • 1,2,5,10-NN 	<ul style="list-style-type: none"> • 1,2,5,10-NN 	<ul style="list-style-type: none"> • 1,2,5,10-NN
Performance Measure	<ul style="list-style-type: none"> • MAER, MSD, MAD, SD • Kernel Density Estimation 	<ul style="list-style-type: none"> • MAER, MSD, MAD, SD 	<ul style="list-style-type: none"> • MAER, MSD, MAD, SD

- 1) Case-base of multi-family housings are normalized by five normalization methods (interval, Gaussian, Z-score, logistic, and ratio)
- 2) The each normalized five case-bases are utilized to retrieve similar cases of k-NN (Nearest Neighbor) of 1, 2, 5, and 10 using Euclidean distance-based similarity measures and genetic algorithms as an attribute weight assigning method, respectively.

The K-nearest neighbor principle is an approach used to compare the effectiveness of different models using some specified criterion such as accuracy estimation. The principle involves using a distance measure to locate the k nearest cases with respect to the current input case (Beyer et al. 1999). Then, a retrieval case is selected. This is the class selected from the majority of the k cases of the entire population.

- 3) The retrieved cases of 1, 2, 5, and 10-NN are reused to estimate the average costs.
- 4) Furthermore, based on the retrieved cases of k-NN, performance of the CBR cost estimation models are evaluated and compared in terms of accuracy and stability
- 5) Specifically, the comparisons of MAER, MSD, MAD, and SD of cost estimation results as performance indicators are defined in Equations 6-1~6-4.

$$MAER(\%) = \sum_{i=1}^n \frac{1}{n} \left| \frac{c_i - \hat{c}_i}{c_i} \right| \times 100, \hat{c}_i: \text{estimated or hypothetical cost} \quad (\text{Eq. 6-1})$$

MAER are the ratios use to quantify how close the predictions or estimates are as compared to the target data. MAER is a simple method for evaluating the accuracy of single sequences and is easy to comprehend and calculate. MAER is scale dependent and cannot be interrelated across series (Black 2014).

$$MSD(\%) = \sum_{i=1}^n \frac{1}{n} (c_i - \hat{c}_i)^2, \hat{c}_i: \text{estimated or hypothetical cost} \quad (\text{Eq.6-2})$$

MSD can be expressed as the second moment of a given set of observations made from an arbitrary origin. If the stated source represents the mean of the set, then the equivalent of the variance computed from the set of observations is the MSD (Black 2014).

$$MAD(\%) = \sum_{i=1}^n \frac{1}{n} |c_i - \hat{c}_i|, \hat{c}_i: \text{estimated or hypothetical cost} \quad (\text{Eq. 6-3})$$

MAD is simply the estimated average distance for all the elements in the data set from the cumulative mean computed from the same data set. MAD is obtained in a three-step approach: calculation of the total mean; finding the absolute deviation; and finally computing the mean deviation using the absolute deviation figures obtained (Black 2014).

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^n (x_i - \mu)^2}, \sigma: \text{standard deviation} \quad (\text{Eq. 6-4})$$

SD is a measure used to determine the level of variation in a group of data values. SD is based on the mean. If the obtained SD is closer to '0', it shows that the data sets are closer to the mean. A dataset whose values are dispersed over a wider range of values generates a high SD (Black 2014).

- 6) The LOOCV method based on the case-base of 100 multi-family housings is adopted in this study to validate the five different normalization method-based CBR cost models. In statistics, there are different techniques used to validate models and assess the possibility of data analysis to generalize and give an independent set of data. Cross-validation is one of the widely used models mostly in predictions and estimations (Black 2014).

Specifically, LOOCV is k-fold cross validation where k equals to the total number of examples in dataset. For example, with total number of N cases in the case-base, N experiments are conducted to test a model. For each attempt of experiment to test a case, N-1 cases are used for performance measuring. LOOCV validation method provides reasonable criterion of model selection since this method reduces randomness; thus lead to unbiased results (Cawley and Talbot 2004). LOOCV method allows increase of experiment attempts using large numbers of training data. However, this method requires more computing time to run N experiments.

- 7) Kernel density estimation (KDE) constitutes one of the ways to evaluate the function of a probability density function. KDE is a non-parametric technique and is used mostly for random variables. KDE is an essential tool in data smoothening, which enables the population inference to be performed on any fixed data sample (Sheather and Jones 1991).

KDE are related to histograms, although they are superior. In fact, KDE is a preferred technique because it alleviates estimation problems

encountered when using histograms. The results using histograms are not smooth as compared to those of KDE. Furthermore, histograms depend on the end points and width of the data points. With kernel density estimators, the resulting curves are smooth with no end points and greatly depend on the bandwidths (Botev et al. 2010).

The use of kernel density estimators has proved effective in bypassing challenges exhibited by histograms. For instance, the KDE solves the problem of roughness in histograms. KDE is also less dependent on end points of bins as opposed to histograms. To remove the problem of reliance on end points, KDE focuses on each data point instead of fixing endpoints of the blocks. However, the KDE approach cannot be used to remove reliance on bandwidth. The choice of the correct bandwidth can thus be accomplished through various methods. For instance, a diagonal bandwidth matrix can be used.

6.2 Multi-Family Housing (MFH)

6.2.1 Case-base Profile

Korean housing supply legislation requires the use of unit gross area in the construction of apartment houses in Korea. Thus, contractors can analyze the cost data of apartment building projects and can build a database according to unit types, since we can expect similar patterns. We base the data analysis of the research on building cost data of 100 multi-family housing buildings from 15 housing complex projects in Korea.

Data come from the Seoul Housing Corporation, a Korean public enterprise established in 1989. To elaborate, public owners prepare priced bills of quantities, which contain the total expenditures including all input, such as labor and materials, to use as a vital standard to estimate a precise budget. Construction firms also use these priced bills of quantities to determine a total fixed price for their bid proposals, and they receive awards based on their bids.

Through data analysis, we find four types of apartment households (type 49m², 59m², 84m², and 114m²), and either a single type or a combination of different types make up the apartment buildings. To improve the convenience of data analysis and the accuracy of estimation of building cost, we separate all of the historical data and classify it into the singular type of unit gross area, according to each unit type. Furthermore, the gross floor area ratio, which represents an area ratio of a certain unit type that takes up among an apartment building, is adopted into takeoff priced quantities of each item as priced bills of

quantities are structured with trade sections and grouped by the sum of quantities of each input item (Eq. 6-5).

$$\text{Cost of Each Type} = \frac{\text{Total cost of apartment building} \times \text{GFA ratio}}{\text{Number of households for each type}} \quad (\text{Eq. 6-5})$$

After analyzing the cost data, we need to normalize this data in terms of escalation, regional location, and system specification. However, we only conduct normalization regarding escalation in this research as there is little point in normalizing for regional location and/or system specification, mainly because of Korea's relatively small territory. We execute normalization apropos of escalation on each building type using a Korean construction cost index published every four months by the Korea Institute of Construction Technology (KICT). Using the index provided, we normalize cost data for the years 2005 and 2007 to the year 2008 by multiplying the index of 122.17 and 115.53, respectively.

Extracted twelve attributes are as follows: 1) number of households, 2) gross floor area, 3) number of unit floor households, 4) number of elevators, 5) number of floors, 6) number of piloti with household scale, 7) number of households of unit floor per elevator, 8) height between stories, 9) depth of pit, 10) roof type, 11) hallway type, and 12) Structure type.

Table 6-2. Attributes for Multi-Family Housing

No.	Attribute Name	Scale Type
X1	No. of Households	Ratio Scale
X2	Gross Floor Area	Ratio Scale
X3	No. of Unit Floor Households	Ratio Scale
X4	No. of Elevators	Ratio Scale
X5	No. of Floors	Ratio Scale
X6	No. of Piloti with Household Scale	Ratio Scale
X7	No. of Households of Unit Floor per Elevator	Ratio Scale
X8	Height between Stories	Ratio Scale
X9	Depth of Pit	Ratio Scale
X10	Roof Type	Nominal Scale
X11	Hallway Type	Nominal Scale
X12	Structure Type	Ratio Scale (index)
X13	Cost	Ratio Scale

Attribute weights obtained by AI, entropy, FC, and GA for multi-family housing are summarized in Table 6-3.

Table 6-3. Attribute Weights by AI, Entropy, FC, and GA (MFH)

No.	AI	Entropy	FC	GA
X1	0.107	0.089	0.083	0.019
X2	0.444	0.090	0.083	0.361
X3	0.041	0.090	0.083	0.176
X4	0.002	0.088	0.083	0.004
X5	0.161	0.090	0.083	0.292
X6	0.091	0.072	0.083	0.007
X7	0.010	0.088	0.083	0.041
X8	0.030	0.091	0.083	0.002
X9	0.006	0.090	0.083	0.006
X10	0.026	0.059	0.083	0.007
X11	0.002	0.063	0.083	0.024
X12	0.081	0.090	0.083	0.061

The matrix plot from the case-base of multi-family housings to describe relationships among attributes are analyzed in Figure 6-1. The figure indicates thirteen attributes with their corresponding type of scale. Two of the attributes have nominal scale while the others have ratio. Each attribute is assigned a number to help in its representation in the matrix plot.

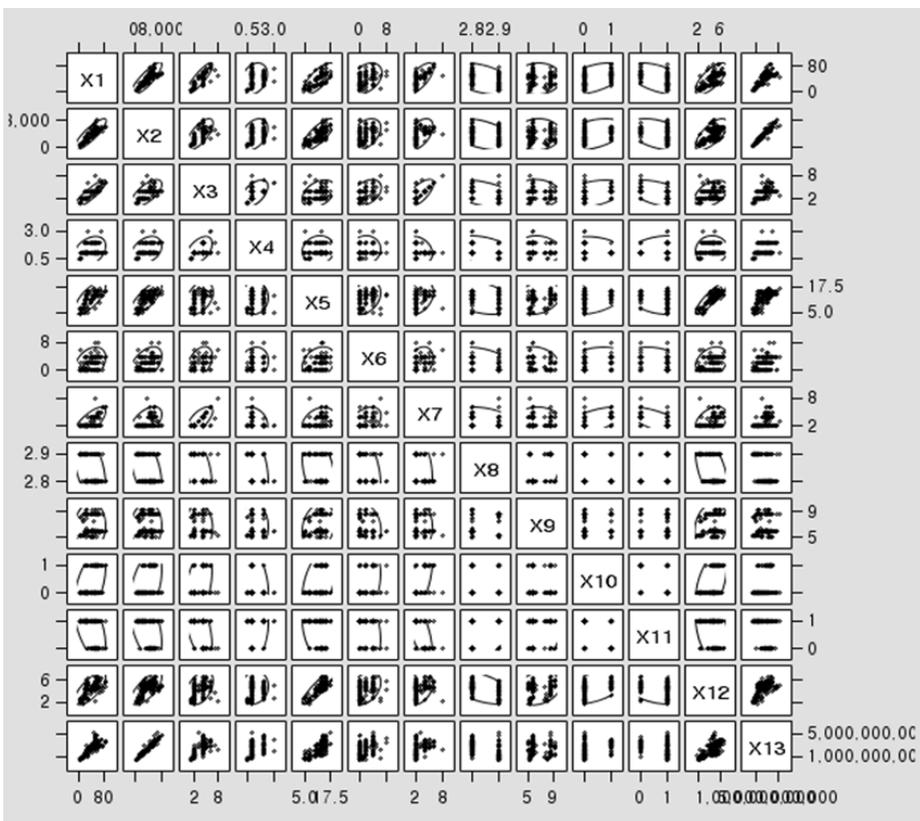


Figure 6-1. Matrix Plot (MFH)

For instance, in the matrix plot the attribute for the number of housing is represented by X1 while the attribute for number of unit floor households is represented by X3. On the same note, X13 is used to represent the cost. Notably, the relationships of all the attributes relate to multi-family housings are described in the matrix plot. Ideally, the matrix plot makes analysis of the relationship that exists between similar attributes in different multi-family housing cases.

As from the matrix plot, the number of households ranges from 0 to 80, and there is no ratio between the attributes. X2 represents the gross floor area which is the total area on which the households are built. Notably, the ratio between the number of the households and the area occupied is 0.800 which indicate that more lands is used to build the households. X3 represents the number of unit floor households and from the matrix, it is apparent that the largest house has eight units while the smallest has two units. The number of elevators is represented by X4 in the matrix, and their ratio ranges from 0.5 to 3.0. The minimum number of floors (X5) in each house is five while the maximum number of houses is 17.5. Additionally, the maximum number of the pilot with household scale (X6) is eight. Similarly, the highest number of households of unit floor in the elevator is eight while the lowest is two. The height between stories (X8) ranges from 2.8 to 2.9 while the depth of pit (X9) ranges from five to nine units. The roof type (X10) is 1 unit just like the hallway type (X11). The structure type (X12) varies from two to six.

6.2.2 Results and Discussions

MARE, MSD, MAD, and SD for Norm. Method (MFH)

Table 6-4 shows the unique characteristics resulting from various methods of normalization. All the MAER calculated show positive values except the Z-score method. The Gaussian normalization approach had the highest values of MAER while the MAER calculated with the ratio method has the lowest values when $k=1, 2, 5$ and 10 . The interval values ranged from 0.173 to 0.2 did not have consistency with the increase in the k -values. The Gaussian normalization, on that other hand had values ranging from 0.244 and 0.374 and there was no consistent as well. The Z-score values ranged from -0.497 and -0.080 . The logistic normalization values ranging from 0.127 to 0.145 and the values rose with an increase in the k -value while those of ratio normalization varied between 0.077 and 0.078 . Therefore, the ratio normalization method had the highest accuracy and consistence compared to the interval, Gaussian, Z-score and logistic methods.

Table 6-4. MAER for Normalization Methods (MFH)

MAER	k=1	k=2	k=5	k=10
Interval	0.173	0.181	0.176	0.200
Gaussian	0.244	0.316	0.280	0.374
Z-score	-0.497	-0.340	-0.080	-0.183
Logistic	0.127	0.130	0.130	0.145
Ratio	0.078	0.077	0.078	0.078

The MSD values in table 6-5 show that the ratio method had the lowest values while the Z-score had the highest ones (Weir and Cockerham 1984). There were no negative values in the approach results when k=1, 2, 5 and 10. Also, no correlation was established between the increase/decrease of values obtained with the increase/decrease in the value of k. The logistic and ratio approaches showed a stagnation in the values obtained from k=2 to k=10. The values of Interval method ranged from 0.007 to 0.011 while those of the Gaussian method ranged from 0.011 to 0.014. The Z-score test gave results ranging from 0.210 to 0.306 that the values of the ratio method ranged from 0.004 to 0.006. Consistency was established in all the methods other than the Z-score method. The low MAD value is an indication of a narrow the range of the estimate errors (Coscoy et al. 2007). Being stable and consistent and having the lowest values indicated that the ratio method gave the best accuracy under CBR model.

Table 6-5. MSD for Normalization Methods (MFH)

MSD	k=1	k=2	k=5	k=10
Interval	0.011	0.007	0.008	0.007
Gaussian	0.011	0.014	0.013	0.012
Z-score	0.306	0.220	0.225	0.210
Logistic	0.008	0.007	0.007	0.007
Ratio	0.006	0.004	0.004	0.004

In the MAD data analysis as shown in Table 6-6, the ratio method gave the

lowest values ranging from 0.043 and 0.044. The ratio, Gaussian, interval and logistic tests show low values that are inconsistent with the value of k hence the MAD based on them could give relatively accurate results. The Z-score test gave the highest values in the normalization method with values ranging from 0.286 and 0.298 from k=1, 2, 5 and 10 making it the most erroneous model to use in the MAD. There was no correlation between the values obtained and the k values in all the tests. Having the lowest values that indicate a low range of errors, the ratio normalization-based CBR cost model for multi-family housing was the most accurate among the five methods in fitting models, smoothing and forecasting (Soukoreff and MacKenzie 2003).

Table 6-6. MAD for Normalization Methods (MFH)

MAD	k=1	k=2	k=5	k=10
Interval	0.056	0.058	0.055	0.055
Gaussian	0.063	0.082	0.079	0.076
Z-score	0.298	0.293	0.297	0.286
Logistic	0.051	0.058	0.057	0.058
Ratio	0.044	0.043	0.044	0.043

The SD results for normalization method in Table 6-7 shows the values obtained using all the normalization methods are stale. There is no method that gives negative results too. The Z-score normalization method has the highest values ranging from 0.459 and 0.556 hence a model created from it has the highest errors. The ratio method, on the other hand has the lowest values

ranging from 0.064 to 0.080 indicating that the ratio-based model gives values closest to the mean among the five methods .The values of all tests fluctuate inconsistently and are independent of the size of the k-values. The higher the values of the SD, the higher the dispersion between the datasets from the mean and vice versa. The datasets that is closer to the mean represents a more stable CBR cost model. Having the lowest value, the ratio normalization-based CBR model is more accurate than all other methods.

Table 6-7. SD for Normalization Methods (MFH)

SD	k=1	k=2	k=5	k=10
Interval	0.104	0.087	0.087	0.087
Gaussian	0.105	0.120	0.115	0.115
Z-score	0.556	0.471	0.477	0.459
Logistic	0.087	0.084	0.083	0.083
Ratio	0.080	0.064	0.067	0.067

The ratio normalization method had the highest accuracy in the CBR cost model for multi-family housing compared to the interval, Gaussian, Z-score and logistic methods in MAER. Also, the ratio method yielded the lowest values for MSD and MAD. The ratio model was the most accurate among the five methods.

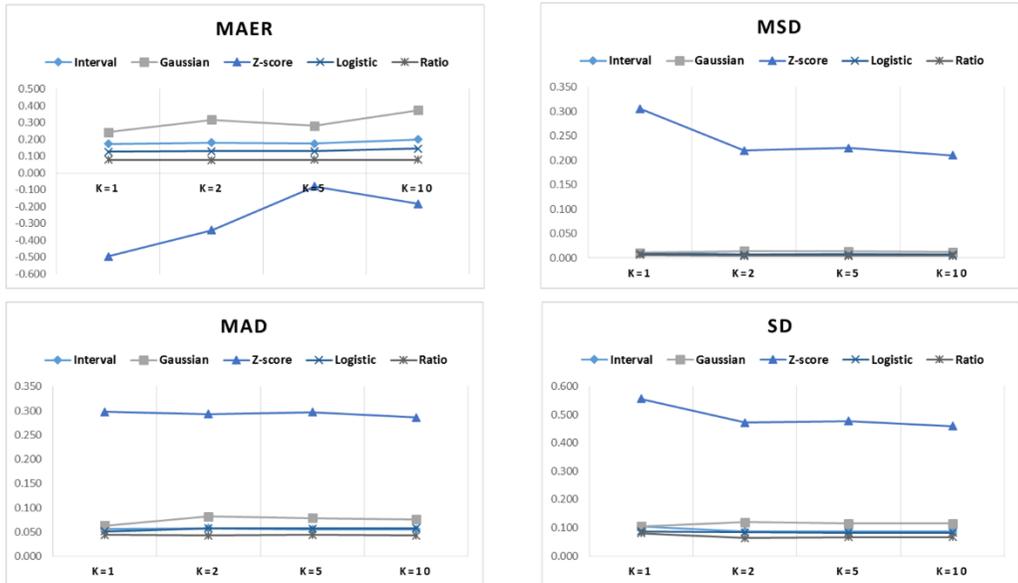


Figure 6-2. MAER, MSD, MAD, and SD for Norm. Methods (MFH)

Kernel Density Estimation (MFH)

As illustrated in Figure 6-3, density estimations have been conducted using different methods including the Gaussian, logistic, Z-score, ratio, and interval normalization. In regards to the original score, various methods display different graph trends. For example, using the Gaussian model, the density obtained is under-smoothed because the bandwidth is too small with six modes. Similarly, in the case of the logistic form, the bandwidth is increased and thus the estimate is flatter with three modes. This situation is considered overestimated since the bandwidth is too large and obscures most of the data structure (Sheather and Jones 1991). However, graphs using the ratio norm and interval norm provide an optimally smooth kernel estimates with fewer modes. In this case, the value of bandwidth minimizes the error between the estimated density and the actual density (Sheather and Jones 1991; Wand and Jones 1995).

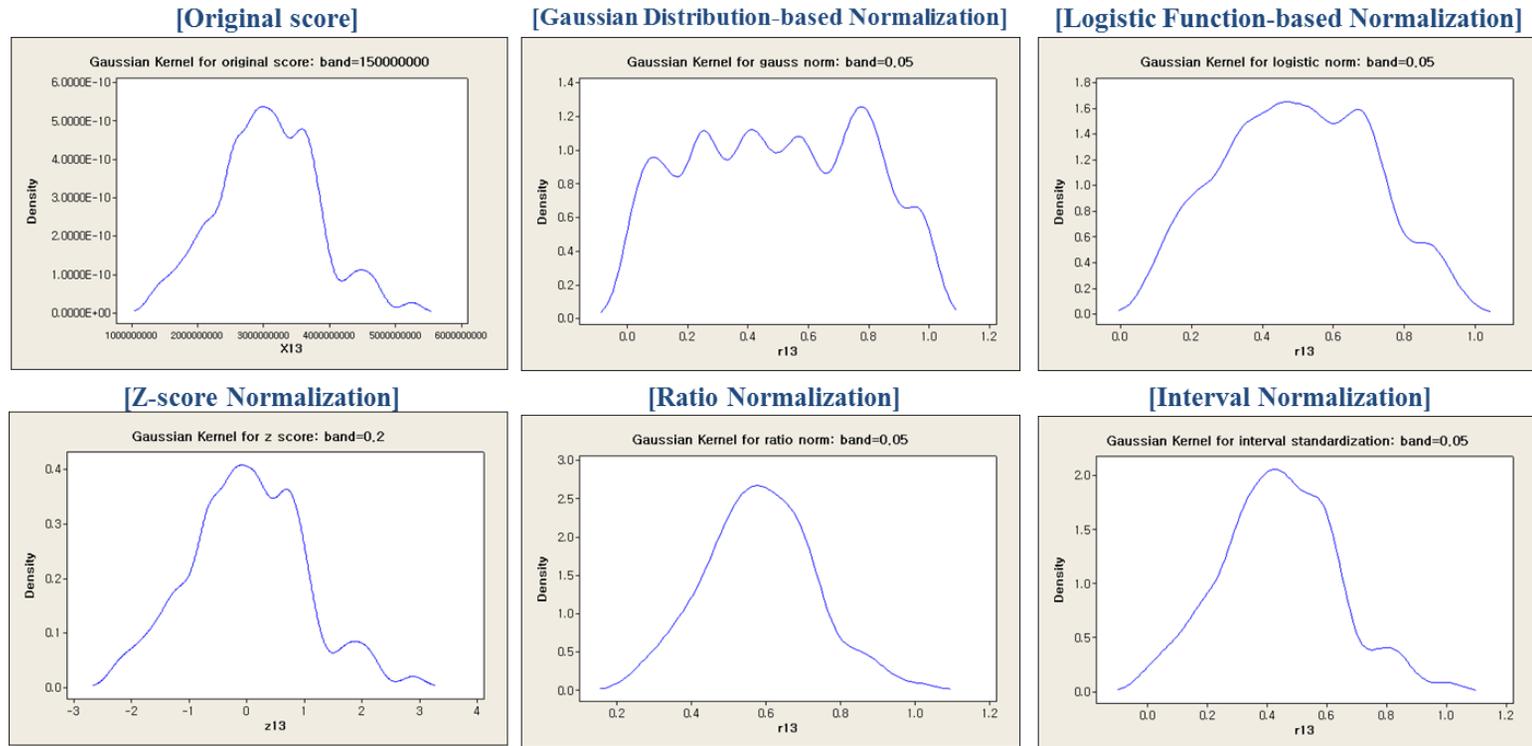


Figure 6-3. Kernel Density Estimation for Normalized Case-Bases (MFH)

MARE, MSD, MAD, and SD for AW Method (MFH)

In the attribute weighting method, in Table 6-8, the AI, GA and entropy methods have the lowest values. However, the FC method shows significantly high values for k=1, 2, 5 and 10. The higher the value, the lower the accuracy in the weight assignment. The GA method, having the lowest values that range between 0.077 and 0.078 has the lowest errors. The FC values are the greatest ones ranging from 0.113 and 0.142. All the methods within the attribute weighting show no correlation between the value obtained and the value of k. The AI, entropy and GA methods have low values hence can be used for accurate weight assignment. Having the lowest value, the GA is the best method of assigning weight in CBR model for multi-family housing under MAER.

Table 6-8. MAER for Attribute Weighting Method (MFH)

MAER	k=1	k=2	k=5	k=10
AI	0.092	0.083	0.084	0.090
Entropy	0.102	0.113	0.119	0.134
FC	0.113	0.114	0.123	0.142
GA	0.078	0.077	0.078	0.078

Attribute weight assignment under MSD in Table 6-9 indicates that all the methods have low and stable values. The AI and GA methods have lower values than the FC and entropy methods. The GA method has the lowest values ranging between 0.004 and 0.006 among the four methods when k=1, 2, 5 and 10. On the other hand, the FC method has the highest values under MSD. The entropy

and FC values fluctuate while the GA and AI methods' show a constant values between k=2 to k=10. Having the lowest values under MSD, the GA is the best method of assigning weight because it has the lowest errors in the CBR model.

Table 6-9. MSD for Attribute Weighting Method (MFH)

MSD	k=1	k=2	k=5	k=10
AI	0.008	0.005	0.005	0.005
Entropy	0.010	0.009	0.010	0.010
FC	0.013	0.009	0.010	0.011
GA	0.006	0.004	0.004	0.004

Table 6-10. MAD for Attribute Weighting Method (MFH)

MAD	k=1	k=2	k=5	k=10
AI	0.049	0.046	0.046	0.049
Entropy	0.056	0.063	0.066	0.071
FC	0.063	0.063	0.067	0.075
GA	0.044	0.043	0.044	0.043

Table 6-10 shows the MAD values obtained from various methods for attribute weight assignment. From the results, all the methods yielded low and stable values hence they are likely to be accurate in weight assignment (Jain et al. 2005). The FC method has the highest values while the GA method has the lowest ones. All the tests show insignificant variation in the values from k=2 to

k=10. The AI and GA methods show the slightest variation. Therefore, GA-based CBR model under MAD for multi-family housing is GA.

In Table 6-11, the SD values for attribute weight assignment for four methods are given. All the methods display significantly low and stable values for k=1, 2, 5 and 10. Among the four methods, the GA displayed the lowest values while the FC method displays the highest ones. The values for entropy method are constant in all the values of k while the values of FC and AI method fall and become constant at k=5. Given that GA has the lowest SD values, the GA-based CBR model is expected to yield more stable cost estimation results for multi-family housing.

Table 6-11. SD for Attribute Weighting Method (MFH)

SD	k=1	k=2	k=5	k=10
AI	0.090	0.074	0.073	0.073
Entropy	0.098	0.098	0.098	0.098
FC	0.115	0.097	0.100	0.100
GA	0.080	0.064	0.067	0.067

The GA method of Attribute Weight Assignment is the most accurate method under MAER, MSD and SD since it has the lowest values that indicate the lowest errors in its application. In the MAD, both the AI and GA methods showed the highest stability and accuracy; however, GA gives the highest accuracy during application.

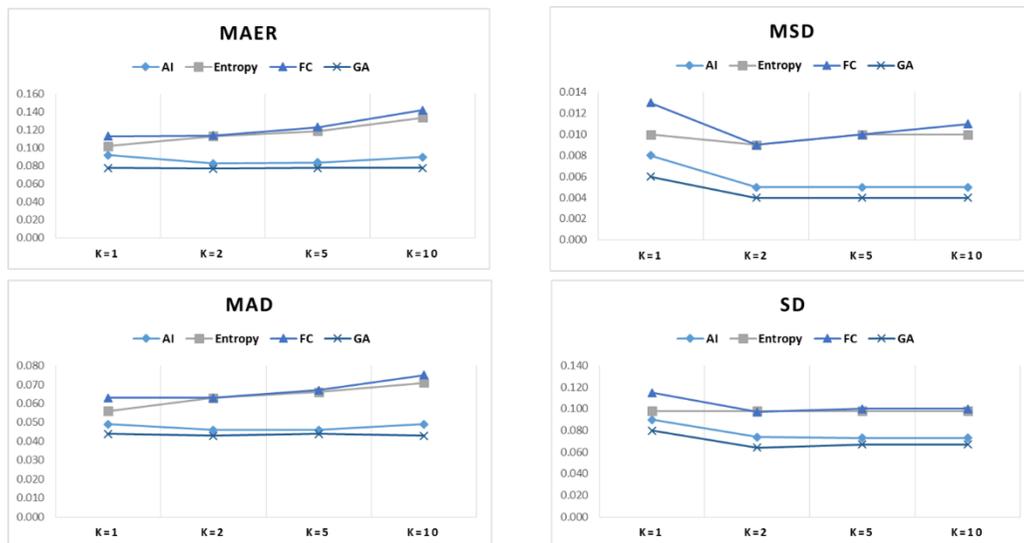


Figure 6-4. MAER, MSD, MAD, and SD for AW Method (MFH)

MARE, MSD, MAD, and SD for SM Method (MFH)

Among the MAER of similarity measurement methods shown in Table 6-12, the arithmetic method has the lowest values ranging from 0.070 to 0.074. The Mahalanobis similarity measurement method, on the other hand, has the highest values ranging from 0.076 to 0.095. The Euclidian, Mahalanobis and arithmetic methods shows lack of correlation between the method values and the values of k. On the other hand, the fractional method has the highest value at k=1 while the other values are similar regardless of the value of k (Diebold and Mariano 2012). Given that the arithmetic method has the lowest values; it is the most accurate method of similarity measurement under MAER.

Table 6-12. MAER for Similarity Measurement Methods (MFH)

MAER	k=1	k=2	k=5	k=10
Mahalanobis	0.076	0.082	0.089	0.095
Euclidean	0.078	0.077	0.078	0.078
Arithmetic	0.072	0.070	0.072	0.074
Fractional	0.083	0.073	0.073	0.073

Under MSD of similarity measurement in Table 6-13, the Euclidian, arithmetic and fractional methods have small and equivalent values for k=1, 2, 5 and 10; hence, they can yield low errors. The Mahalanobis value are different but are higher than the other methods and can have a higher error rate when employed to measure similarity than the other three methods. Therefore, all the similarity measurement methods are appropriate to use under MSD; however, the Mahalanobis-based CBR model can have higher errors.

Table 6-13. MSD for Similarity Measurement Methods (MFH)

MSD	k=1	k=2	k=5	k=10
Mahalanobis	0.006	0.005	0.006	0.005
Euclidean	0.006	0.004	0.004	0.004
Arithmetic	0.006	0.004	0.004	0.004
Fractional	0.006	0.004	0.004	0.004

The results in Table 6-14 show the MAD values obtained in the similarity measurement. All the methods yield significantly small values hence their error

rates in the CBR model are also low. The Euclidian and Mahalanobis values are equivalent for k=1, 2, 5 and 10. On the other hand, the fractional method gives the highest values for k=1 (Chiarandini et al. 2005). The Arithmetic method that has the lowest value is the most appropriate to employ because of its highest stability and accuracy among all the methods.

Table 6-14. MAD for Similarity Measurement Methods (MFH)

MAD	k=1	k=2	k=5	k=10
Mahalanobis	0.043	0.046	0.049	0.051
Euclidean	0.043	0.046	0.049	0.051
Arithmetic	0.040	0.039	0.040	0.041
Fractional	0.045	0.041	0.041	0.040

The SD values in Table 6-15 show that the arithmetic methods have low and stable values. Euclidian and Mahalanobis similarity measurements values are equivalent. The values of arithmetic method begin with 0.075 at k=1 but drops to 0.062 for k=2, 5 and 10.

Table 6-15. SD for Similarity Measurement Methods (MFH)

SD	k=1	k=2	k=5	k=10
Mahalanobis	0.080	0.071	0.075	0.075
Euclidean	0.080	0.071	0.075	0.075
Arithmetic	0.075	0.062	0.062	0.062
Fractional	0.080	0.062	0.060	0.060

The fractional method values begin from 0.80 and drop to 0.62, 0.60 and 0.60 for k=1, 3, 5 and 10 respectively. All the methods have low SD values hence can yield accurate results when used for weight measurement. However, given that the arithmetic method of SD has the lowest values, it can yield the most stable estimate results.

Given that the arithmetic method has the lowest values in the MAER. All the similarity measurement methods are appropriate to use under MSD model. The Arithmetic method is the most appropriate to employ in MAD and SD. It can be established that the arithmetic method-based CBR cost model can yield the highest cost estimate accuracy for multi-family housing.

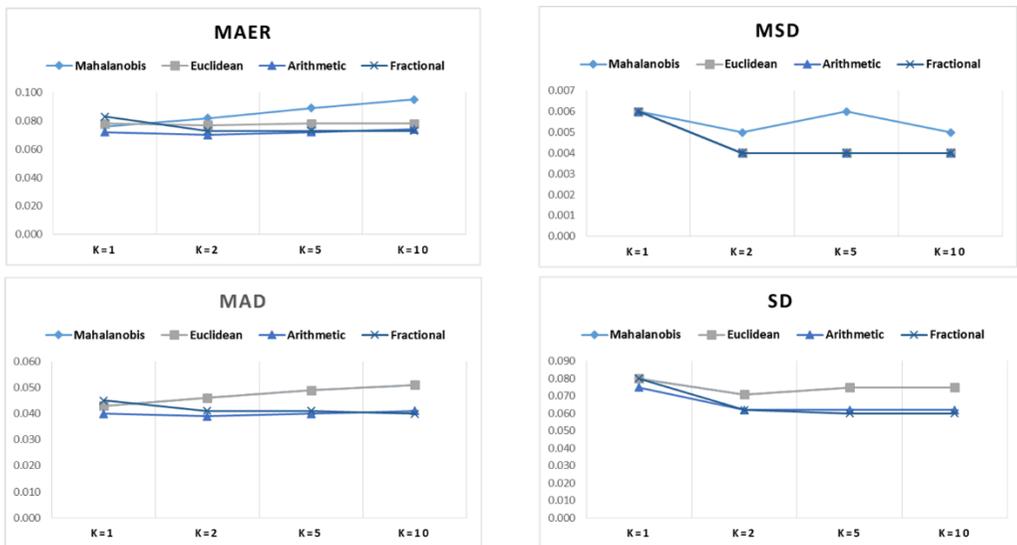


Figure 6-5. MAER, MSD, MAD, and SD for SM Methods (MFH)

6.3 Military Barrack (MB)

6.3.1 Case-base Profile

117 cases of military barrack projects¹ were utilized for validation. Extracted eight attributes are as follows: 1) number of beds, 2) number of floors, 3) gross floor area, 4) unit floor area, 5) quarter area ratio, 6) office area ratio, 7) underground floor, and 8) pit foundation.

Table 6-16. Attributes for Military Barrack

No.	Attribute Name	Scale Type
X1	No. of Beds	Ratio Scale
X2	No. of Floors	Ratio Scale
X3	Gross Floor Area	Ratio Scale
X4	Unit Floor Area	Ratio Scale
X5	Quarter Area Ratio	Ratio Scale
X6	Office Area Ratio	Ratio Scale
X7	Underground Floor	Nominal Scale
X8	Pit Foundation	Nominal Scale

¹ Refer Ji et al. (2011a) for data acquisition and analysis of military barrack projects

Attribute weights obtained by AI, entropy, FC, and GA for military barrack are summarized in Table 6-17.

Table 6-17. Attribute Weights by AI, Entropy, FC, and GA (MB)

No.	AI	Entropy	FC	GA
X1	0.208	0.133	0.125	0.000
X2	0.142	0.149	0.125	0.027
X3	0.293	0.133	0.125	0.780
X4	0.264	0.142	0.125	0.176
X5	0.000	0.147	0.125	0.000
X6	0.014	0.120	0.125	0.000
X7	0.026	0.046	0.125	0.000
X8	0.053	0.129	0.125	0.018

The matrix plot for the Figure 6-6 indicates the relationship between the attribute and their cost. In the matrix plot, the value of the x axis represents the attribute while at the same time they may indicate the quantity of the attribute. Generally, there is a positive relationship between the number of the attributes and the cost incurred. The cost incurred increases as the number of the attributes increases and vice-versa.

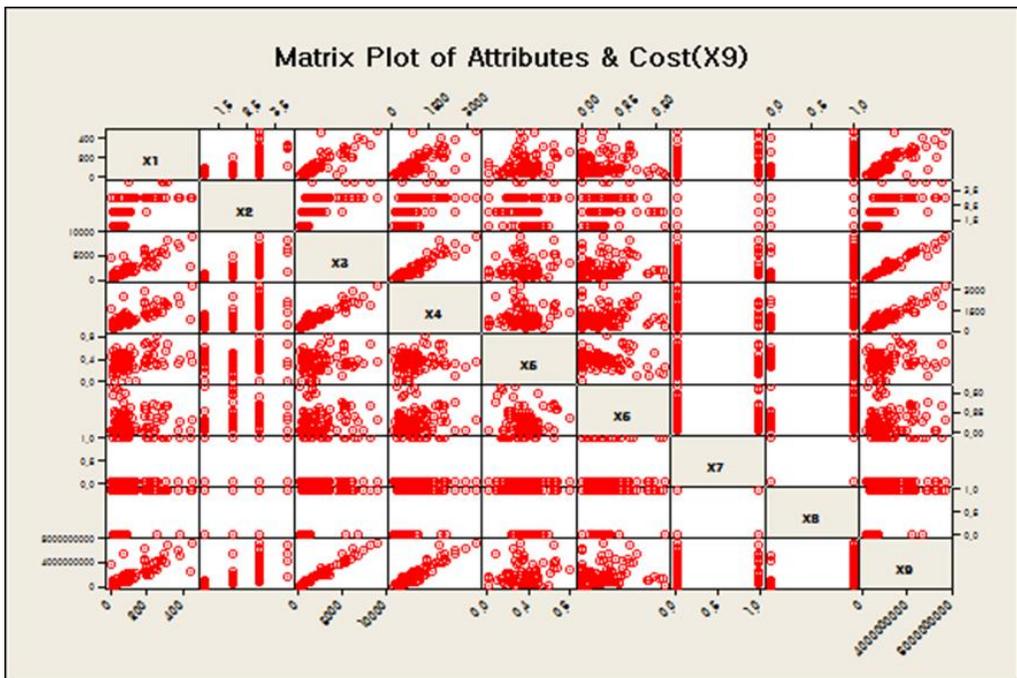


Figure 6-6. Matrix Plot (MB)

6.3.2 Results and Discussions

MARE, MSD, MAD, and SD for Norm. Method (MB)

In the military barrack, the MAER of normalization method shows the logistic and Z-score methods having the lowest values while the greatest value is recorded in the ratio method. There is no negative value in any of the MAER of normalization methods. Therefore, having the lowest value, the logistic and Z-score methods are the most accurate in the MAER for the military barrack.

Table 6-18. MAER for Normalization Methods (MB)

MAER	k=1	k=2	k=5	k=10
Interval	0.149	0.144	0.142	0.166
Gaussian	0.092	0.084	0.081	0.089
Z-score	0.377	0.067	0.068	0.067
Logistic	0.079	0.074	0.080	0.091
Ratio	0.150	0.147	0.152	0.178

In table 6-19, the values of normalization under various methods are established for MSD. The results show that the interval and ratio methods have similar values for k=1, 2, 5 and 10. Among the four normalization methods, the two have the lowest values. The Gaussian method has slightly different results but the difference is insignificant. Therefore, the most appropriate method of normalization to use under MSD in the military barrack are the interval and ratio methods. The best alternative for the two is the Gaussian method and the worst is the Z-score.

Table 6-19. MSD for Normalization Methods (MB)

MSD	k=1	k=2	k=5	k=10
Interval	0.003	0.002	0.002	0.003
Gaussian	0.004	0.002	0.002	0.002
Z-score	0.068	0.043	0.046	0.067
Logistic	0.003	0.003	0.004	0.005
Ratio	0.003	0.002	0.002	0.003

Under MAD, as illustrated in table 6-20, the Z-score normalization method yields the highest values while the ratio scale has the lowest for k=1, 3, 5 and 10. None of the methods shows a correlation between the value of k and the values of normalization obtained. Therefore, in the military barrack, the ratio scale is the most appropriate method to use under MAD and the worst of them is the Z-score method.

Table 6-20. MAD for Normalization Methods (MB)

MAD	k=1	k=2	k=5	k=10
Interval	0.031	0.028	0.028	0.032
Gaussian	0.038	0.033	0.032	0.034
Z-score	0.155	0.131	0.130	0.146
Logistic	0.037	0.036	0.042	0.046
Ratio	0.031	0.027	0.028	0.031

Table 6-21. SD for Normalization Methods (MB)

SD	k=1	k=2	k=5	k=10
Interval	0.051	0.044	0.045	0.045
Gaussian	0.062	0.049	0.046	0.046
Z-score	0.261	0.205	0.211	0.255
Logistic	0.055	0.055	0.064	0.064
Ratio	0.054	0.043	0.045	0.045

Figure 6-21 shows the values obtained for different methods of normalization methods. In the table, the Z-score yield the largest normalization values while the interval method yields the lowest hence the most accurate. However, the ratio scaled results are also closely related to those of the interval scale for k=1, 2, 5 and 10. None of the methods used yields negative values. Therefore, the best method to use in the SD normalization is the interval scale and the best alternative for it is the ratio scale because the two methods yield the lowest results.

The logistic and Z-score methods are the most accurate for the military barrack under MAER. On the other hand, the ratio and interval methods are appropriate under MSD and MAD. The best method to use in the SD is the interval method but the ratio method makes the best alternative because the two methods yield the lowest results.

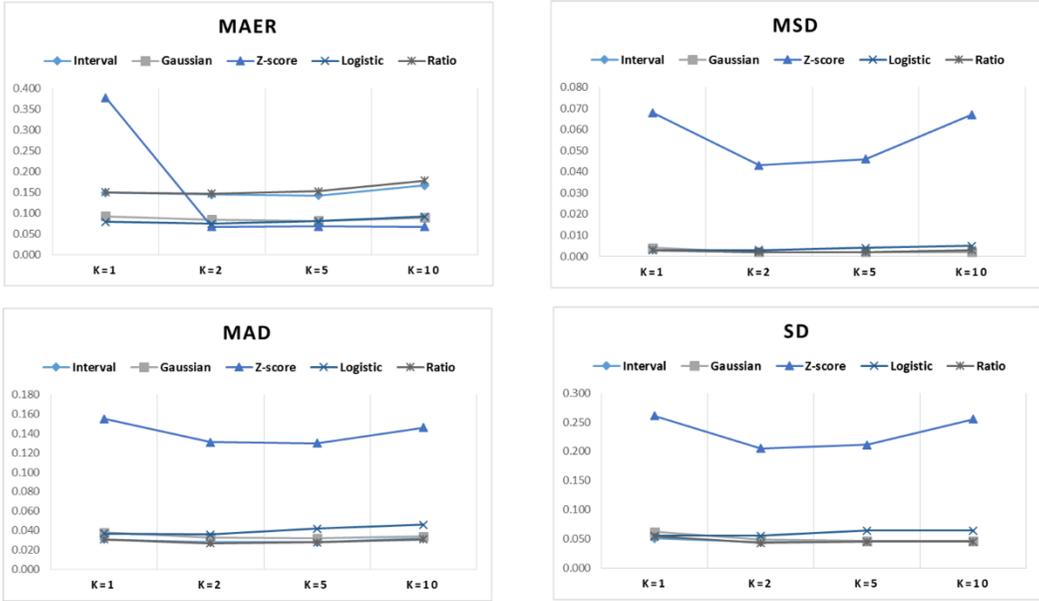


Figure 6-7. MAER, MSD, MAD, and SD for Norm. Methods (MB)

Kernel Density Estimation (MB)

Figure 6-8 indicates comparative use of different methods of kernel density estimation. The specific methods used include Gaussian distribution, ratio normalization, interval normalization, the z-score normalization and the logistic function. While the resultant curves exhibit general resemblance to the original score, considerable deviations exist in the nature of the crests and the smoothness of the curves (Zhang 2015). The original score indicates existence of six crests. On the contrary, the Gaussian normalization results in three crests, with the middle crests being oversmoothed. On the other hand, the logistic normalization produces two crests, the middle crest in the logistic normalization being smoothed out. This extremely smooth curve indicates

overestimation which results from the use of extremely large bandwidths. The large bandwidth thus obscures data structure. The use of z-score results in under-smoothed curve with many peaks (Zhang 2015). This could indicate that the bandwidth used is too small. Ratio normalization results in optimally smooth curve. Similarly, interval normalization produces relatively smooth curves. In these last scenarios, the correct choice of the bandwidth alleviates errors in density estimation. Therefore, ratio and interval normalization techniques are ideal in analysis of results obtained in Figure 6-8.

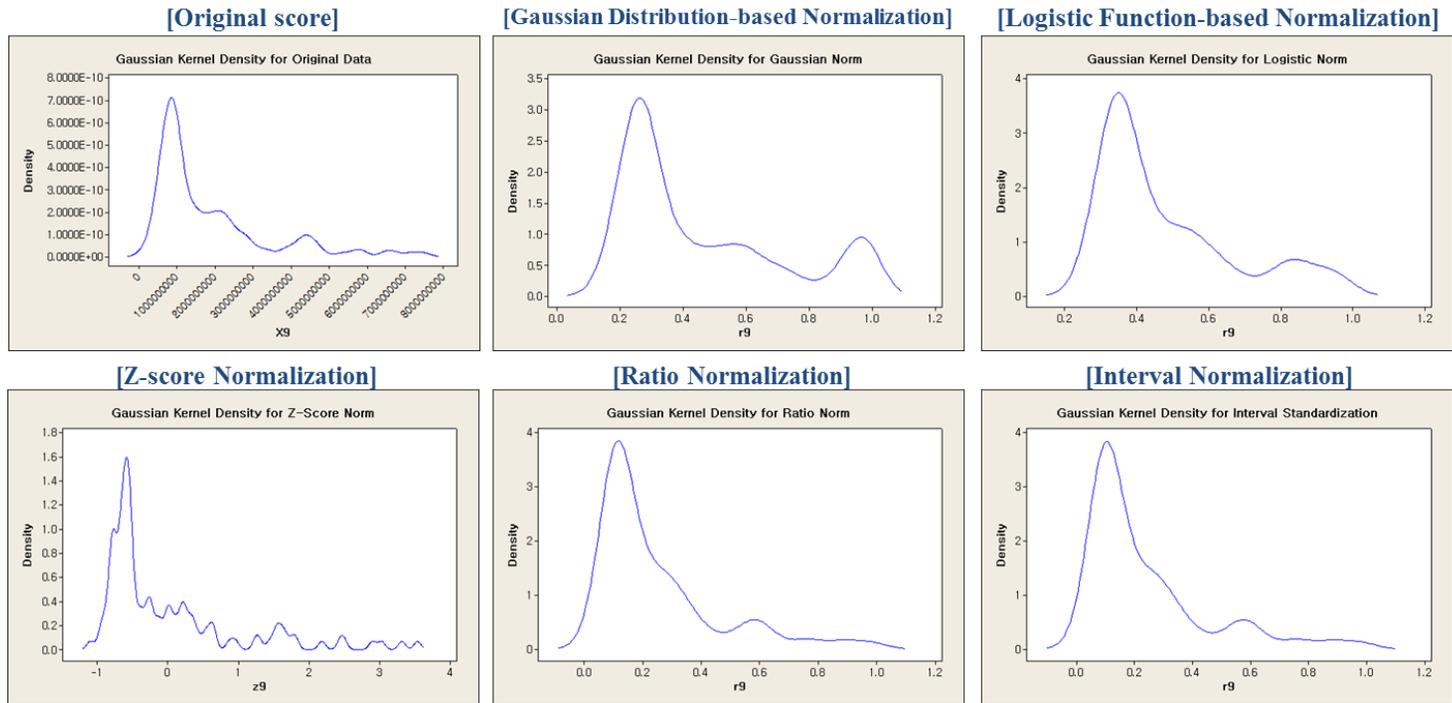


Figure 6-8. Kernel Density Estimation for Normalized Case-Bases (MB)

MARE, MSD, MAD, and SD for AW Method (MB)

For attribute weight measurement, in Table 6-22, the MAER shows that the results for all methods are relatively unstable. Among the four methods of attribute weigh assignment, the GA has the lowest values. On the other hand, the FC method has the highest values for k=1, 2, 5, and 10. Given that it has the lowest values, the GA method is the best for attribute weight assignment for the barrack. The entropy method, on the other hand, is the most inappropriate because it yields the highest values among the four.

Table 6-22. MAER for Attribute Weighting Methods (MB)

MAER	k=1	k=2	k=5	k=10
AI	0.195	0.183	0.198	0.206
Entropy	0.238	0.234	0.248	0.284
FC	0.243	0.238	0.260	0.317
GA	0.150	0.147	0.152	0.178

Table 6-23 shows the values of the weight assignment methods under the MSD. The AI and GA methods show stability in the values obtained for k=1, 2, 5 and 10 while there is instability in the entropy and FC methods. The AI values range from 0.003 to 0.004 while the GA values vary from 0.002 to 0.003. Given that the GA has the lowest values, it is therefore the most appropriate method of assigning weight under MSD because it can guarantee the highest accuracy.

Table 6-23. MSD for Attribute Weighting Methods (MB)

MSD	k=1	k=2	k=5	k=10
AI	0.004	0.003	0.003	0.004
Entropy	0.006	0.005	0.008	0.010
FC	0.008	0.007	0.010	0.014
GA	0.003	0.002	0.002	0.003

In figure 6-24, the values of weight assignment methods under MAD are established. All the methods show stability and there is no method that yields negative values. The FC method of weight assignment has the largest values while the GA values are the lowest. For k=1, 2, 5 and 10. The methods can be used for attribute weight assignment because of their low error rate. Given that that GA has the lowest values, it yields the lowest variability in weight assignment hence it is the best method of weight assignment under the MAD.

Table 6-24. MAD for Attribute Weighting Methods (MB)

MAD	k=1	k=2	k=5	k=10
AI	0.040	0.032	0.035	0.039
Entropy	0.046	0.043	0.051	0.059
FC	0.050	0.044	0.054	0.069
GA	0.031	0.027	0.028	0.031

The SD values shown in figure 6-25 shows that the values obtained among various methods for k=1, 2, 5 and 10. All the methods show instability in the

values obtained for various k-values. The FC method yields the highest values while the GA method yields the lowest ones. There is no method that gives negative values among the four. Therefore, having the lowest values, the GA method is the best method for assigning weight under the SD.

Table 6-25. SD for Attribute Weighting Methods (MB)

SD	k=1	k=2	k=5	k=10
AI	0.066	0.056	0.055	0.055
Entropy	0.079	0.071	0.088	0.088
FC	0.088	0.081	0.096	0.096
GA	0.054	0.043	0.045	0.045

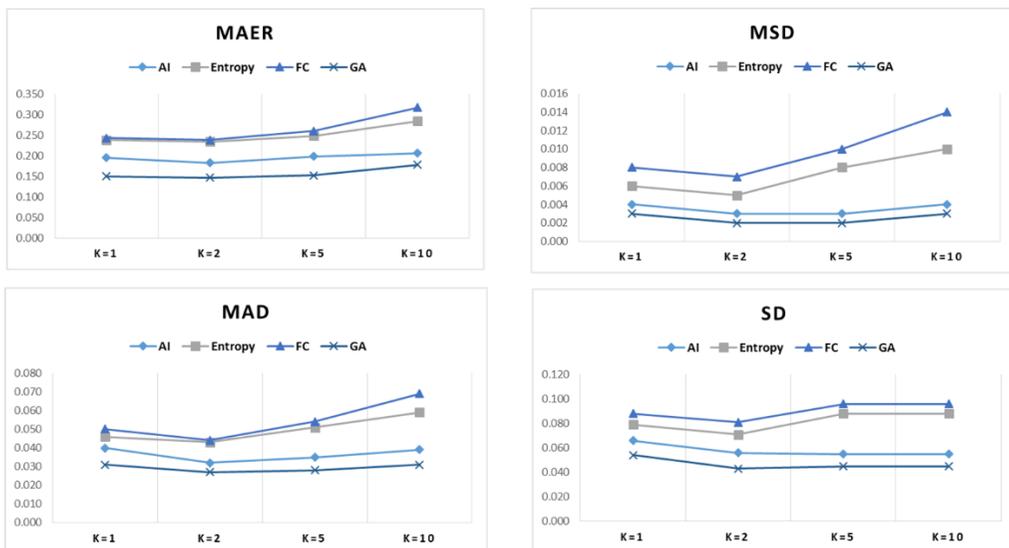


Figure 6-9. MAER, MSD, MAD, and SD for AW Methods (MB)

MARE, MSD, MAD, and SD for SM Method (MB)

Table 6-26 shows the similarity measurement values obtained for various k-values under the MAER. None of the methods yielded stable or negative values for k=1, 2, 5 and 10. There was no correlation between the increases of k. The Mahalanobis method gave the highest values while the Euclidian method yielded the lowest ones among the four methods. Therefore, having the lowest values, the Euclidean method is the best for similarity measurement in the MAER approach for the barrack.

Table 6-26. MAER for Similarity Measurement Methods (MB)

MAER	k=1	k=2	k=5	k=10
Mahalanobis	0.268	0.287	0.289	0.340
Euclidean	0.150	0.147	0.152	0.178
Arithmetic	0.222	0.203	0.237	0.259
Fractional	0.223	0.203	0.223	0.239

The MSD of similarity measurement methods are shown in Table 6-27. The results are relatively stable and all the values are low signifying that their errors are low. The Mahalanobis method yields the largest values among the four methods while the Euclidian method has the lowest for k=1, 2, 5 and 10. Therefore, having the lowest values and being the most stable among the four, the most appropriate method for similarity measurement under MSD is the Euclidean method.

Table 6-27. MSD for Similarity Measurement Methods (MB)

MSD	k=1	k=2	k=5	k=10
Mahalanobis	0.007	0.010	0.012	0.016
Euclidean	0.003	0.002	0.002	0.003
Arithmetic	0.004	0.005	0.007	0.010
Fractional	0.009	0.005	0.007	0.009

Table 6-28 shows the relative results they yielded with various k-values under the MAD in similarity measurement. None of the methods shows stability or a correlation between the k-value and the MAD values. Also, there is no method that gives negative values. The Mahalanobis-based CBR model gave the highest values while the Euclidean method gave the lowest ones for k=1, 2, and 10. Having the lowest values, the best method for similarity measurement under the MAD technique is the Euclidean method.

Table 6-28. MAD for Similarity Measurement Methods (MB)

MAD	k=1	k=2	k=5	k=10
Mahalanobis	0.053	0.056	0.060	0.073
Euclidean	0.031	0.027	0.028	0.031
Arithmetic	0.041	0.037	0.047	0.056
Fractional	0.046	0.038	0.045	0.051

The SD values for similarity measurements are shown in Table 6-29. None of the methods has stable results or negative ones for k=1, 2, 5, and 10. Among the four methods, the Mahalanobis has the highest values while the Euclidean

method has the lowest ones. Therefore, with the lowest values, the Euclidean method is the most appropriate one for similarity measurement under the SD.

Table 6-29. SD for Similarity Measurement Methods (MB)

SD	k=1	k=2	k=5	k=10
Mahalanobis	0.085	0.095	0.105	0.105
Euclidean	0.054	0.043	0.045	0.045
Arithmetic	0.067	0.069	0.082	0.082
Fractional	0.093	0.070	0.080	0.080

The Euclidean-based CBR model has the lowest and most stable values under the MAER, MSD, MAD, and SD. It is therefore the most appropriate method for similarity measurement for the military barrack project.

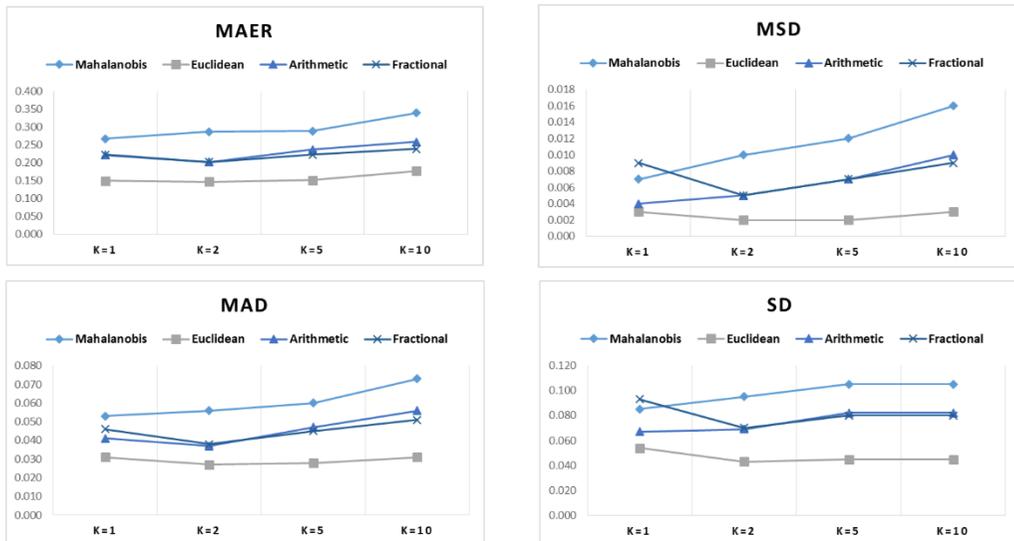


Figure 6-10. MAER, MSD, MAD, and SD for SM Methods (MB)

6.4 Government Office (GO)

6.4.1 Case-Base Profile

52 cases of government office projects² were utilized for validation. Extracted seven attributes are as follows: 1) gross floor area, 2) number of underground floor, 3) number of ground floor, 4) structure type (reinforced concrete), 5) structure type (steel reinforced concrete), 6) external material (metal), and 7) external material (stone).

Table 6-30. Attributes for Government Office

No.	Attribute Name	Scale Type
X1	Gross Floor Area	Ratio Scale
X2	No. of Underground Floor	Ratio Scale
X3	No. of Ground Floor	Ratio Scale
X4	Structure Type (reinforced concrete)	Nominal Scale
X5	Structure Type (steel reinforced concrete)	Nominal Scale
X6	External Material (Metal)	Nominal Scale
X7	External Material (Stone)	Nominal Scale

² Refer Koo et al. (2010b) for relating information about government office

Attribute weights obtained by AI, entropy, FC, and GA for government office building are summarized in Table 6-31.

Table 6-31. Attribute Weights by AI, Entropy, FC, and GA (GO)

No.	AI	Entropy	FC
X1	0.246	0.137	0.143
X2	0.003	0.162	0.143
X3	0.183	0.169	0.143
X4	0.234	0.129	0.143
X5	0.289	0.104	0.143
X6	0.040	0.154	0.143
X7	0.004	0.144	0.143

The matrix plot from the case-base of government office building to describe relationships of various attributes and their costs are analyzed in Figure 6-11. The matrix indicates variation in the relationship between the attributes and the cost. In some attributes, the relationship is positive while in others it is negative. Ostensibly, some do not exhibit any relationship.

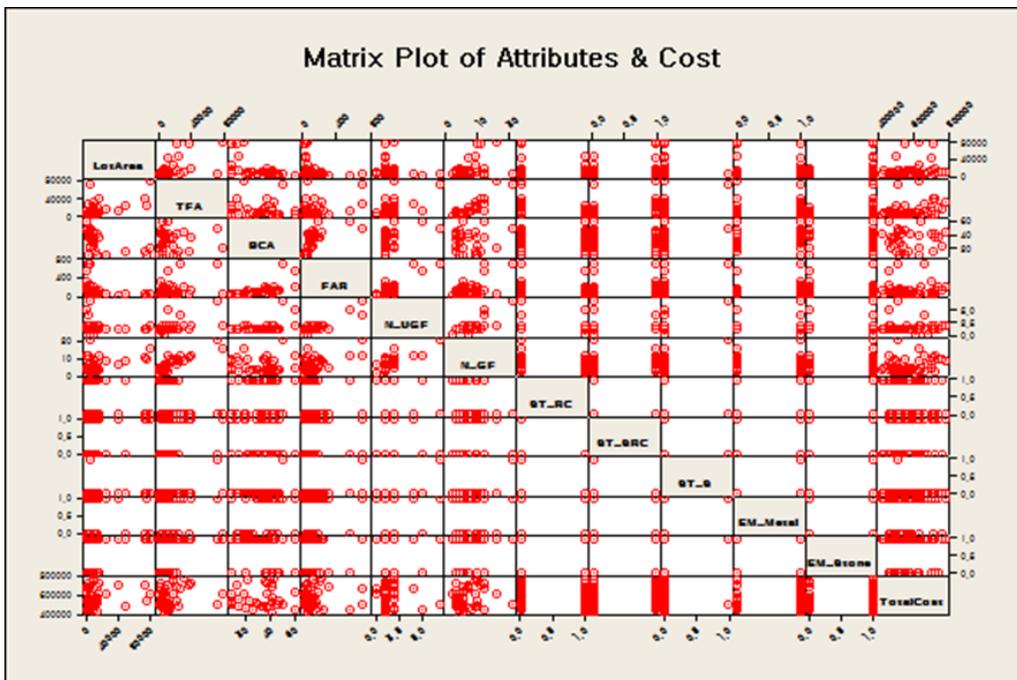


Figure 6-11. Matrix Plot (GO)

6.4.2 Results and Discussions

MARE, MSD, MAD, and SD for Norm. Method (GO)

Table 6-32 shows the results of MAER values obtained by various normalization methods. It can be observed that the interval results are correlated with the values of k in the all the methods. As the value of k increases, the values obtained in the results decrease. There is no stability in the methods. The ratio method that yields the lowest results for k=1, 2, 5 and 10 is the most appropriate because it yields results with low error rate. From the MAER table, The Gaussian method has the highest values while the Ratio scale has the lowest. Therefore, the ratio-based CBR model is the most accurate to use for government office projects.

Table 6-32. MAER for Normalization Methods (GO)

MAER	k=1	k=2	k=5	k=10
Interval	1.360	1.048	0.949	0.908
Gaussian	1.379	1.122	0.993	1.030
Logistic	0.623	0.510	0.436	0.444
Ratio	0.187	0.151	0.142	0.135

In table 6-33, the results for normalization under MSD are presented. It can be observed that there is a negative correlation between the values of k with the values obtained from the test. There is no normalization method that gives a negative result in the MSD. Among the four methods, the Gaussian method

yields the highest results while the ratio method gives the lowest ones. Therefore, when constructing a CBR model for government office building, the most appropriate normalization method under the MSD is the Ratio method because the results obtained using the method has the lowest errors.

Table 6-33. MSD for Normalization Methods (GO)

MSD	k=1	k=2	k=5	k=10
Interval	0.140	0.082	0.077	0.064
Gaussian	0.202	0.126	0.091	0.088
Logistic	0.102	0.060	0.045	0.044
Ratio	0.030	0.018	0.017	0.014

Table 6-34 compares the results of MAD from four normalization methods. It can be observed that the values obtained with all the methods are fairly stable. There is no clarity in the correlation between k values and the results obtained in the Gaussian, logistic and ratio methods. Also, none of the methods gives negative results. Among the four methods, the Gaussian gives the highest values while the ratio method gives the lowest for k=1, 2, 5 and 10. Therefore, the most appropriate method to use when for a government office building is the ratio-based CBR model because the model obtained has the lowest error rate.

Table 6-34. MAD for Normalization Methods (GO)

MAD	k=1	k=2	k=5	k=10
Interval	0.294	0.233	0.223	0.210
Gaussian	0.355	0.290	0.249	0.255
Logistic	0.253	0.195	0.173	0.178
Ratio	0.134	0.110	0.009	0.012

In table 6-35, the results for SD from four different normalization methods are shown. The Gaussian, logistic and ratio-based CBR models show a negative correlation between the values of k and the results obtained while the results for interval method do not correlate. Among the four methods, the Gaussian method yields the highest results while the ratio method yields the lowest. When the results obtained are high, it implies that the model obtained by using the method has high error rate and vice versa. It can therefore be established that the ratio-based CBR model is the most accurate for government office projects under MSD performance measure because of the low error rate.

Table 6-35 SD for Normalization Methods (GO)

SD	k=1	k=2	k=5	k=10
Interval	0.378	0.290	0.190	0.230
Gaussian	0.454	0.359	0.305	0.299
Logistic	0.322	0.247	0.213	0.212
Ratio	0.174	0.137	0.130	0.120

When estimating construction cost for government office buildings, the best normalization method to employ is the ratio method because it has the lowest error rate in MAER, MAD, MSD and MAD. On the other hand, using the Gaussian method can lead to high errors because it yields the highest values in the four approaches for $k=1, 2, 5$ and 10 .

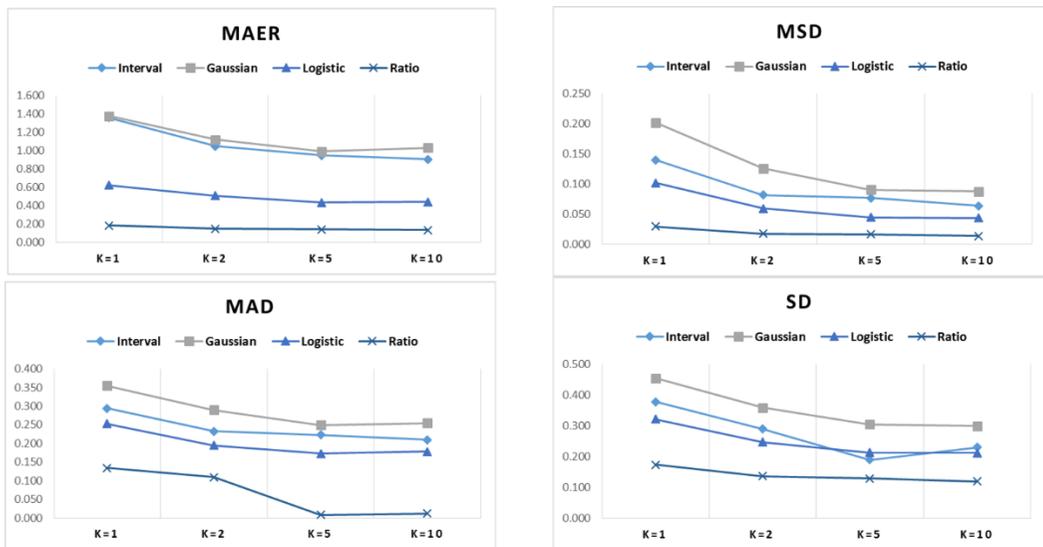


Figure 6-12. MAER, MSD, MAD, and SD for Norm. Methods (GO)

Kernel Density Estimation (GO)

The basic reason for normalization procedures is to limit discontinuities and produce smoother curves that can provide an accurate reflection of the data trends. Based on figure three, the original data indicates existence of four crests. The normalization of this data through Gaussian normalization procedure is ineffective. This could be a result of wrong bandwidth (Ziegler 2006). The

result in Gaussian normalization is underestimation leading to a curve with four crests concentrated in the middle zone. Similar trend is exhibited when logistic function is used for normalization. Figure 6-13 indicates that ratio normalization is the best strategy. It neither leads to underestimation or overestimation of the crests. The curve produced through ratio normalization is smooth and the trend exhibited aligns to the trends in the original score (Zhang and Wang 2009). In the absence of ratio normalization, the z-score can be used. The z-score normalization curve closely resembles the original score curve but is relatively smoother. The interval procedure is ineffective since it does not smooth the crests. It shows that the bandwidth is too small.

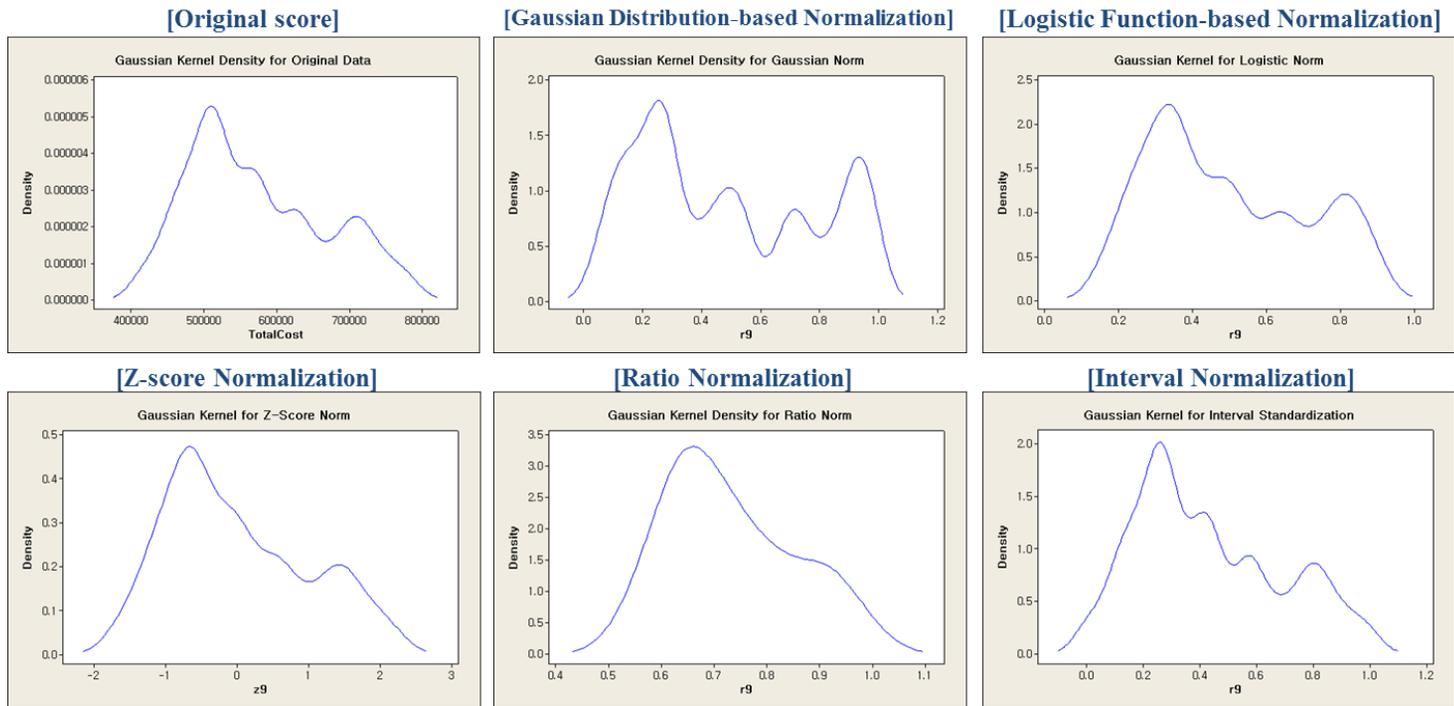


Figure 6-13. Kernel Density Estimation for Normalized Case-Bases (GO)

MARE, MSD, MAD, and SD for AW Method (GO)

Table 6-36 shows the results obtained for attribute weight assignment for various methods under MAER can be observed that stability is higher in the FC and GA methods than the AI method. However, the two stable methods (FC and GA) have relatively higher values than the AI method in the weight assignment for k=1, 2, 5 and 10. The three methods are appropriate for weight assignment because the values they yield are almost equivalent. However, the best one to use is the AI because it can give a model whose results are the most accurate among the three.

Table 6-36. MAER for Attribute Weighting Methods (GO)

MAER	k=1	k=2	k=5	k=10
AI	0.187	0.151	0.142	0.135
FC	0.196	0.156	0.141	0.141
GA	0.196	0.155	0.140	0.143

In Table 6-37, the estimation MSD values obtained from the GA, AI and FC methods are shown. It can be established the FC and GA values are equivalent and that all the methods yield low and stable values for k=1, 2, 5 and 10 hence are appropriate for weight assignment. The values of AI are lower than the ones obtained in the FC and GA. Therefore, an AI-based CBR model for government office has lower errors than the one constructed using either FC or GA.

Table 6-37. MSD for Attribute Weighting Methods (GO)

MSD	k=1	k=2	k=5	k=10
AI	0.030	0.018	0.017	0.014
FC	0.031	0.020	0.016	0.016
GA	0.031	0.020	0.016	0.016

Table 6-38 shows the values obtained for attribute weight assignment under MAD. From the table, one can establish all the methods yield results that are relatively stable. It can also be noted that that the FC and GA values are almost equivalent while those of the AI show a slight variation for k=1, 2, 5 and 10. A slightly negative correlation exists in the FC and GA but the correlation diminishes with the increase in the value of K in the GA model. Among the three methods, The FC has the highest results while AI has the lowest. Errors in estimation decrease with a decrease in the values for all k. Therefore, having the lowest results, the AI model is the most appropriate to apply when constructing for government office under MAD (Nielsen 2007).

Table 6-38. MAD for Attribute Weighting Methods (GO)

MAD	k=1	k=2	k=5	k=10
AI	0.134	0.110	0.104	0.099
FC	0.139	0.115	0.105	0.103
GA	0.139	0.114	0.104	0.104

In table 6-39, results are shown reflecting the estimation values obtained when attribute weight assignment is done using various methods under SD. From the table one can established that the results are stable in all the methods. The results for FC and GA are almost equivalent too. The AI results which are different from the other two methods are the lowest while the FC values are the highest for k=1, 2, 5 and 10. A method constructed using the method with the lowest SD values has the highest accuracy. Therefore, AI is the most appropriate method of attribute weight assignment for government office projects among the three methods under SD.

Table 6-39. SD for Attribute Weighting Methods (GO)

SD	k=1	k=2	k=5	k=10
AI	0.174	0.137	0.130	0.120
FC	0.176	0.141	0.127	0.126
GA	0.176	0.140	0.127	0.127

In the attribute weight assignment, it can be determined that all the three methods are appropriate. Their yield is almost equivalent for k=1, 2, 5 and 10. However, it can be established that the AI yields the lowest errors while the FC yields the highest. Therefore, the assignment of attribute weight in the government housing can be done best using the AI model.

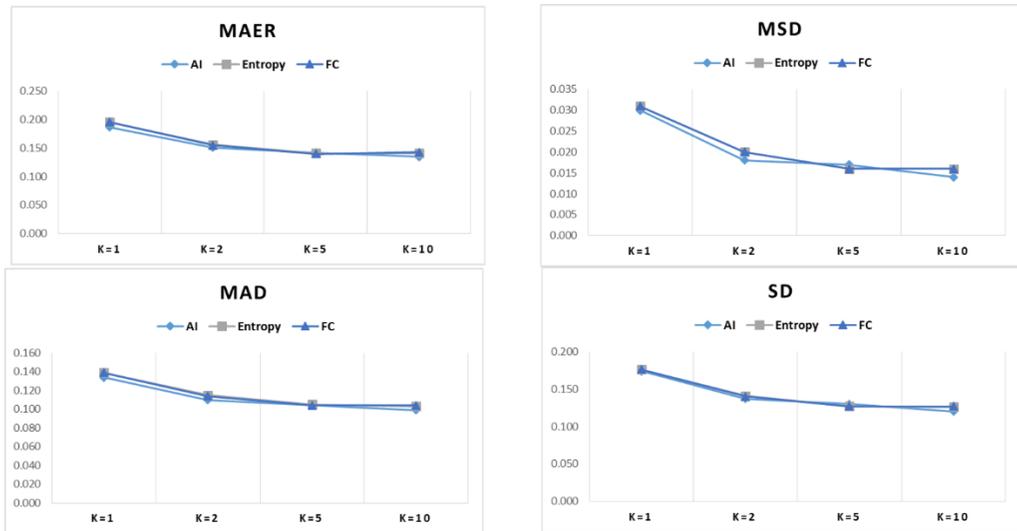


Figure 6-14. MAER, MSD, MAD, and SD for AW Methods (GO)

MARE, MSD, MAD, and SD for SM Method (GO)

Table 6-40 shows the MAER results obtained when measuring similarity using various methods. It can be established that the three methods have stable results for $k=1, 2, 5$ and 10 . However, the Arithmetic method has higher results than all the three while the Fractional method has the lowest. An MAER that has high results has high rate of errors than the one with the low results (Wang et al. 2012). Therefore, from the table results, the most appropriate method of similarity measurement under MAER is therefore the fractional method.

Table 6-40. MAER for Similarity Measurement Methods (GO)

MAER	k=1	k=2	k=5	k=10
Euclidean	0.187	0.151	0.142	0.135
Arithmetic	0.190	0.148	0.140	0.135
Fractional	0.184	0.150	0.130	0.131

In table 6-41, the values obtained from different methods under MSD for similarity measurement are shown. It can be established that all the methods give results that are stable and positive. The Euclidean method yields the highest results while the lowest values are obtained using the fractional method for k=1, 2, 5 and 10. Large values imply high error rate and vice versa. Therefore, in the creation of a CBR model for government office under MSD can be done best using the fractional method because the results obtained have the lowest error rate.

Table 6-41. MSD for Similarity Measurement Methods (GO)

MSD	k=1	k=2	k=5	k=10
Euclidean	0.030	0.018	0.017	0.014
Arithmetic	0.031	0.018	0.015	0.014
Fractional	0.030	0.018	0.014	0.014

Table 6-42 shows the various values obtained under MAD using various methods. From the table, it can be observed that all the results are positive and

stale. The Arithmetic model gives the highest values while the fractional method has the lowest for $k=1, 2, 5$ and 10 . Low values in the MAD are an indication of low error rate in the model. Therefore, the best method to use for similarity measurement when creating a CBR model for government office under MAD approach is the fractional model because it has the lowest error rate.

Table 6-42. MAD for Similarity Measurement Methods (GO)

MAD	k=1	k=2	k=5	k=10
Euclidean	0.134	0.110	0.104	0.099
Arithmetic	0.138	0.106	0.100	0.099
Fractional	0.133	0.107	0.094	0.097

In table 6-43, the SD results for similarity measurement under different methods are shown. It can be established that the different methods have stability. The highest values are obtained using the Arithmetic method while the lowest ones are obtained using the Fractional method. For $k=1, 2, 5$ and 10 . When a method yields high values, it implies that the estimations obtained have high errors and vice versa (Elsas and Florysiak 2015). Therefore, the best similarity measurement method under SD for government office is the Fractional method.

Table 6-43. SD for Similarity Measurement Methods (GO)

SD	k=1	k=2	k=5	k=10
Euclidean	0.174	0.137	0.130	0.120
Arithmetic	0.178	0.136	0.122	0.119
Fractional	0.175	0.134	0.118	0.118

The three methods of measuring similarities are stable. They also tend to have a similar correlation with the changes in the values for k. However, the arithmetic method tends to yield the highest values while the fractional method yields the lowest under the MAER, MAD, MSD, and SD. Therefore, when measuring similarities for government office projects, the best method to employ is the fractional method because its estimations have the lowest errors.

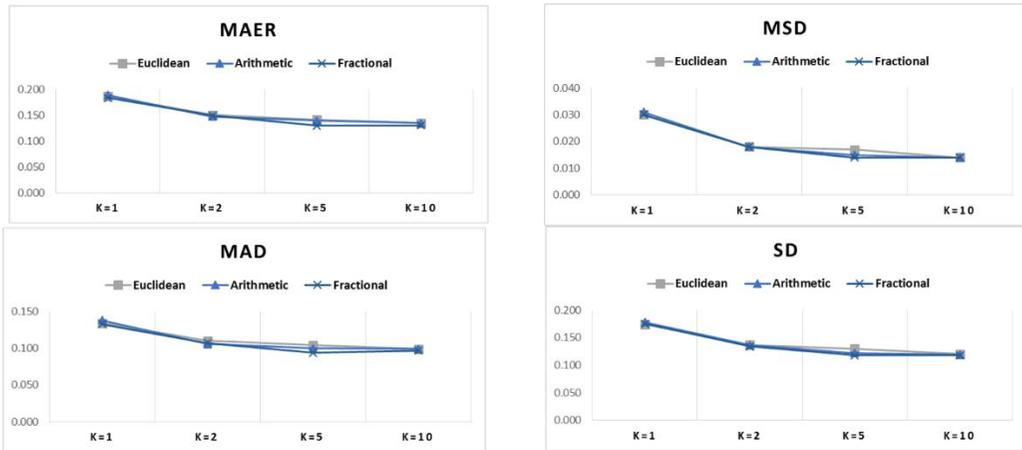


Figure 6-15. MAER, MSD, MAD, and SD for SM Methods (GO)

6.5 Summary

This chapter validated the suggested front-end cost estimation methodology by selective CBR for multi-family housing, military barrack, and government office projects. For each project type, MEAR, MSD, MAD for the estimate accuracy, SD for estimate stability, KDE for appropriateness of selecting normalization method was examined.

The mechanism of the CBR model for selecting the most accurate and stable normalization methods from sub-module 1, attribute weighting method from sub-module 2, and similarity measurement method from sub-module 3 responding to different types of building projects (or different characteristics of case-bases) was validated and summarized in Table 6-44~52.

Case Study 1: Multi-Family Housing Projects

To summarize, ratio normalization method, GA attribute weighting method, and arithmetic summation similarity measurement method-based CBR cost model was proposed to be the most accurate and stable for multi-family housing projects.

Specifically, the most accurate normalization method in terms of MAER, MSD, and MAD was ratio method whereas the least accuracy was derived from Z-score. In terms of SD, the most stable accuracy result was yielded from ratio method while the least stable result was obtained from Z-score. In terms of

appropriateness of normalization method to be used, both ratio, interval, and Z-score normalization methods were analyzed to be appropriate whereas Gaussian and logistic methods were evaluated to be inappropriate. Therefore, in overall, ratio normalization method-based CBR cost model for multi-family housing achieved the most accurate and stable cost estimate results and is to be the most appropriate method.

Table 6-44. Summary of Norm. Method Selection (MFH)

Norm. Method	MAER	MSD	MAD	SD	KDE
Most Accurate	Ratio	Ratio	Ratio		
Least Accurate	Z-score	Z-score	Z-score		
Most Stable				Ratio	
Least Stable				Z-score	
Appropriate					Ratio/ Interval/ Z-score
Inappropriate					Gaussian/ Logistic/

Regarding an attribute weighting method in terms of MAER, MSD, MAD, and SD, the most accurate and stable results were derived from GA whereas the least accuracy was resulted from FC method for the multi-family housing case-base. Hence, GA attribute weighting method-based CBR cost model is considered to be the most effective in achieving high level of cost estimate accuracy.

Table 6-45. Summary of AW Method Selection (MFH)

AW Method	MAER	MSD	MAD	SD
Most Accurate	GA	GA	GA	
Least Accurate	FC	FC/ Entropy	FC	
Most Stable				GA
Least Stable				FC

Concerning the similarity measurement methods, the most accurate and stable cost estimate results were obtained by arithmetic summation method in MAER, MSD, MAD, and SD for multi-family housing projects whereas the least accuracy and stability were achieved by Mahalanobis method.

Table 6-46. Summary of SM Method Selection (MFH)

SM Method	MAER	MSD	MAD	SD
Most Accurate	Arithmetic	Euclidean/ Arithmetic/ Fractional	Arithmetic	
Least Accurate	Mahalanobis	Mahalanobis	Euclidean/ Mahalanobis	
Most Stable				Arithmetic/
Least Stable				Euclidean/ Mahalanobis

Case Study 2: Military Barrack Projects

In short, interval/ratio normalization method, GA attribute weighting method, and Euclidean similarity measurement method-based CBR cost model was derived to be the most accurate and stable for military barrack projects.

Table 6-47. Summary of Norm. Method Selection (MB)

Norm. Method	MAER	MSD	MAD	SD	KDE
Most Accurate	Logistic/ Z-score	Interval/ Ratio	Ratio		
Least Accurate	ratio	Z-score	Z-score		
Most Stable				Interval/ Ratio	
Least Stable				Z-score	
Appropriate					Interval/ Ratio
Inappropriate					Gaussian/ Logistic/ Z-score

Concretely, in terms of MSD, MAD, and SD, the highest level of accuracy and stability was achieved using interval/ratio normalization methods whereas the least accuracy and stability was obtained by Z-score. Under MEAR, logistic and Z-score methods achieved the most accurate results whereas ratio method yielded the least accuracy. From the kernel density estimation as means of distinguishing appropriate normalization method selection, both interval and

ratio methods were extracted to be appropriate whereas Gaussian, logistic, and Z-score methods were analyzed to be inappropriate methods. To sum up, interval/ratio normalization method-based CBR cost model for military barrack attained the most accurate and stable cost estimate results; and these methods are also considered to be appropriate to use.

Concerning an attribute weighting method under the performance measures of MAER, MSD, MAD, and SD, the most accurate and stable cost estimate results were attained from GA whereas the least accuracy and stability were resulted from FC method. Hence, GA attribute weighting method-based CBR cost model is effective in establishing the most accurate and stable cost estimate results for the military barrack projects.

Table 6-48. Summary of AW Method Selection (MB)

AW Method	MAER	MSD	MAD	SD
Most Accurate	GA	GA	GA	
Least Accurate	FC	FC	FC	
Most Stable				GA
Least Stable				FC

In terms of the similarity measurement methods, the most accurate and stable cost estimation results were obtained by Euclidean method under MAER, MSD, MAD, and SD for military barrack projects whereas the least accuracy and stability were obtained by Mahalanobis method. Hence, the Euclidean

similarity measurement method-based CBR cost model is suggested to be the most effective approach.

Table 6-49. Summary of SM Method Selection (MB)

SM Method	MAER	MSD	MAD	SD
Most Accurate	Euclidean	Euclidean	Euclidean	
Least Accurate	Mahalanobis	Mahalanobis	Mahalanobis	
Most Stable				Euclidean
Least Stable				Mahalanobis

Case Study 3: Government Office Projects

In overall, ratio normalization method, AI attribute weighting method, and fractional function similarity measurement method-based CBR cost model was validated to be the most accurate and stable for government office projects.

Specifically, the most accurate cost estimation results under MAER, MSD, and MAD, and stable results under SD were achieved from ratio normalization method whereas the least accuracy and stability was attained by Gaussian method. In terms of appropriateness of normalization methods, ratio and Z-score methods were analyzed to be appropriate whereas Gaussian, logistic, and interval were inappropriate to be selected as effective normalization methods. Based on the analysis, ratio normalization method-based CBR cost model for government office projects achieved the most accurate and stable cost estimate results which satisfied appropriateness as normalization method as well.

Table 6-50. Summary of Norm. Method Selection (GO)

Norm. Method	MAER	MSD	MAD	SD	KDE
Most Accurate	Ratio	Ratio	Ratio		
Least Accurate	Gaussian	Gaussian	Gaussian		
Most Stable				Ratio	
Least Stable				Gaussian	
Appropriate					Ratio/ Z-score
Inappropriate					Gaussian/ Logistic/ Interval

In terms of attribute weighting method, cost estimation results from AI method was considered to be the most accurate and stable under MAER, MSD, MAD, and SD whereas the least accuracy and stability were obtained by FC method. Thus, AI attribute weighting method-based CBR cost model is considered to be the most accurate and stable for government office projects.

Table 6-51. Summary of AW Method Selection (GO)

AW Method	MAER	MSD	MAD	SD
Most Accurate	AI	AI	AI	
Least Accurate	FC	FC	FC	
Most Stable				AI
Least Stable				FC

For government office projects, fractional function-based similarity measurement method yielded the most accurate and stable estimate results under MAER, MSD, MAD, and SD whereas the least accuracy and stability were attained by the arithmetic summation method. Therefore, the fractional function similarity measurement method needs to be applied to CBR cost model to achieve the high level of cost estimate for government office projects.

Table 6-52. Summary of SM Method Selection (GO)

SM Method	MAER	MSD	MAD	SD
Most Accurate	Fractional	Fractional	Fractional	
Least Accurate	Arithmetic	Euclidean	Arithmetic	
Most Stable				Fractional
Least Stable				Arithmetic

Chapter 7. Conclusions

The success of every construction project has significant dependence on the high level of cost estimate accuracy in the front-end stage. However, the current construction industry has encountered problems of inaccurate budgeting, limited information availability, limited usage of unit price of actual construction cost, and lack of flexibility of a cost model for various building projects.

The CBR, which utilizes and adjusts past cases to solve given problems, is an effective approach to be applied for accurate front-end cost estimation; and it is especially supportive of public owners' decision-makings in budgeting and cost planning. Importantly, since CBR relies on the past historical data, it is significant to perform effective case-base development to obtain high quality of case-bases. Furthermore, to improve the reliability of front-end cost estimation results, an explanatory power and accuracy of CBR-based cost model needs to be enhanced by adopting advanced statistical methods and techniques and designing flexible reasoning environments.

This chapter summarizes the research results, clarifies research contributions to the body of knowledge, and clearly elaborates the limitations and future research.

7.1 Research Results

Improved Flexibility and Accuracy of Front-End Cost Estimation using Selective CBR Model for Different Types of Building Projects

This research proposed the front-end cost estimation methodology by selective CBR in dealing with various building types or different characteristics of dataset more flexibly and accurately. The *Method Selection Module* comprised the three sub-modules that are *Sub-Module 1: Normalization Method Selection Module* (interval, Gaussian, Z-score, logistic, and ratio), *Sub-Module 2: Attribute Weighting Method Selection Module* (Attribute Impact, entropy, feature counting, and genetic algorithms), and *Sub-Module 3: Similarity Measurement Method Selection Module* (Mahalanobis, Euclidean, arithmetic summation, fractional function).

The results of case studies for the validation of the proposed methodology are summarized as below: For the multi-family housing project, ratio normalization method, GA attribute weighting method, and arithmetic summation similarity measurement method-based CBR cost model was proposed to be the most accurate and stable. For the military barrack project, interval/ratio normalization method, GA attribute weighting method, and Euclidean similarity measurement method-based CBR cost model was suggested to be the most accurate and stable. For the government office project, ratio normalization method, AI attribute weighting method, and fractional function similarity measurement method-based CBR cost model was derived to be the most accurate and stable.

Effective Case-Base Establishment for Acquiring Qualified Datasets

This research suggested the *Module 1: Case-Base Development* for the effective case-base development to obtain improved quality of datasets. The procedures were data acquisition, design of case-base structure, attribute extraction, data storing, analysis of matrix plot, data cleaning, data integration, data transformation, and data reduction. The importance of preprocessed or treated case-bases was emphasized to secure the validity and reliability of the CBR cost model. Furthermore, the *Module 1* is expected to improve the reliability of public owners/cost estimators' trust and transparency of estimation process.

CBR Model Design Experiment for Examining and Verifying Model Components and Validation Procedures

Issues in normalization, attribute weight assignment, and similarity measurement were discussed. Accuracy and stability of cost estimation were tested according to different normalization, attribute weighting, and similarity measurement methods. In terms of normalization methods, this research verified the two hypothesis (in Chapter 4.1): 1) accuracy (MAER, MSD, and MAD) and stability (SD) of CBR cost estimation can be enhanced by applying statistically advanced normalization methods. 2) A CBR cost model has appropriate normalization methods to be applied.

In terms of attribute weighting methods (in Chapter 4.2), the use of the proposed AI which considered the ranges of the attributes in measuring the weights of attributes quantitatively was validated, and it yielded reliable

estimate accuracy and stability in CBR and parametric cost estimations. This experiment contributes the body of knowledge where accurately assigning the weights of attributes is required, and it remains a challenging issue especially in machine learning areas such as attribute weight assignment in CBR and weights of connection strength of neural network method.

In terms of similarity measurement method (in Chapter 4.3), both cost estimation results of the theoretical examination based on simulation data conditions and the applicability test by the case study supported that Euclidean distance, arithmetic summation, and fractional function can yield relatively high level of cost estimate accuracy in retrieving similar cases whereas the Mahalanobis distance based similarity measurement which considered the influence of covariance among attributes achieved an overall higher MAERs and standard deviations. Furthermore, it is important to note that lower MAERs of the Mahalanobis distance method were resulted when the simulation data based experiment was executed compared to MAERs from the case studies using multi-family housings. This was mainly because limitations existed where a large number of attributes were used to compute variance-covariance matrix. The twelve attributes were used for the case study of multi-family housings whereas only three or five attributes were used for simulation test. Therefore, they contained highly correlated information and yielded less accurate results.

7.2 Research Contributions

With the proposed cost estimation methodology by selective CBR, realistic and accurate cost estimation in the conceptual stage can be made based on past historical data for multi-family housing, military barrack, and government office building projects. This can support decision-makings of public institutions where high level of accuracy is required for budgeting. By establishing stable budgeting, this can ultimately reduce change of orders, increase of cost, and delay of projects.

With minimal information in the front-end stage, the suggested CBR model can satisfy public owners with cost estimation results with accuracy, reliability of human trust, and transparency of cost estimation process. Thus public owners and cost estimators can obtain highly effective strategies by obtaining accurate cost estimation results with attribute information of historical data of similar projects. Based on preprocessed reliable data, cost estimator can persuade project owners for the initiation of construction projects. Furthermore, as the suggested CBR model is reasoning on accumulated data for various building projects, dependency on experienced experts for estimating accurate cost can be reduced; and less participation of participants for cost evaluation can be made.

This research can also compensate the limitations of unit price for actual construction cost and standard of construction estimate by providing accurate

cost estimate results of the most similar past cases in the conceptual stage. The utilization of the standards containing unit price of construction materials, labors, equipment, and etc. are focused on detailed design or documentation stage. Without detailed design information, these standards have very limited usages.

However, the proposed case-based reasoning cost estimation model can provide accurate cost estimate results using basic and only several information. Moreover, updates of unit prices of the standards require much effort and times; and calculating process of unit prices is blackbox as historical data used are unrevealed. Unlike this, only with several attribute information, new cases can be stored in CBR model; and cost estimation can be made with transparency of estimate process knowing which cases are selected for computations.

This research ultimately contributes improved accuracy of CBR model by selecting the most accurate normalization, attribute weighting, and similarity measurement methods according to different building construction projects. The proposed cost estimation methodology using selective CBR can reflect different characteristics of database. This research improved flexibility of cost estimation model compared to the existing fixed model.

7.3 Limitations and Future Research

1) The proposed cost estimation methodology was validated by performing case studies of multi-family housing, military barrack, and government public office projects. To examine the flexibility of the model, other types of construction projects such as plants, roads, airports need to be tested.

2) Since cost data of construction projects are confidential and thus very difficult to obtain from public institutions or construction companies, continuous storing of new cases could not be achieved. CBR methodology can be highly effective especially when various range of new cases are kept updated; otherwise, CBR may not reflect and reason for new problems/projects. Therefore, how to extend and retain the cases of CBR database need to be considered.

3) A sensitivity analysis in terms of how many attributes should be used and their accuracy result comparisons need to be examined. Once case-bases are constructed, it is very difficult to add new attributes and their information.

4) To improve quality of case-bases, more logical case-base construction process needs to be developed. Furthermore, advanced data preprocessing techniques need to be applied.

5) Apart from interval, Gaussian, Z-score, logistic, and ratio methods, other normalization methods need to be examined to improve cost estimate accuracy.

6) Other than AI, entropy, FC, and GA methods, various attribute weighting methods should be test to enhance cost estimate accuracy.

7) Regarding similarity measurement methods, other methods except Mahalanobis, Euclidean, arithmetic summation, and fractional function need to be tested for improving cost estimation.

8) A flexible cost estimation model dealing with various types of construction project using combinational methods such as CBR with neural networks or others needs to be further examined.

9) Instead of correlation coefficients or coefficient of determinations as the weights of the attributes to calculate values of AI, we need to use other weight-assigning methods such as standardized regression coefficients or genetic algorithm to verify the robustness of AI. More importantly, the AI concept itself needs further examination and development.

Bibliography

AACE International, Skills & Knowledge of Cost Engineering, 4th ed, AACE, Morgantown, WV, USA, 1999.

R. Adhikari, A neural network based linear ensemble framework for time series forecasting, *Neurocomputing* 157 (2015) 231-242.

A. Aamodt, E. Plaza, Case-based reasoning: foundational issues, methodological variations and system approaches, *AI Communications* 7 (1) (1994) 39-59.

D.W. Aha, L.A. Breslow, H. Muñoz-Avila, Conversational case-based reasoning, *Applied Intelligence* 14 (1) (2001) 9-32.

H. Ahn, K.J. Kim, I. Han, Global optimization of feature weights and the number of neighbors that combine in a case-based reasoning system, *Expert System* 23 (5) (2006) 290-301.

J. Ahn, S.H. Ji, M. Park, H.S. Lee, S. Kim, S.W. Suh, The attribute impact concept: applications in case-based reasoning and parametric cost estimation, *Automation in Construction* 43 (2014) 195-203.

S.J. Ahn, *Statistical Decision Theory*, Freeacademy, 2007.

K.D. Althoff, R. Bergmann, M. Minor, A. Hanft, *Advances in case-based reasoning: 9th European Conference, ECCBR 2008, Trier, Germany, September 1-4, 2008: proceedings*, Springer, 2008.

S.H. An, G.H. Kim, K.I. Kang, A case-based reasoning cost-estimating model using experience by analytic hierarchy process, *Building and Environment* 42 (2007) 2573–2579.

S. Anderson, K. Molenaar, C. Schexnayder, *Guidance for cost estimation and management for highway projects during planning, programming, and preconstruction*, Washington, D.C: Transportation Research Board, 2007.

D. Arditi, O. Tokdemir, Comparison of case-based reasoning and artificial neural networks, *Journal of Computing in Civil Engineering* 13 (3) (1999) 162-169.

A. Ashworth, *Building Economics and Cost Control: Worked Solutions*, 1st ed, Butterworth, London, 1983.

K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is “nearest neighbor” meaningful?, *Database Theory-ICDT’99* (1999) 217-235.

M. Bill, P.M. Joseph, J.P. Joseph, *Cost Estimating*, 2010.

K. Black, *Business Statistics: For Contemporary Decision Making*, John Wiley & Sons, 2014.

E. Blocher, D. Stout, P. Juras, G. Cokins, *Cost Management: A Strategic Emphasis*, 6th ed., McGraw-Hill Education, 2012.

R. Bod, J. Hay, S. Jannedy, *Probabilistic Linguistics*, MIT Press, Cambridge, Massachusetts, 2003.

J. Bode, Neural networks for cost estimation: simulations and pilot application, *International Journal of Production Research* 38 (6) (2000) 25–30.

- S.A. Book, Prediction bounds for general-error-regression cost-estimating relationships, *Journal of Cost Analysis and Parametrics* 5 (1) (2012) 25-51.
- Z.I. Botev, J.F. Grotowski, D.P. Kroese, Kernel density estimation via diffusion, *Annals of Statistics* 38 (5) (2010) 2916-2957.
- J. Van den Broeck, S.A. Cunningham, R. Eeckels, K. Herbst, Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Medicine*, 2 (10) (2005) 966-970.
- M.F. Bryant, A comparison of the rule and case-based reasoning approaches for the automation of help-desk operations at the tier-two level, Ph.D. Dissertation, Nova Southeastern University, 2009.
- H.D. Burkhard, Similarity and distance in case-based reasoning, *Fundamenta Informaticae* 47 (3-4) (2001) 201-215.
- S.R. Carroll, D.J. Carroll, *Statistics Made Simple for School Leaders: Data-Driven Decision Making*. R&L Education, New York, 2002.
- G.C. Cawley, N.L.C. Talbot, Fast exact leave-one-out cross-validation of sparse least-squares support vector machines, *Neural Networks* 17 (2004) 1467-1475.
- M.Y. Cheng, H.C. Tsai, W.S. Hsieh, Web-based conceptual cost estimates for construction projects using evolutionary fuzzy neural inference model, *Automation in Construction* 18 (2) (2009) 164–172.
- M. Chiarandini, D. Basso, T. Stützle, Statistical methods for the comparison of stochastic optimizers, *The Sixth Metaheuristics International Conference* (2005) 189-196.

J.S. Chou, Web-based CBR system applied to early cost budgeting for pavement maintenance project, *Expert System with Applications* 39 (2009) 2947-2960.

W.S. Cleveland, *Visualizing data*. Hobart Press, 1993.

W.S. Cleveland, M.E. McGill, *Dynamic Graphics for Statistics*, Chapman and Hall/CRC, 1988.

S. Coscoy, E. Huguet, F. Amblard, Statistical analysis of sets of random walks: how to resolve their generating mechanism, *Bulletin of mathematical biology* 69 (8) (2007) 2467-2492.

R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, The mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems* 50 (1) (2000) 1-18.

S.J. Delany, S. Ontañón, Case-based reasoning research and development: 21st International Conference, ICCBR 2013, Saratoga Springs, NY, USA, July 8-11, 2013. *Proceedings*, Springer Berlin Heidelberg, 2013.

E. Deza, M.M. Deza, *Encyclopedia of Distances*, Springer, 2009.

F.X. Diebold, R.S. Mariano, Comparing predictive accuracy, *Journal of Business & economic statistics*, 20 (1) (2012) 134-144.

S.Z. Doğan, D. Arditi, H.M. Günadın, Determining attribute weights in a CBR model for early cost prediction of structural system, *Journal of Construction Engineering and Management* 132 (10) (2006) 1092-1098.

- J. Du, J. Bormann, Improved similarity measure in case-based reasoning with global sensitivity analysis: an example of construction quantity estimating, *Journal of Computing in Civil Engineering* 28 (6) (2012) 04014020.
- T. Duong, ks: Kernel density estimation and kernel discriminant analysis for multivariate data in R, *Journal of Statistical Software*, 21 (7) (2007).
- P. Duverlie, J.M. Castelain, Cost estimation during design step: parametric method versus case-based reasoning method, *International Journal of Advanced Manufacturing Technology* 15 (12) (1999) 895–906.
- A.S.C. Ehrenberg, Data Reduction, *Journal of Empirical Generalisations in Marketing Science* 5 (1) (2000) 1-391.
- K.R. Ellsworth, Cost-to-capacity analysis for estimating waste-to-energy facility costs, *Cost Engineering* 40 (6) (1998) 27–30.
- R. Elsas, D. Florysiak, Dynamic capital structure adjustment and the impact of fractional dependent variables, *Journal of Financial and Quantitative Analysis*, 50 (5) (2015) 1105-1133.
- Esteem Software, Esteem 1.4: Case Based Reasoning Development Tool, San Mateo, Calif, 1996.
- B.S. Everitt, S. Landau, M. Leese, D. Stahl, *Miscellaneous Clustering Methods in Cluster Analysis*, 5th Edition, John Wiley & Sons, Ltd, Chichester, UK, 2011.
- F. Famili, W.M. Shen, R. Weber, E. Simoudis, Data pre-processing and intelligent data analysis, *Intelligent Data Analysis* 1 (1) (1997) 3-23.

D.E. Farrar, R.R. Glauber, Multicollinearity in regression analysis: the problem revisited, *Review of Economics and Statistics* 49 (1967) 92-107.

R. Fellows, A. Liu, *Research Methods for Construction*, 2nd ed, Blackwell Science, 2003.

GAO, *GAO Cost Estimating and Assessment Guide: Best Practices for Developing and Managing Capital Program Costs*, GAO-09-3SP, Washington, D.C., 2004.

S. García, J. Luengo, F. Herrera, *Data preprocessing in data mining*, Springer international publishing, 2015.

H.G. Gauch, J.T.G. Hwang, G.W. Fick, Model evaluation by comparison of model-based predictions and measured values, *Agronomy Journal* 95 (6) (2003) 1442-1446.

A. M. Gerrard, *Guide to Capital Cost Estimating*, 4th ed., IChemE, 2000

N.A. Gershenfeld, *The Nature of Mathematical Modeling*, Cambridge University Press, Cambridge, UK, 1999.

C. Globig, K.P. Jantke, S. Lange, Y. Sakakibara, On case-based learnability of languages, *New Generation Computing* 15 (1) (1997) 59-83.

D.E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley, 2006.

A.R. Golding, A review of case-based reasoning, *American Association for Artificial Intelligence*, 16 (2) (1995) 85-86.

- G. Góra, A. Wojna, A new classification system combining rule induction and instance-based learning, *Fundamenta Informaticae* 51 (4) (2002) 369-390.
- P. Hall, B.U. Park, R.J. Samworth, Choice of neighbor order in nearest-neighbor classification, *Annals of Statistics* 36 (5) (2008) 2135-2152.
- J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kauffman, San Francisco, 2003.
- J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kauffman, San Francisco, 2006.
- J. Havil, *Gamma: Exploring Euler's Constant*, Princeton University Press, Princeton, NJ, 2003.
- T. Hegazy, A. Ayed, Neural network model for parametric cost estimating of highway projects, *Journal of Construction Engineering and Management* 124 (3) (1998) 210–218.
- ISPA, *Parametric Estimating Handbook*, 4th edition, Vienna, VA, 2008.
- A. Jain, K. Nandakumar, A. Ross, Score normalization in multimodal biometric systems, *Pattern recognition* 38 (12) (2005) 2270-2285.
- A. Jarde, S. Alkass, Computer-integrated system for estimating the costs of building project, *Journal of Construction Engineering and Management* 13 (4) (2007) 205–223.
- J. Jarmulak, S. Craw, R. Rowe, Genetic algorithms to optimise CBR retrieval, *Advances in Case-Based Reasoning*, 1898 (2000) 136-147.

S.H. Ji, M. Park, H.S. Lee, Data preprocessing-based parametric cost model for building projects: with case studies of Korean construction projects, *Journal of Construction Engineering and Management* 136 (8) (2010) 844–853.

S.H. Ji, M. Park, H.S. Lee, J. Ahn, N. Kim, B. Son, Military facility cost estimation system (MilFaCE) using case-based reasoning in Korea, *Journal of Computing in Civil Engineering* 25 (3) (2011a) 218-231.

S.H. Ji, M. Park, H.S. Lee, Cost estimation model for building projects using case-based reasoning, *Canadian Journal of Civil Engineering* 38 (5) (2011b) 570-581.

S.H. Ji, M. Park, H.S. Lee, Case adaptation method of case-based reasoning for construction cost estimation in Korea, *Journal of Construction Engineering and Management* 138 (1) (2012) 43-52.

R. Jin, Y. Breitbart, C. Muoh, Data discretization unification, *Knowledge and Information Systems* 19 (1) (2009) 1-29.

R.Z. Jin, K.M. Cho, C.T. Hyun, M.J. Son, MRA-based revised CBR model for cost prediction in the early stage of construction projects, *Expert Systems with Applications* 39 (2012) 5214-5222.

R.A. Johnson, D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1998.

S. Karshenas, J. Tse, A case-based reasoning approach to construction cost estimating, *Information Technology 2002, Computing in Civil Engineering* (2002) 113–123.

R. Kesavan, C. Elanchezian, B.V. Ramnath, *Process Planning and Cost Estimation*, New Age International Pvt Ltd Publishers, 2008.

B.S. Kim, T.H. Hong, Revised case-based reasoning model development based on multiple regression analysis for railroad bridge construction, *Journal of Construction Engineering and Management* 138 (1) (2012) 154-162.

G. Kim, K. Kang, A study on predicting construction cost of apartment housing projects based on case based reasoning technique at the early project stage, *Journal of Architectural Institute of Korea* 20 (5) (2004) 83-92.

H.J. Kim, Y.C. Seo, C.T. Hyun, A hybrid conceptual cost estimating model for large building projects, *Automation in Construction* 25 (9) (2012) 72-81.

K. Kim, K. Kim, Preliminary cost estimation model using case-based reasoning and genetic algorithms, *Journal of Computing in Civil Engineering* 24 (6) (2010) 499-505.

M.H. Kim, *Cost Planning in Architecture*, Kimoondang, Seoul, 2005.

S. Kim, Hybrid forecasting system based on case-based reasoning and analytic hierarchy process for cost estimation, *Journal of Civil Engineering and Management* 19 (1) (2013) 86-96.

R. Kirkham, Ferry and Brandon's *Cost Planning of Buildings*, 9th Ed., Wiley-Blackwell, 2014.

J.L. Kolodner, C.E. Hmelo, N.H. Narayanan, Problem-based learning meets case-based reasoning, *ICLS '96 Proceedings of the 1996 international conference on Learning sciences*, (1996) 188-195.

C.W. Koo, T.H. Hong, C.T. Hyun, The development of a construction cost prediction model with improved prediction capacity using the advanced CBR approach, *Expert Systems with Applications* 38 (7) (2011) 8597-8606.

C.W. Koo, T.H. Hong, C.T. Hyun, K.J. Koo, A CBR-based hybrid model for predicting a construction duration and cost based on project characteristics in multi-family housing projects, *Canadian Journal of Civil Engineering* 37 (2010a) 739-752.

C.W. Koo , T.H. Hong , C.T. Hyun , S.H. Park, J. Seo, A study on the development of a cost model based on the owner's decision making at the early stages of a construction project, *International Journal of Strategic Property Management*, 14 (2) (2010b) 121-137.

Korea Institute of Construction Technology, *Construction Cost Index*, 2010.

S.B. Kotsiantis, D. Kanellopoulos, P.E. Pintelas, Data preprocessing for supervised learning, *International Journal of Computer Science* 1 (2) (2006) 111-117.

E. Kreyszig, *Advanced Engineering Mathematics*, 4th ed., Wiley, 1979.

D. Leake, *Case-Based Reasoning: Experience, Lessons, and Future Directions*, AAAI Press/MIT Press, Menlo Park, NJ, 1996.

G.H. Lee, Rule-based and case-based reasoning approach for internal audit of bank, *Knowledge-Based Systems*, 21 (2) (2008) 140-147.

H. Liu, H. Metoda, *Instance Selection and Constructive Data Mining*, Kluwer, Boston, MA, 2001.

P.C. Mahalanobis, On the generalized distance in statistics, *Proceedings of the National Institute of Sciences (Calcutta)* (1936) 49-55.

S.L. Mansar, F. Marir, H.A. Reijers, Case-based reasoning as a technique for knowledge management in business process redesign, *Electronic Journal on Knowledge Management*, 1 (2) (2003) 113-124.

R. Mantaras et al., Retrieval, reuse, revision and retention in case-based reasoning, *The Knowledge Engineering Review*, 20 (3) (2005) 215-240.

F. Marir, M. Watson, Case-based reasoning: A review, *The Knowledge Engineering Review* 9 (4) (1994) 327-354.

C. Marling, M. Sqalli, E. Rissland, H. Muñoz-Avila, D. Aha, Case-based reasoning integrations, *AI magazine*, 23 (1) (2002) .

J.F. McCarthy, E.J. McCarthy, *Construction Project Management: A Managerial Approach*, Pareto, 2011.

J.H. McDonald, *Handbook of Biological Statistics*, Baltimore, MD: Sparky House Publishing, 2009.

G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Interscience, 1992.

B. Melnyk, D. Morrison-Beedy, *Intervention Research: Designing, Conducting, Analyzing, and Funding*, Springer Publishing Company, 2012.

R.E. Meyer, J.T. Burns, Facility parametric cost estimating, *Proceedings of the Trans American Association of Cost Engineers*, 43rd Annual Meeting, Denver, EST.02.1-EST.02.6 (1999).

T.M. Mitchell, *Machine Learning*, Mcgraw Hill, 1997.

S. Montani, L.C. Jain, *Successful Case-Based Reasoning Applications: 2*, Springer Berlin Heidelberg, 2013.

H.R. Nemati, C.D. Barko, A. Moosa, E-CRM analytics: The Role of Data Integration, Business Intelligence in the Digital Economy: Opportunities, Limitations, and Risks, Idea group Inc., 2004.

A. Neumaier, Vienna Proposal for Interval Standardization, Universit Wien Nordbergstr, Wien, 2008.

A.A. Nielsen, The regularized iteratively reweighted MAD method for change detection in multi-and hyperspectral data, Image Processing, IEEE Transactions on, 16 (2) (2007) 463-478.

J.W. Osborne, Notes on the use of data transformations, Practical Assessment, Research and Evaluation 8 (6) (2002).

B. Ozorhon, L. Dikmen, M. Birgonul, Case-based reasoning model for international market selection, Journal of Construction Engineering and Management 132 (9) (2006) 940-948.

S.K. Pal, S.C.K. Shiu, Foundations of Soft Case-Based Reasoning, Wiley Interscience, Hoboken, NJ, 2004.

J.K. Patel, C.B. Read, Handbook of the Normal Distribution. 2nd eds., CRC Press, New York, 1996.

J. Peltier, D. Zahay, A.S. Krishen, A hierarchical IMC data integration and measurement framework and its impact on CRM system quality and customer performance, Journal of Marketing Analytics 1 (1) (2013) 32-48.

P. Perner, Case-based reasoning and the statistical challenges, Quality and Reliability Engineering International, 24 (6) (2008) 705-720.

M. Pica, Project Life Cycle Economics: Cost Estimation, Management and Effectiveness in Construction Projects, Gower Publishing, 2015.

A. Poluektov, Kernel density estimation of a multidimensional efficiency profile, Journal of Instrumentation, 10 (2) (2015) 2011.

K. Potts, N. Ankrah, Construction Cost Management: Learning from Case Studies, 2nd ed., Routledge, 2014.

J. Prentzas, I. Hatzilygeroudis, Categorizing approaches combining rule-based and case-based reasoning, Expert Systems, 24 (2) (2007) 97-122.

D. Pyle, Data Preparation for Data Mining, Morgan Kaufmann Publishers, Los Altos, CA, 1999.

Z. Qian, W.S. Gao, F. Wang, Z. Yan, A case-based approach to power transformer fault diagnosis using dissolved gas analysis data, European Transactions on Electrical Power 19 (3) (2009) 518-530.

E. Rahm, H.H. Do, Data Cleaning: Problems and current approaches, IEEE Data Eng. Bull 23 (4) (2000) 3-13.

I. Ragnemalm, the Euclidean distance transform in arbitrary dimensions, Pattern Recognition Letters 14 (11) (1993) 883-888.

A. Ram, N. Wiratunga, Case-based reasoning research and development, ICCBR 2011, Springer-Verlag Berlin Heidelberg , 2011.

M.S. Ramabodu, Procurement guidelines for project success in cost planning of construction projects, Ph.D. Dissertation, University of the Free State, 2014.

M.M. Richter, R.O. Weber, *Case-Based Reasoning: A Textbook*, Springer Berlin Heidelberg, 2013.

C.K. Riesbeck, R.C. Schank, *Inside Case-Based Reasoning*, Lawrence Erlbaum Associates, Hillsdale, NJ, 1989.

H. Robinson, B. Symonds, B. Gilbertson, B. Ilozor, *Design Economics for the Built Environment: Impact of Sustainability on Project Evaluation*, Wiley-Blackwell, 2015.

D.C. Rubin, A basic-systems approach to autobiographical memory, *American Psychological Society*, 14 (2) (2005) 79-83.

H.G. Ryu, H.S. Lee, M. Park, Construction planning method using case-based reasoning (CONPLA-CBR), *Journal of Computing in Civil Engineering*, 21 (6) (2007) 410-422.

R.C. Schank, *Dynamic Memory Revisited*, Cambridge University Press, 1999.

R.C. Schank, A. Kass, C.K. Riesbeck, *Inside Case-Based Explanation*, ed., Lawrence Erlbaum Associates, Hillsdale, NJ, 1994.

A. Schirmer, Case-Based reasoning and improved adaptive search for project scheduling, *Naval Research Logistics* 47 (3) (2000) 201-222.

S.D. Schuette, R.W. Liska, *Building Construction Estimating*, McGraw-Hill, New York, 1994.

K.H. Seong, M. Park, H.S. Lee, S.H. Ji, Cost estimating in early stage using parametric method for apartment construction projects, 2008 Annual Conference, Korea Institute of Construction Engineering and Management (2008) 219–223.

- A. R. Serway, C. Vuille , College Physics, 9th edition, Brooks/Cole, Boston, 2012.
- Z.D. Sevgi, D. Arditi, G. Murat, Using decision trees for determining attribute weights in a case-based model of early cost prediction, *Journal of Construction Engineering and Management* 134 (2) (2008) 146-152.
- L.A. Shalabi, Z. Shaaban, B. Kasasbeh, Data mining: A preprocessing engine, *Journal of Computer Science* 2 (9) (2006) 735-739.
- S.J. Sheather, M.C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, *Journal of the Royal Statistical Society* 53 (3) (1991) 683-690.
- K.S. Shin, I. Han, Case-based reasoning supported by genetic algorithms for corporate bond rating, *Expert Systems with Applications* 16 (2) (1999) 85-95.
- A.E. Smith, A.K. Mason, Cost estimation predictive modeling: regression versus neural network, *Engineering Economist* 42 (2) (1997) 137-161.
- J. Smith, D. Jaggar, *Building cost planning for the design team*, Butterworth-Heinemann, 2007.
- N.J. Smith, *Project Cost Estimating*, Thomas Telford, London, 1995.
- B. Smyth, M.T. Keane, Adaptation-guided retrieval: questioning the similarity assumption in reasoning, *Artificial Intelligence* 102 (2) (1998) 249-293.
- B.S. Son, H.S. Lee, M.S. Park, D.Y. Han, J. Ahn, Quantity based active schematic estimating (Q-BASE) model, *KSCE Journal of Civil Engineering* 17 (1) (2013) 9-21.

R. Sonmez, Parametric range estimating of building costs using regression models and bootstrap, *Journal of Construction Engineering and Management* 134 (12) (2008) 1011–1016.

R.W. Soukoreff, I.S. MacKenzie, Metrics for text entry research: an evaluation of MSD and KSPC, and a new unified error metric, *CHI '03 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2003) 113-120.

M. Soutos, D.J. Lowe, Procost—towards a powerful early stage cost estimating tool, *International Conference on Computing in Civil Engineering*, July 12-15 in Mexico, ASCE (2005) 1-12.

R.D. Stewart, R.M. Wyskida, J.D. Johannes, *Cost Estimator's Reference Manual*, Wiley-Interscience, 1995.

J.B. Stiff, P.A. Mongeau, *Persuasive Communication*, Guilford Press, New York, 2002.

H.W. Stoll, *Product Design Methods and Practices*, CRC Press, 1999.

D. Towey, *Cost Management of Construction Projects*, Wiley-Blackwell, 2013.

S.M. Trost, G.D. Oberlender, Predicting accuracy of early cost estimates using factor analysis and multivariate regression, *Journal of Construction Engineering and Management*, 129 (2) (2003) 198-203.

E. Tulving, *Organization of Memory*, Academic Press, New York and London, 1972.

M.P. Wand, M.C. Jones, *Kernel Smoothing*, Chapman & Hall/CRC, London, 1995.

- D. Wang, D.S. Yeung, E.C.C. Tsang, Weighted Mahalanobis distance kernels for support vector machines, *IEEE Transactions on Neural Networks* 18 (5) (2007) 1453-1462.
- X.V. Wang, N. Blades, J. Ding, R. Sultana, G. Parmigiani, Estimation of sequencing error rates in short reads, *BMC bioinformatics*, 13 (1) (2012).
- I. Watson, *Applying Case-Based Reasoning: Techniques for Enterprise System*, Morgan Kaufmann Publishers, 1997.
- B.S. Weir, C.C. Cockerham, Estimating F-statistics for the analysis of population structure, *Society for the Study of Evolution*, 38 (6) (1984) 1358-1370.
- G.M. Winch, *Managing Construction Projects*, 2nd Edition, Wiley-Blackwell, 2009.
- N.J. Yau, J.B. Yang, Case-based reasoning in construction management, *Computer Aided Civil and Infrastructure Engineering* 13 (1998) 143–150.
- K. Younis, M. Karim, R. Hardie, J. Loomis, S. Rogers, M. DeSimio, Cluster merging based on weighted Mahalanobis distance with application in digital mammograph, *Aerospace and Electronics Conference, Proceedings of the IEEE 1998 National* (1998).
- J. Zhang, Generalized least squares cross-validation in kernel density estimation. *Statistica Neerlandica*, *Statistica Neerlandica*, 69 (3) (2015) 315-328.
- J. Zhang, X. Wang, Robust normal reference bandwidth for kernel density estimation. *Statistica Neerlandica*, 63 (1) (2009) 13-23.
- K. Ziegler, On local bootstrap bandwidth choice in kernel density estimation, *Statistics & Decisions*, 24 (2) (2006) 11.

Appendix

Appendix: Source Code of Macros

Appendix: Source Code of Macros

Macro # for calculating the AI weight column

AiW x.1-x.m; # input data, including standardized cost variable

weight wt. # output weights

mcolumn x.1-x.m r1.1-r1.m r2 wt

mmatrix r # correlation matrix

name wt 'w_ai'

corr x.1-x.m r

copy r r1.1-r1.m

copy r1.m r2;

exclude;

rows m.

let wt = r2*r2

let wt = wt/sum(wt)

Endmacro

Macro # for calculating the Entropic weight column

EntropyW x.1-x.m; # input data, excluding the cost variable

weight wt. # output weights

mcolum x.1-x.m p.1-p.m wt

mconstant w.1-w.m e.1-e.m u.1-u.m n d.1-d.m d_sum j i

name wt 'w_entropy'

let n = count(x.1)

do j = 1:m # column number

let p.j = x.j/sum(x.j)

enddo

do j = 1:m # column number

let e.j = 0

do i=1:n # row number

if p.j(i) = 0

let e.j = e.j + 0

elseif p.j(i) ~= 0

let e.j = e.j - p.j(i)*loge(p.j(i))

endif

enddo # i

let u.j = e.j/loge(m)

let d.j = 1 - u.j

enddo # j

let d_sum = 0

do j = 1:m

let d_sum = d_sum + d.j

```
enddo
```

```
do j = 1:m
```

```
    let w.j = d.j/d_sum
```

```
enddo
```

```
do j = 1:m
```

```
    let wt(j) = w.j
```

```
enddo
```

```
Endmacro
```

Macro # Gaussian distribution Normalization

```
gaussnrm x.1-x.m; # input
```

```
    save y.1-y.n. # output
```

```
mcolumn x.1-x.m y.1-y.n z.1-z.n
```

```
mconstant j
```

```
center x.1-x.m z.1-z.n # Standardization
```

```
do j=1:n
```

```
    cdf z.j y.j
```

```
enddo
```

```
Endmacro
```

Macro # Interval standardization or Grey relational generating (GRG)

```
intstd x.1-x.m; # input
  save y.1-y.n; # output
  direct bct; # "b": benefit, "c": cost, "t": target
  target t.

mcolumn x.1-x.m y.1-y.n
mconstant bct t j

default bct="b" t=0

if bct = "b"
  do j=1:m
    let y.j = (x.j - min(x.j))/(max(x.j) - min(x.j))
  enddo
elseif bct = "c"
  do j=1:m
    let y.j = (max(x.j) - x.j)/(max(x.j) - min(x.j))
  enddo
elseif bct = "t"
  do j=1:m
    let y.j = abs(x.j - t)/rmax((max(x.j) - t), (t - min(x.j)))
  enddo
endif

Endmacro
```

Macro # Logistic function Normalization

```
logisnrm x.1-x.m; # input
```

```
    save y.1-y.n. # output
```

```
mcolumn x.1-x.m y.1-y.n z.1-z.n
```

```
mconstant j
```

```
center x.1-x.m z.1-z.n # Standardization
```

```
do j=1:n
```

```
    let y,j = (1 + expo(-z,j))**(-1)
```

```
enddo
```

```
Endmacro
```

Macro # loocv for Arithmetic summation based similarity measure

```
loocv_an x.1-x.m c c1_hat d1 maer1 msd1 mad1 m_d1 sd_d1 &  
          c2_hat d2 maer2 msd2 mad2 m_d2 sd_d2 &  
          c3_hat d3 maer3 msd3 mad3 m_d3 sd_d3 &  
          c4_hat d4 maer4 msd4 mad4 m_d4 sd_d4; #
```

standardized or normed data

weight w;

knn knn1 knn2 knn3 knn4. # knn1, knn2, knn3 knn4: positive integers less than or equal to half the size of the case-base

```
Mcolumn x.1-x.m x1.1-x1.m w c xi.1-xi.m x0.1-x0.m simil &  
          c1 s_c1 c1_hat d1 knn1_c c2_hat d2 knn2_c c3_hat d3 knn3_c &  
          c4_hat d4 knn4_c# k_sim
```

```
Mconstant rowi row0 j max_x.1-max_x.m min_x.1-min_x.m n n1 i h sim &  
          knn1 maer1 msd1 mad1 m_d1 sd_d1 &  
          knn2 maer2 msd2 mad2 m_d2 sd_d2 &  
          knn3 maer3 msd3 mad3 m_d3 sd_d3 &  
          knn4 maer4 msd4 mad4 m_d4 sd_d4
```

```
Mmatrix x0_v xi_v
```

brief 0 # brief 1

```
name knn1 'knn1' c1_hat 'c1_hat' d1 'diff1' maer1 'mean of |d1|/c' msd1 'mean  
of d1^2' mad1 'mean of |d1|' &
```

```
          m_d1 'mean of d1' sd_d1 'stdev of d1'
```

```
name knn2 'knn2' c2_hat 'c2_hat' d2 'diff2' maer2 'mean of |d2|/c' msd2 'mean  
of d2^2' mad2 'mean of |d2|' &
```

```
          m_d2 'mean of d2' sd_d2 'stdev of d2'
```

```
name knn3 'knn3' c3_hat 'c3_hat' d3 'diff3' maer3 'mean of |d3|/c' msd3 'mean  
of d3^2' mad3 'mean of |d3|' &
```

```
          m_d3 'mean of d3' sd_d3 'stdev of d3'
```

```

name knn4 'knn4' c4_hat 'c4_hat' d4 'diff4' maer4 'mean of |d4|/c' msd4 'mean
of d4^2' mad4 'mean of |d4|' &
    m_d4 'mean of d4' sd_d4 'stdev of d4'
let n = count(x.1)

####

name simil 'sim_a'
note 'Arithmetic summation based similarity measure'

do i=1:n
    let row0 = i # putative new case
    copy x.1-x.m x0.1-x0.m;
    include;
    rows row0.
    copy x0.1-x0.m x0_v # {row} new vector

    copy x.1-x.m c x1.1-x1.m c1;
    exclude;
    rows row0.

let n1 = count(x1.1)
do h=1:n1
    let rowi = h # comparative case
    copy x1.1-x1.m xi.1-xi.m;
    include;
    rows rowi.
    copy xi.1-xi.m xi_v # {row} comparative vector

%simil_a1 x1.1-x1.m sim;
weight w;

```

```

        comv xi_v;
        newv x0_v.
    let simil(h) = sim
    erase sim
enddo # h

sort c1 s_c1;
    by simil;
    descending simil.

copy s_c1 knn1_c;
    include;
        rows 1:knn1. # rows n1-knn1+1:n1 for ascending order.
let c1_hat(i) = mean(knn1_c)

copy s_c1 knn2_c;
    include;
        rows 1:knn2. # rows n1-knn2+1:n1 for ascending order.
let c2_hat(i) = mean(knn2_c)

copy s_c1 knn3_c;
    include;
        rows 1:knn3. # rows n1-knn3+1:n1 for ascending order.
let c3_hat(i) = mean(knn3_c)

copy s_c1 knn4_c;
    include;
        rows 1:knn4. # rows n1-knn4+1:n1 for ascending order.
let c4_hat(i) = mean(knn4_c)

```

```
enddo # i
```

```
let d1 = c - c1_hat # the difference of cost and the (mean) retrieved cost(s)
let maer1 = mean(abs(d1)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd1 = mean(d1**2) # mean squared deviation(MSD)
let mad1 = mean(abs(d1)) # mean absolute absolute deviation(MAD)
let m_d1 = mean(d1)
let sd_d1 = stdev(d1)
let d1(n+2) = maer1
let d1(n+3) = msd1
let d1(n+4) = mad1
let d1(n+5) = m_d1
let d1(n+6) = sd_d1
print knn1 maer1 msd1 mad1 m_d1 sd_d1
```

```
let d2 = c - c2_hat # the difference of cost and the (mean) retrieved cost(s)
let maer2 = mean(abs(d2)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd2 = mean(d2**2) # mean squared deviation(MSD)
let mad2 = mean(abs(d2)) # mean absolute absolute deviation(MAD)
let m_d2 = mean(d2)
let sd_d2 = stdev(d2)
let d2(n+2) = maer2
let d2(n+3) = msd2
let d2(n+4) = mad2
let d2(n+5) = m_d2
let d2(n+6) = sd_d2
print knn2 maer2 msd2 mad2 m_d2 sd_d2
```

```

let d3 = c - c3_hat # the difference of cost and the (mean) retrieved cost(s)
let maer3 = mean(abs(d3)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd3 = mean(d3**2) # mean squared deviation(MSD)
let mad3 = mean(abs(d3)) # mean absolute absolute deviation(MAD)
let m_d3 = mean(d3)
let sd_d3 = stdev(d3)
let d3(n+2) = maer3
let d3(n+3) = msd3
let d3(n+4) = mad3
let d3(n+5) = m_d3
let d3(n+6) = sd_d3
print knn3 maer3 msd3 mad3 m_d3 sd_d3

```

```

let d4 = c - c4_hat # the difference of cost and the (mean) retrieved cost(s)
let maer4 = mean(abs(d4)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd4 = mean(d4**2) # mean squared deviation(MSD)
let mad4 = mean(abs(d4)) # mean absolute absolute deviation(MAD)
let m_d4 = mean(d4)
let sd_d4 = stdev(d4)
let d4(n+2) = maer4
let d4(n+3) = msd4
let d4(n+4) = mad4
let d4(n+5) = m_d4
let d4(n+6) = sd_d4
print knn4 maer4 msd4 mad4 m_d4 sd_d4

```

Endmacro

Macro # loocv for Euclidean distance based similarity measure

```
loocv_en x.1-x.m c c1_hat d1 maer1 msd1 mad1 m_d1 sd_d1 &  
          c2_hat d2 maer2 msd2 mad2 m_d2 sd_d2 &  
          c3_hat d3 maer3 msd3 mad3 m_d3 sd_d3 &  
          c4_hat d4 maer4 msd4 mad4 m_d4 sd_d4; #
```

standardized or normed data

weight w;

knn knn1 knn2 knn3 knn4. # knn1, knn2, knn3, knn4: positive integers less than or equal to half the size of the case-base

```
Mcolumn x.1-x.m x1.1-x1.m w c xi.1-xi.m x0.1-x0.m simil &  
          c1 s_c1 c1_hat d1 knn1_c c2_hat d2 knn2_c c3_hat d3 knn3_c &  
          c4_hat d4 knn4_c # k_sim
```

```
Mconstant rowi row0 j max_x.1-max_x.m min_x.1-min_x.m n n1 i h wdis_sq  
wdis sim &
```

```
knn1 maer1 msd1 mad1 m_d1 sd_d1 &  
knn2 maer2 msd2 mad2 m_d2 sd_d2 &  
knn3 maer3 msd3 mad3 m_d3 sd_d3 &  
knn4 maer4 msd4 mad4 m_d4 sd_d4
```

```
Mmatrix x0_v xi_v
```

brief 0 # brief 1

```
name knn1 'knn1' c1_hat 'c1_hat' d1 'diff1' maer1 'mean of |d1|/c' msd1 'mean  
of d1^2' mad1 'mean of |d1|' &
```

```
m_d1 'mean of d1' sd_d1 'stdev of d1'
```

```
name knn2 'knn2' c2_hat 'c2_hat' d2 'diff2' maer2 'mean of |d2|/c' msd2 'mean  
of d2^2' mad2 'mean of |d2|' &
```

```
m_d2 'mean of d2' sd_d2 'stdev of d2'
```

```
name knn3 'knn3' c3_hat 'c3_hat' d3 'diff3' maer3 'mean of |d3|/c' msd3 'mean  
of d3^2' mad3 'mean of |d3|' &
```

```

    m_d3 'mean of d3' sd_d3 'stdev of d3'
name knn4 'knn4' c4_hat 'c4_hat' d4 'diff4' maer4 'mean of |d4|/c' msd4 'mean
of d4^2' mad4 'mean of |d4|' &
    m_d4 'mean of d4' sd_d4 'stdev of d4'

```

```

let n = count(x.1)

```

```

####

```

```

    name simil 'sim_e'
    note 'Euclidean distance based similarity measure'

```

```

do i=1:n

```

```

    let row0 = i # putative new case
    copy x.1-x.m x0.1-x0.m;
    include;
        rows row0.
    copy x0.1-x0.m x0_v # {row} new vector

```

```

    copy x.1-x.m c x1.1-x1.m c1;
    exclude;
        rows row0.

```

```

let n1 = count(x1.1)

```

```

do h=1:n1

```

```

    let rowi = h # comparative case
    copy x1.1-x1.m xi.1-xi.m;
    include;
        rows rowi.
    copy xi.1-xi.m xi_v # {row} comparative vector

```

```

%simil_e1  x1.1-x1.m wdis_sq wdis sim;
    weight w;
    comv xi_v;
    newv x0_v.
let simil(h) = sim
erase wdis_sq wdis sim
enddo # h

sort c1 s_c1;
    by simil;
    descending simil.

copy s_c1 knn1_c;
    include;
        rows 1:knn1. # rows n1-knn1+1:n1 for ascending order.
let c1_hat(i) = mean(knn1_c)

copy s_c1 knn2_c;
    include;
        rows 1:knn2. # rows n1-knn2+1:n1 for ascending order.
let c2_hat(i) = mean(knn2_c)

copy s_c1 knn3_c;
    include;
        rows 1:knn3. # rows n1-knn3+1:n1 for ascending order.
let c3_hat(i) = mean(knn3_c)

copy s_c1 knn4_c;
    include;
        rows 1:knn4. # rows n1-knn4+1:n1 for ascending order.

```

```

let c4_hat(i) = mean(knn4_c)

enddo # i

let d1 = c - c1_hat # the difference of cost and the (mean) retrieved cost(s)
let maer1 = mean(abs(d1)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd1 = mean(d1**2) # mean squared deviation(MSD)
let mad1 = mean(abs(d1)) # mean absolute absolute deviation(MAD)
let m_d1 = mean(d1)
let sd_d1 = stdev(d1)
let d1(n+2) = maer1
let d1(n+3) = msd1
let d1(n+4) = mad1
let d1(n+5) = m_d1
let d1(n+6) = sd_d1
print knn1 maer1 msd1 mad1 m_d1 sd_d1

let d2 = c - c2_hat # the difference of cost and the (mean) retrieved cost(s)
let maer2 = mean(abs(d2)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd2 = mean(d2**2) # mean squared deviation(MSD)
let mad2 = mean(abs(d2)) # mean absolute absolute deviation(MAD)
let m_d2 = mean(d2)
let sd_d2 = stdev(d2)
let d2(n+2) = maer2
let d2(n+3) = msd2
let d2(n+4) = mad2
let d2(n+5) = m_d2
let d2(n+6) = sd_d2

```

```
print knn2 maer2 msd2 mad2 m_d2 sd_d2
```

```
let d3 = c - c3_hat # the difference of cost and the (mean) retrieved cost(s)  
let maer3 = mean(abs(d3)/c) # mean absolute error rate(MAER) or mean  
absolute percentage error(MAPE)
```

```
let msd3 = mean(d3**2) # mean squared deviation(MSD)
```

```
let mad3 = mean(abs(d3)) # mean absolute absolute deviation(MAD)
```

```
let m_d3 = mean(d3)
```

```
let sd_d3 = stdev(d3)
```

```
let d3(n+2) = maer3
```

```
let d3(n+3) = msd3
```

```
let d3(n+4) = mad3
```

```
let d3(n+5) = m_d3
```

```
let d3(n+6) = sd_d3
```

```
print knn3 maer3 msd3 mad3 m_d3 sd_d3
```

```
let d4 = c - c4_hat # the difference of cost and the (mean) retrieved cost(s)  
let maer4 = mean(abs(d4)/c) # mean absolute error rate(MAER) or mean  
absolute percentage error(MAPE)
```

```
let msd4 = mean(d4**2) # mean squared deviation(MSD)
```

```
let mad4 = mean(abs(d4)) # mean absolute absolute deviation(MAD)
```

```
let m_d4 = mean(d4)
```

```
let sd_d4 = stdev(d4)
```

```
let d4(n+2) = maer4
```

```
let d4(n+3) = msd4
```

```
let d4(n+4) = mad4
```

```
let d4(n+5) = m_d4
```

```
let d4(n+6) = sd_d4
```

```
print knn4 maer4 msd4 mad4 m_d4 sd_d4
```

```
Endmacro
```

Macro # loocv for Fractional function based similarity measure

```
loocv_fn x.1-x.m c c1_hat d1 maer1 msd1 mad1 m_d1 sd_d1 &  
          c2_hat d2 maer2 msd2 mad2 m_d2 sd_d2 &  
          c3_hat d3 maer3 msd3 mad3 m_d3 sd_d3 &  
          c4_hat d4 maer4 msd4 mad4 m_d4 sd_d4; #
```

standardized or normed data

weight w;

knn knn1 knn2 knn3 knn4. # knn1, knn2, knn3 knn4: positive integers less than or equal to half the size of the case-base

```
Mcolumn x.1-x.m x1.1-x1.m w c xi.1-xi.m x0.1-x0.m simil &  
          c1 s_c1 c1_hat d1 knn1_c c2_hat d2 knn2_c c3_hat d3 knn3_c &  
          c4_hat d4 knn4_c# k_sim
```

```
Mconstant rowi row0 j max_x.1-max_x.m min_x.1-min_x.m n n1 i h sim &  
          knn1 maer1 msd1 mad1 m_d1 sd_d1 &  
          knn2 maer2 msd2 mad2 m_d2 sd_d2 &  
          knn3 maer3 msd3 mad3 m_d3 sd_d3 &  
          knn4 maer4 msd4 mad4 m_d4 sd_d4
```

```
Mmatrix x0_v xi_v
```

brief 0 # brief 1

```
name knn1 'knn1' c1_hat 'c1_hat' d1 'diff1' maer1 'mean of |d1|/c' msd1 'mean  
of d1^2' mad1 'mean of |d1|' &
```

```
          m_d1 'mean of d1' sd_d1 'stdev of d1'
```

```
name knn2 'knn2' c2_hat 'c2_hat' d2 'diff2' maer2 'mean of |d2|/c' msd2 'mean  
of d2^2' mad2 'mean of |d2|' &
```

```
          m_d2 'mean of d2' sd_d2 'stdev of d2'
```

```
name knn3 'knn3' c3_hat 'c3_hat' d3 'diff3' maer3 'mean of |d3|/c' msd3 'mean  
of d3^2' mad3 'mean of |d3|' &
```

```
          m_d3 'mean of d3' sd_d3 'stdev of d3'
```

```

name knn4 'knn4' c4_hat 'c4_hat' d4 'diff4' maer4 'mean of |d4|/c' msd4 'mean
of d4^2' mad4 'mean of |d4|' &
    m_d4 'mean of d4' sd_d4 'stdev of d4'
let n = count(x.1)

####
name simil "sim_f"
note 'Fractional function based similarity measure'

do i=1:n
    let row0 = i # putative new case
    copy x.1-x.m x0.1-x0.m;
    include;
    rows row0.
    copy x0.1-x0.m x0_v # {row} new vector

    copy x.1-x.m c x1.1-x1.m c1;
    exclude;
    rows row0.

let n1 = count(x1.1)
do h=1:n1
    let rowi = h # comparative case
    copy x1.1-x1.m xi.1-xi.m;
    include;
    rows rowi.
    copy xi.1-xi.m xi_v # {row} comparative vector

%simil_f1 x1.1-x1.m sim;
weight w;

```

```

        comv xi_v;
        newv x0_v.
    let simil(h) = sim
    erase sim
enddo # h

sort c1 s_c1;
    by simil;
    descending simil.

copy s_c1 knn1_c;
    include;
        rows 1:knn1. # rows n1-knn1+1:n1 for ascending order.
let c1_hat(i) = mean(knn1_c)

copy s_c1 knn2_c;
    include;
        rows 1:knn2. # rows n1-knn2+1:n1 for ascending order.
let c2_hat(i) = mean(knn2_c)

copy s_c1 knn3_c;
    include;
        rows 1:knn3. # rows n1-knn3+1:n1 for ascending order.
let c3_hat(i) = mean(knn3_c)

copy s_c1 knn4_c;
    include;
        rows 1:knn4. # rows n1-knn4+1:n1 for ascending order.
let c4_hat(i) = mean(knn4_c)

```

```
enddo # i
```

```
let d1 = c - c1_hat # the difference of cost and the (mean) retrieved cost(s)
let maer1 = mean(abs(d1)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd1 = mean(d1**2) # mean squared deviation(MSD)
let mad1 = mean(abs(d1)) # mean absolute absolute deviation(MAD)
let m_d1 = mean(d1)
let sd_d1 = stdev(d1)
let d1(n+2) = maer1
let d1(n+3) = msd1
let d1(n+4) = mad1
let d1(n+5) = m_d1
let d1(n+6) = sd_d1
print knn1 maer1 msd1 mad1 m_d1 sd_d1
```

```
let d2 = c - c2_hat # the difference of cost and the (mean) retrieved cost(s)
let maer2 = mean(abs(d2)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd2 = mean(d2**2) # mean squared deviation(MSD)
let mad2 = mean(abs(d2)) # mean absolute absolute deviation(MAD)
let m_d2 = mean(d2)
let sd_d2 = stdev(d2)
let d2(n+2) = maer2
let d2(n+3) = msd2
let d2(n+4) = mad2
let d2(n+5) = m_d2
let d2(n+6) = sd_d2
print knn2 maer2 msd2 mad2 m_d2 sd_d2
```

```

let d3 = c - c3_hat # the difference of cost and the (mean) retrieved cost(s)
let maer3 = mean(abs(d3)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd3 = mean(d3**2) # mean squared deviation(MSD)
let mad3 = mean(abs(d3)) # mean absolute absolute deviation(MAD)
let m_d3 = mean(d3)
let sd_d3 = stdev(d3)
let d3(n+2) = maer3
let d3(n+3) = msd3
let d3(n+4) = mad3
let d3(n+5) = m_d3
let d3(n+6) = sd_d3
print knn3 maer3 msd3 mad3 m_d3 sd_d3

```

```

let d4 = c - c4_hat # the difference of cost and the (mean) retrieved cost(s)
let maer4 = mean(abs(d4)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd4 = mean(d4**2) # mean squared deviation(MSD)
let mad4 = mean(abs(d4)) # mean absolute absolute deviation(MAD)
let m_d4 = mean(d4)
let sd_d4 = stdev(d4)
let d4(n+2) = maer4
let d4(n+3) = msd4
let d4(n+4) = mad4
let d4(n+5) = m_d4
let d4(n+6) = sd_d4
print knn4 maer4 msd4 mad4 m_d4 sd_d4
Endmacro

```

Macro # loocv for Mahalanobis distance based similarity measure

```
loocv_mn x.1-x.m c c1_hat d1 maer1 msd1 mad1 m_d1 sd_d1 &
                c2_hat d2 maer2 msd2 mad2 m_d2 sd_d2 &
                c3_hat d3 maer3 msd3 mad3 m_d3 sd_d3 &
                c4_hat d4 maer4 msd4 mad4 m_d4 sd_d4; #
standardized or normed data
weight w;
knn knn1 knn2 knn3 knn4. # knn1, knn2, knn3 knn4: positive integers less
than or equal to half the size of the case-base

Mcolumn x.1-x.m  x1.1-x1.m w c xi.1-xi.m x0.1-x0.m simil &
                c1 s_c1 c1_hat d1 knn1_c c2_hat d2 knn2_c c3_hat d3 knn3_c &
                c4_hat d4 knn4_c # k_sim
Mconstant rowi row0 j max_x.1-max_x.m min_x.1-min_x.m n n1 i h
wmdis_sq wmdis sim &
                knn1 maer1 msd1 mad1 m_d1 sd_d1 &
                knn2 maer2 msd2 mad2 m_d2 sd_d2 &
                knn3 maer3 msd3 mad3 m_d3 sd_d3 &
                knn4 maer4 msd4 mad4 m_d4 sd_d4
Mmatrix x0_v xi_v

brief 0 # brief 1
name knn1 'knn1' c1_hat 'c1_hat' d1 'diff1' maer1 'mean of |d1|/c' msd1 'mean
of d1^2' mad1 'mean of |d1|' &
                m_d1 'mean of d1' sd_d1 'stdev of d1'
name knn2 'knn2' c2_hat 'c2_hat' d2 'diff2' maer2 'mean of |d2|/c' msd2 'mean
of d2^2' mad2 'mean of |d2|' &
                m_d2 'mean of d2' sd_d2 'stdev of d2'
name knn3 'knn3' c3_hat 'c3_hat' d3 'diff3' maer3 'mean of |d3|/c' msd3 'mean
of d3^2' mad3 'mean of |d3|' &
```

```

    m_d3 'mean of d3' sd_d3 'stdev of d3'
name knn4 'knn4' c4_hat 'c4_hat' d4 'diff4' maer4 'mean of |d4|/c' msd4 'mean
of d4^2' mad4 'mean of |d4|' &
    m_d4 'mean of d4' sd_d4 'stdev of d4'

let n = count(x.1)
#let n = n - 1

####
    name simil 'sim_m'
    note 'Mahalanobis distance based similarity measure'

do i=1:n
    let row0 = i # putative new case
    copy x.1-x.m x0.1-x0.m;
    include;
        rows row0.
    copy x0.1-x0.m x0_v # {row} new vector

    copy x.1-x.m c x1.1-x1.m c1;
    exclude;
        rows row0.

let n1 = count(x1.1)
do h=1:n1
    let rowi = h # comparative case
    copy x1.1-x1.m xi.1-xi.m;
    include;
        rows rowi.
    copy xi.1-xi.m xi_v # {row} comparative vector

```

```

%simil_m1  x1.1-x1.m wmdis_sq wmdis sim;
    weight w;
    comv xi_v;
    newv x0_v.

let simil(h) = sim
erase wmdis_sq wmdis sim
enddo # h

sort c1 s_c1;
    by simil;
    descending simil.

copy s_c1 knn1_c;
    include;
        rows 1:knn1. # rows n1-knn1+1:n1 for ascending order.
let c1_hat(i) = mean(knn1_c)

copy s_c1 knn2_c;
    include;
        rows 1:knn2. # rows n1-knn2+1:n1 for ascending order.
let c2_hat(i) = mean(knn2_c)

copy s_c1 knn3_c;
    include;
        rows 1:knn3. # rows n1-knn3+1:n1 for ascending order.
let c3_hat(i) = mean(knn3_c)

copy s_c1 knn4_c;

```

```

include;
    rows 1:knn4. # rows n1-knn4+1:n1 for ascending order.
let c4_hat(i) = mean(knn4_c)

enddo # i

#let n = n + 1
let d1 = c - c1_hat # the difference of cost and the (mean) retrieved cost(s)
let maer1 = mean(abs(d1)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd1 = mean(d1**2) # mean squared deviation(MSD)
let mad1 = mean(abs(d1)) # mean absolute absolute deviation(MAD)
let m_d1 = mean(d1)
let sd_d1 = stdev(d1)
let d1(n+2) = maer1
let d1(n+3) = msd1
let d1(n+4) = mad1
let d1(n+5) = m_d1
let d1(n+6) = sd_d1
print knn1 maer1 msd1 mad1 m_d1 sd_d1

let d2 = c - c2_hat # the difference of cost and the (mean) retrieved cost(s)
let maer2 = mean(abs(d2)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd2 = mean(d2**2) # mean squared deviation(MSD)
let mad2 = mean(abs(d2)) # mean absolute absolute deviation(MAD)
let m_d2 = mean(d2)
let sd_d2 = stdev(d2)
let d2(n+2) = maer2
let d2(n+3) = msd2

```

```

let d2(n+4) = mad2
let d2(n+5) = m_d2
let d2(n+6) = sd_d2
print knn2 maer2 msd2 mad2 m_d2 sd_d2

```

```

let d3 = c - c3_hat # the difference of cost and the (mean) retrieved cost(s)
let maer3 = mean(abs(d3)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd3 = mean(d3**2) # mean squared deviation(MSD)
let mad3 = mean(abs(d3)) # mean absolute absolute deviation(MAD)
let m_d3 = mean(d3)
let sd_d3 = stdev(d3)
let d3(n+2) = maer3
let d3(n+3) = msd3
let d3(n+4) = mad3
let d3(n+5) = m_d3
let d3(n+6) = sd_d3
print knn3 maer3 msd3 mad3 m_d3 sd_d3

```

```

let d4 = c - c4_hat # the difference of cost and the (mean) retrieved cost(s)
let maer4 = mean(abs(d4)/c) # mean absolute error rate(MAER) or mean
absolute percentage error(MAPE)
let msd4 = mean(d4**2) # mean squared deviation(MSD)
let mad4 = mean(abs(d4)) # mean absolute absolute deviation(MAD)
let m_d4 = mean(d4)
let sd_d4 = stdev(d4)
let d4(n+2) = maer4
let d4(n+3) = msd4
let d4(n+4) = mad4
let d4(n+5) = m_d4

```

```
let d4(n+6) = sd_d4
print knn4 maer4 msd4 mad4 m_d4 sd_d4
```

Endmacro

Macro # Ratio standardization or maximum score standardization

```
ratiostd x.1-x.m; # input
  save y.1-y.n; # output
  direct bc. # "b": benefit, "c": cost
```

```
mcolum x.1-x.m y.1-y.n
mconstant bc j
```

```
default bc="b"
```

```
if bc = "b"
  do j=1:m
    let y.j = x.j/max(x.j)
  enddo
elseif bc = "c"
  do j=1:m
    let y.j = min(x.j)/x.j
  enddo
endif
```

Endmacro

Macro

simil_a1 x.1-x.m sim_a; # standardized or normed data

weight w;

comn rowi; # row number for comparative case

use only one of these two comcases

comv xi_v; # row vector for comparative case

newn row0; # row number for new case

use only one of these two newcases

newv x0_v. # row vector for new case

Mmatrix xi_v x0_v

Mcolumn x.1-x.m w xi.1-xi.m x0.1-x0.m d

Mconstant sim_a rowi row0 j max_x.1-max_x.m min_x.1-min_x.m

name sim_a 'sim_a'

if comv = 1

copy xi_v xi.1-xi.m

elseif comn = 1

copy x.1-x.m xi.1-xi.m;

include;

rows rowi.

endif

if newv = 1

copy x0_v x0.1-x0.m

elseif newn = 1

copy x.1-x.m x0.1-x0.m;

include;

rows row0.

```

endif

let sim_a = 0
do j=1:m
let max_x.j = max(x.j)
let min_x.j = min(x.j)
let sim_a = sim_a + w(j)*(1-abs((xi.j - x0.j)/(max_x.j - min_x.j)))
enddo

# Print sim_a

Endmacro

```

Macro

simil_e1 x.1-x.m wdis_sq wdis sim_e; # standardized or normed data

```

weight w;
comn rowi; # row number for comparative case
           # use only one of these two comcases
comv xi_v; # row vector for comparative case
newn row0; # row number for new case
           # use only one of these two newcases
newv x0_v. # row vector for new case

```

Mcolumn x.1-x.m w xi.1-xi.m x0.1-x0.m d

Mmatrix W_m dW d_t xi_v xi_t x0_v x0_t

Mconstant wdis wdis_sq sim_e rowi row0

```
# name wdis 'wdis' wdis_sq 'wdis_sq' sim_e 'sim_e'
```

```

if comn = 1
copy x.1-x.m xi.1-xi.m;
include;
rows rowi.
copy xi.1-xi.m xi_v
endif

if newn = 1
copy x.1-x.m x0.1-x0.m;
include;
rows row0.
copy x0.1-x0.m x0_v
endif

transpose xi_v xi_t
transpose x0_v x0_t

Subtract xi_t x0_t d
Transpose d d_t

Diagonal w W_m
Multiply d_t W_m dW
Multiply dW d wdis_sq
Let wdis = sqrt(wdis_sq)
Let sim_e = 1- wdis
# Print wdis_sq wdis sim_e

Endmacro

```

Macro

simil_f1 x.1-x.m sim_f; # standardized or normed data

```
weight w;  
comn rowi; # row number for comparative case  
# use only one of these two comcases  
comv xi_v; # row vector for comparative case  
newn row0; # row number for new case  
# use only one of these two newcases  
newv x0_v. # row vector for new case
```

```
Mmatrix xi_v x0_v
```

```
Mcolumn x.1-x.m w xi.1-xi.m x0.1-x0.m d
```

```
Mconstant sim_f rowi row0 j max_x.1-max_x.m min_x.1-min_x.m
```

```
# name sim_f 'sim_f'
```

```
if comv = 1
```

```
copy xi_v xi.1-xi.m
```

```
elseif comn = 1
```

```
copy x.1-x.m xi.1-xi.m;
```

```
include;
```

```
rows rowi.
```

```
endif
```

```
if newv = 1
```

```
copy x0_v x0.1-x0.m
```

```
elseif newn = 1
```

```
copy x.1-x.m x0.1-x0.m;
```

```
include;
```

```
rows row0.
```

```
endif

let sim_f = 0
do j=1:m
let max_x.j = max(x.j)
let min_x.j = min(x.j)
let sim_f = sim_f + w(j)*(1+abs((xi.j - x0.j)/(max_x.j - min_x.j)))**(-1)
enddo

# Print sim_f

Endmacro
```

Macro

simil_m1 x.1-x.m wmdis_sq wmdis sim_m; # standardized or normed

data

irows i.1-i.n; # rows to be included for calculating the covariance matrix S

erows e.1-e.b; # rows to be excluded

weight w;

comn rowi; # row number for comparative case

use only one of these two comcases

comv xi_v; # row vector for comparative case

newn row0; # row number for new case

use only one of these two newcases

newv x0_v. # row vector for new case

Mcolumn x.1-x.m w xi.1-xi.m x0.1-x0.m d

Mmatrix S W_m S_inv WS_inv dWS_inv d_t xi_v xi_t x0_v x0_t

Mconstant wmdis wmdis_sq sim_m rowi row0 # i.1-i.n

name wmdis 'wmdis' wmdis_sq 'wmdis_sq' sim_m 'sim_m'

if comn = 1

copy x.1-x.m xi.1-xi.m;

include;

rows rowi.

copy xi.1-xi.m xi_v

endif

if newn = 1

copy x.1-x.m x0.1-x0.m;

include;

rows row0.

```
copy x0.1-x0.m x0_v
endif
```

```
transpose xi_v xi_t
transpose x0_v x0_t
```

```
Subtract xi_t x0_t d
Transpose d d_t
```

```
Covariance x.1-x.m S
# Rows i.1-i.n.
```

```
Diagonal w W_m
Invert S S_inv
```

```
Multiply W_m S_inv WS_inv
Multiply d_t WS_inv dWS_inv
Multiply dWS_inv d wmdis_sq
Let wmdis = sqrt(wmdis_sq)
Let sim_m = 1- wmdis
```

```
# Info S wmdis sim_m
# Print S # wmdis sim_m
```

```
Endmacro
```


國文抄錄

建築 建設工事의 選別的 事例基盤推論에 의한 初期 工事費 豫測

건축 건설공사의 성공여부는 초기단계 공사비 예측의 높은 정확도에 달려있다. 특히, 초기단계의 정확한 공사비 예측은 향후 설계 및 시공단계에서의 공사비 절감을 도모하고, 효율적인 원가관리를 가능하게 한다. 하지만, 건축 건설공사의 부정확한 예산 산정과 초기단계 건설공사 관련 정보 부족, 그리고 공사비 관련 단가집의 제한된 활용성 및 다양한 건축 건설공사에 대응할 수 있는 공사비 예측모델의 유연성 부족으로 인해, 발주자와 견적 업무 담당자는 이에 대응할 수 있는 효과적인 공사비 예측 전략 수립이 필요한 상황이다.

전술된 문제를 해결 하기 위해, 본 연구는 건축 건설공사의 선별적 사례기반추론에 의한 초기 공사비 예측 방법론을 제안하였다. 본 연구에서 제안하는 공사비 예측 방법론을 통해 공사비 예측의 정확도 향상과 예측된 공사비에 대한 발주자 및 견적 업무 담당자의 신뢰성 확보, 그리고 공사비 예측 산정 절차의 투명화를 향상 시키고자 하였다.

제안된 공사비 예측 방법론은 *모듈 1: 사례기반 구축, 모듈 2: 사례기반추론 모델 방법 선정, 모듈 3: 사례기반추론 공사비 예측으로* 구성된다. 특히, *모듈 2: 사례기반추론 모델 방법 선정은 서브모듈 1: 정규화 방법 선정 (구간 정규화, 가우스분포 정규화, Z-점수 정규화, 로지스틱함수 정규화, 비 정규화), 서브모듈 2: 가중치 산정 방법 선정 (속성영향력, 엔트로피, 동일가중치, 유전자 알고리즘), 서브모듈 3: 유사도 산정 방법 선정 (마할라노비스 거리 기반, 유클리디언 거리 기반, 산술합산 기반, 분수함수기반)*으로 구성된다.

본 연구에서 제안한 건축 건설공사의 선별적 사례기반추론에 의한 초기단계 공사비 예측 방법론의 타당성을 검증하기 위해 공공아파트 (100개 사례), 병영생활관 (117개 사례), 정부청사 (52개 사례)를 대상으로 하나씩-빼기 교차타당화를 통한 사례연구를 진행하였다. 공사비 예측값은 절대평균오차율과 평균제곱편차, 그리고 절대평균편차에 의한 정확성을 비교하고, 표준편차에 의한 안정성, 커널밀도추정에 의한 적절성을 비교하였다. 또한, 다양한 건축 건설공사 특성을 반영하여 가장 정확하고 안정적인 정규화 방법, 가중치 산정 방법, 유사도 산정 방법을 제시하는 선별적인 사례기반추론 모델의 유연성을 실험하였다.

검증 결과, 공공 아파트의 경우 비 정규화 방법과 유전자 알고리즘 가중치 산정 방법, 그리고 산술합산 기반 유사도 산정 방법의 사례기반추론 공사비 예측 모델이 정확도 및 안정성이 가장 높은 조합으로 제시되었다. 병영생활관의 경우, 구간 및 비 정규화 방법과 유전자 알고리즘 가중치 산정 방법, 그리고 유클리디언 거리 기반 유사도 산정 방법의 사례기반추론 공사비 예측 모델이 가장 우수한 것으로 도출되었다. 공공청사의 경우, 비 정규화 방법과 속성영향력 가중치 산정 방법, 그리고 분수함수 기반 유사도 산정 방법의 사례기반추론 공사비 예측 모델이 가장 높은 정확도 및 안정성을 가지는 것으로 선별되었다.

본 연구에서 제안한 데이터 전처리 과정을 통한 사례구축 절차를 활용하여 초기단계 공사비 예측 결과의 타당성 및 신뢰성을 향상시킬 수 있을 것으로 기대된다. 또한, 선별적 사례기반추론 모델을 적용하여 다양한 특성을 가진 사례에 대한 모델의 예측 정확도와 설명력 향상을 검증하였으며, 이를 통해 다이내믹한 건축 건설공사의 초기 공사비 예측에 보다 유연하게 대응할 수 있을 것으로 기대된다.

주요어: 초기단계 공사비 예측, 건축 건설공사, 선별적 사례기반추론, 데이터전처리, 정규화, 가중치 산정, 유사도 산정

학 번: 2011-30176