The Hong Kong Polytechnic University

IMPROVING THE LONG-TERM USE OF CASE-BASED REASONING MODEL IN EARLY CONSTRUCTION COST ESTIMATION

XUE XIAO

PhD

This programme is jointly offered by The Hong Kong Polytechnic University and Queensland University of Technology

2021

The Hong Kong Polytechnic University

Department of Building & Real Estate

Queensland University of Technology

School of Civil Engineering and the Built Environment

Improving the long-term use of case-based reasoning model in early construction cost estimation

Xue Xiao

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

June 2020

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

Xue Xiao

_____

Abstract

Construction cost estimation is a significant task because it provides valuable financial information for project decision making. In the early stage, the estimate is typically used to conduct initial feasibility studies. Since the flexibility to adjust the project scope, design, specification, and standards needs to be very high in this stage, construction cost estimation should be made as early as possible. However, it is irrational for construction companies to over-invest their design time and effort in the early stage with limited human resources. An estimator's effort will be more valuable in the later estimation stages. Many researchers therefore have explored ECCE using various techniques. These techniques, which turn out to be helpful in assisting ECCE, are based on the historical database containing previous similar projects and the target project.

By its very nature, early stage estimation involves a large amount of subjectivity, making the estimator's experience of vital importance. This makes case-based reasoning (CBR) an obvious candidate, as it not only includes such basic components as a reasoning cycle and case-base, but also relies on experience or knowledge – making it a perfect match compared with other methods. Most importantly, it is also advantageous for long-term use.

Since data growth has altered the way information is stored and processed, the continuously increasing amount of construction cost information facilitates the amount of data available. This fact creates huge opportunities and challenges for using the CBR model in ECCE. On the one hand, data growth greatly enriches the knowledge and experience in case-base, improving its overall performance. On the other hand, the historical database will continually increase over time as more data is added to it, resulting in a high requirement for updating and maintaining the ability of the case-base. In particular, the changed resource costs, construction methods, design styles and economic conditions create outdated and inconsistent data, which should be carefully handled and eliminated. Without proper handling, these stale data will impair the performance of each component: the typical issues being the unstable knowledge structure and low efficiency because of the continuously increasing size of the case-

base during long-term use. This inevitably raises the problem between the benefits of having more data and the deficiencies of having inappropriate data.

This study therefore aims to improve the practice of long-term use of case-based reasoning in ECCE from several perspectives: understanding the parameter settings of CBR in the existing research and the case-base's influence on accuracy; improving the robustness of the CBR system; and enhancing the efficiency of the CBR system. It firstly identifies which parameter combinations are better and explores the influence the size of the case-base has on the performance of the CBR model. Then, a robustness weight determination method is introduced to improve the robustness of the ECCE CBR model, followed by an original case-base maintenance method based on weight coverage contribution to improve the efficiency of the reasoning process.

The results indicate that the GA-CBR model is more accurate when the size of the case-base is small and there is no significant difference in accuracy between the MRA-CBR model and the GA-CBR model when the size of the case-base becomes large. However, GA does not have the advantage of producing a stable structure in the case-base during long-term use, while the MODAL-CBR model effectively improves the robustness and accuracy of the results. The CBM methods are classified into three strategies according to how the case-base and weight determination are selected. Strategy 1 and Strategy 2 can significantly compress the size of case-base for all CBR models. Strategy 1 generates better results in OLS-CBR while Strategy 2 generates better results in MODAL-CBR. More specifically, Strategy 1 can maintain the OLS-CBR's performance while reducing the size of case-base by 28.13%. There is also a slight improvement in the accuracy of OLS-CBR models when the size of the case-base is slightly reduced.

The study provides valuable knowledge for improving the ECCE CBR model for long-term use, both theoretically and practically. The methods used not only help improve the performance of the CBR system by enhancing the robustness of generating a stable case-base knowledge structure, but also help maintain the efficiency of the case-base during long-term use. Valuable ideas are also provided of how future work can be conducted to improve the ECCE CBR model in the future.

Publications

Xiao, X., Skitmore, M., Li, H., & Xia, B. (2019). Mapping knowledge in the economic areas of green building using scientometric analysis. *Energies, 12*(15), 3011.

Xiao, X., Wang, F., Li, H., & Skitmore, M. (2018). Modelling the stochastic dependence underlying construction cost and duration. *Journal of Civil Engineering Management 24*(6), 444-456.

Xiao, X., Skitmore, M., & Hu, X. (2017). Case-based Reasoning and Text Mining for Green Building Decision Making. *Energy Procedia, 111*, 417-425.

Acknowledgements

I am overwhelmed with gratitude to all those who have helped me along the way to completing this thesis. I would like to thank the Hong Kong Polytechnic University for providing me an academic exchange program. I would like to express my thanks to many people in the Hong Kong Polytechnic University. Firstly, I am thankful to my supervisor, Professor Heng Li, for his guidance and support during my exchange program. Thanks to him, I understand the importance of innovation and self-discipline for achieving academic success. It was a great honour and privilege to study under his guidance. I also want to express my special thanks to Professor Ni Meng, Chairman of the departmental research committee, for his care and support during my exchange at the Hong Kong Polytechnic University, and his encouragement when I encountered difficulties in my Ph.D. journey. As well,, I would like to thank Chloe in the BRE department and Jenny at the research office for their help and support when I needed assistance.

Also, I would like to thank the Queensland University of Technology for providing me an academic exchange program. I would like to thank many people at Queensland University of Technology. Firstly, I would like to express my special thanks to my supervisor, Professor Martin Skitmore, who allowed me to undertake this PhD journey. He has been a fabulous mentor and supporter of my research and career development. His hard-working manner, vision and sincerity have deeply inspired me. I would also like to thank him for his friendship, empathy and his great sense of humour, which helped me to smile at life. I would like to thank my associate supervisor, Associate Professor Bo Xia. I would like to thank Bo for helping me to refine my research topic and guiding my thesis. I also want to thank him for his encouragement when I encountered difficulties in my PhD journey.

I am grateful to the School of Civil Engineering and Built Environment for providing excellent facilities and services. Support from the QUT SEF HDR team is also sincerely appreciated. I acknowledge the services of professional editor, Diane Kolomeitz, who provided copyediting and proofreading services, according to the guidelines laid out in the university-endorsed national ‘Guidelines for editing research theses’.

Apart for the academic advisors mentioned, I made many friends when I visited universities and attended conferences. I would like to express my special thanks to them as well as all my friends for their fellowship and support.

Lastly, and most importantly, I wish to thank my families and my boyfriend for their support and encouragement. They have always given me selfless love, generous support, and timely encouragement throughout my life. They are always happy to hear me tell stories of my experiences. I would like to thank them in Chinese now:

感谢各位一直以来的关爱与支持。是你们让我知道了陪伴和爱是人生中最重要的事情。我一定继续努力，不辜负你们的期望！

Keywords

Early construction cost estimation, case-based reasoning, long-term use, comparative study, robustness, case-base maintenance

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ACCM	Accuracy-Classification Case Memory
AHP	Analytic hierarchy process
AI	Artificial intelligence
ANN	Artificial neural networks
AS	Attribute similarity
ASH	Average storey height
BBNR	Blame Based Noise Reduction
CBM	Case base maintenance
CBR	Case-based reasoning
CRR	Conservative Redundancy Reduction
CC	Coverage contribution
CCR	Coverage contribution ratio
CCW	Coverage contribution weight
CD	Commencement date
CNN	Condensed Nearest Neighbour
CS	Case similarity
D	Duration
DARPA	Defence Advanced Research Projects Agency
DBSCAN	Density Based Spatial Clustering of Application with Noise
ECC	Expansion-contraction compression
ECCE	Early construction cost estimation
ENN	Edited Nearest Neighbour
EWCBR	European Workshop on CBR
FC	Feature counting
FD	Finish date
FT	Foundation type
ICCBR	International Conference on CBR
ICF	Iterative Case Filtering
ICMS	International Construction Measurement Standards

IJCAI	International Joint Conference on Artificial Intelligence Workshop
GA	Genetic algorithms
GDM	Gradient descent method
GM	Gaussian-Means
K-NN	K-nearest neighbour
LBL	Instance base Learning
MAPE	Mean average percent error
MCAS	Minimum criterion for attribute similarity
MCS	Monte Carlo simulation
MODLR	Modal regression
MRA	Multiple regression analysis
MSE	Mean squared errors
NACCM	Negative Accuracy-Classification Case Memory
NN	Nearest neighbour
NSAG	No. of storeys above the ground
NSBG	No. of storeys below the ground
PCA	Principle component analysis
PPP	Public-private partnership
RBM	Reputation-based maintenance
RENN	Repeated Edited Nearest Neighbour
RICS	Royal Institution of Chartered Surveyors
RMSE	Root Mean Squared Error
RNN	Reduced Nearest Neighbour
RST	Rough sets theory
SEM	Storey Enclosure Model
SNN	Selective Nearest Neighbour Rule
STO	Storey
TAFA	Total above floor area
TBFA	Total below floor area
TFA	Total floor area
THB	Total height of building

Chapter 1: Introduction

1.1 BACKGROUND

Construction cost estimation is a significant task because it provides valuable financial information for project decision making (Tahir et al., 2018). The estimated cost helps managers to make justified decisions based on different project stages. Inaccurate cost estimation leads to missed development opportunities, low efficiency in use of resources, and unsuccessful project management (Oberlender & Trost, 2001b; Themsen, 2019). Various construction activities, including the bidding preparation, cost monitoring, and control of projects during the construction stage, as well as financial performance evaluation of the project, are significantly influenced by the result of construction cost estimation (Akintoye & Fitzgerald, 2000).

The role of construction cost estimation differs according to different project stages. In the early stage, the estimate is typically used to conduct initial feasibility studies. The early construction cost estimation (ECCE) helps managers to choose adequate alternatives and to avoid misjudging solutions at the early stage (Arafa & Alqedra, 2011a). ECCE influences various construction-planning activities, including team management, scheduling, and fundraising (Akintoye & Fitzgerald, 2000; Cheng et al., 2009; Larsen et al., 2015b). Since early decisions have the most influence on the overall project performance, the construction cost estimation should be conducted as soon as possible (Lowe, Emsley, & Harding, 2006; Shin, 2015).

However, it is inefficient for construction companies to over-invest their design time and effort in the early stage with limited human resources. An estimator's time and effort will be more valuable in the later estimating stages. Many researchers therefore have explored ECCE using various databased techniques. These techniques turn out to be helpful in assisting ECCE and are based on the historical database containing previous similar projects as the target project.

Compared with other methods, case-based reasoning (CBR) relies on experience and knowledge, making it a perfect match for ECCE. (An et al., 2007; Chen & Burrell, 2001). By recalling previous projects in the case-base and adjusting to new requirements, CBR is considered to be more powerful than other approaches in long-term use (KARANCI, 2010; Kim et al., 2004b). CBR is a problem-solving process

consisting of case-base and reasoning cycle. Case-base is a knowledge and experience container, which is made up of previous structured cases. The reasoning cycle is the process of retrieving and reusing the previous knowledge and experience for solving a target case. As the fundamental part of the CBR model, the case-base significantly influences the reasoning cycle (Smiti & Elouedi, 2018a).

Since data growth has altered the way information is stored and processed, the increasing amount of construction cost information makes more data available (Bilal et al., 2016; García-Gil et al., 2019; Ilyas et al., 2015). This fact creates huge opportunities and challenges for using the CBR model in ECCE. On the one hand, data growth greatly enriches the knowledge and experience in case-base, improving its overall performance. On the other hand, the historical database will continually increase over time as more data is added to it, resulting in a high requirement for updating and maintaining the ability of the database.

The popularity of data warehouses has highlighted the need to address the fitness of data for use. Fitness of data for use also refers to the data quality. Data quality implies that one needs to look beyond traditional concerns with the accuracy of the data (Tayi & Ballou, 1998). There are four dimensions in data quality: accuracy, completeness, consistency, and timeliness (Ballou & Pazer, 1985). Accuray is defined as the correctness of the fact recoding completeness as the relevance of the information recorded, consistency as the uniformity in the information recored, timeliness as the recording of information on time. Poor data quality would result in negative impact on operation of the system (Coetzer & Vlok, 2019). Furthermore, poor data quality cannot fullfil the requirement of the system, consequently resulting in the failure in achieving the expected results (Karkouch et al., 2016; Laranjeiro et al., 2015; Taleb et al., 2016).

Construction cost estimation usually needs the cost data of the historical projects. The historical cost data will be useful for cost estimation only if they are collected and organized in a way that is compatible with future applications. The consistency of the construction cost data is critical, since it provides the reliable baseline for the new project. Therefore, the information must be updated with respect to changes that will inevitably occur (Hendrickson et al., 2008). Without sufficient refining and updating, historical cost data shouldn't be used carelessly. Changes in relative prices may have substantial impacts on construction costs, which have increased in relative price.

Unfortunately, systematic changes over a long period of time for such factors are inevitable. In particular, the changed resource costs, construction methods, design styles and economic conditions create the outdated and inconsistent data. Also, the size of the case-base can grow quickly with the continuous use of the CBR model (Smiti & Elouedi, 2018a). The efficiency of solving a new problem thus becomes increasingly slow, resulting in compromised overall performance of the CBR model (Khan et al., 2019b; Lupiani et al., 2014b). Without proper handling, these problems caused by long-term use will impair the performance of the CBR model, the typical issues being the unstable knowledge structure and low efficiency because of the continuously increasing size of the case-base during long-term use. This inevitably raises the optimization problem between the benefits of having more data and the deficiencies of having inappropriate data.

Therefore, this thesis attempts to improve the long-term use of the CBR model in ECCE. By addressing the gaps and limitations in the current studies, this research aims to answer the research question of how to improve the robustness of the CBR system and maintain case-base, while avoiding excessive storage and time complexity during its long-term use.

1.2 RESEARCH PROBLEM

The main aim of this research is to improve the CBR model of ECCE for long-term use. Based on the previous research background, four particular research problems are proposed in achieving the research aims as follows:

1. *What limitations exist in the current ECCE CBR studies with respect to long-term use?*

A comprehensive literature review on existing ECCE research fields is conducted to find the research gaps and limitations in previous studies. After carefully examining the current ECCE studies, the subsequent research questions are identified and proposed.

2. *What are the main differences caused by different parameter settings of ECCE CBR model?*

The literature review finds various parameters, including weight determination, similarity functions, and case adaptation, which are used in the previous ECCE CBR

model. However, there is no consensus on how to combine these parameters to achieve the optimal results in the CBR model. The literature review also finds that although several studies deem CBR advantageous for long-term use, there is no empirical study illustrating changes in performance of the CBR model with the increase in the number of cases in the case-base.

This begs the question of the role of case-base size. Similar to sample size, case-base size is the number of cases used to establish case-base. Different models may have different performance characteristics when dealing with various sample sizes - thought by some to be more attributable to the differences in results than the models themselves (Ji et al., 2010a; Marshall et al., 2013; Motrenko et al., 2014; RunZhi et al., 2012; Wolf et al., 2013). In ECCE, it is necessary to take the influence of sample size into account when estimating costs by sampling from projects similar to the target project. However, too careful sampling produces less data to synthesize, and therefore less accuracy– resulting in the need to trade-off between the sample size used and its similarity to the target (Skitmore, 2001; Yeung & Skitmore, 2012; Yeung & Skitmore, 2005). This research problem is a necessary part of the main research problem, and the research findings of this research problem facilitate answering the subsequent research question (c) and research question (d).

3. *How to maintain a stable knowledge structure of the CBR model during long-term use?*

The literature review identifies some limitations on weight determination in the current ECCE CBR model. In the CBR model, the solution to a target case generates from most similar previous cases. This process is significantly influenced by the attribute weight. In the CBR system, each attribute can be seen as an index contains a part of the knowledge stored in the case-base. Attribute weight reflects the influence of this knowledge component on case-base. Therefore, attribute weights can be deemed as indicators of the overall knowledge structure of the case-base.

In the ECCE CBR model, attribute weights inevitably change due to updating and refining of the case-base. However, these changes should be minimized because of the consistency requirement of the knowledge structure in case-base. Attribute weights should be maintained as consistent as possible to provide reliable results. The stability of the attribute weights can be considered as the indicators of the CBR model's robustness. Although several attribute weight determination methods have been

exploited in the current ECCE CBR model, how to maintain the stability of attribute weight in ECCE during the long-term use remains a question. When the size of the case-base gets large, it is inevitable to have a few cases which deviate considerably from the bulk of cases (Chan & Wong, 2007). The existing weight determination methods have limitations in being sensitive to the outliers. A single outlier can have a large injury on the parameter estimates, thus may reduce the accuracy of the model. Therefore, improving the robustness of the ECCE CBR is the primary task during long-term use.

4. How to improve the efficiency of the ECCE CBR model for long-term use?

Additionally, the literature review finds some limitations on case-base maintenance (CBM) in the current ECCE CBR model. CBM is the process of refining the case-base to enhance CBR's performance. Since case-base is a fundamental component in the CBR system, numerous studies emphasize that additional maintenance of this case-base is necessary in the CBR system, especially when the knowledge in case-base changes over time (Lupiani et al., 2014a). However, the existing ECCE CBR model extensively focuses on the initial establishment of the reasoning cycle, resulting in ignorance of the case-base maintenance during long-term use.

During the long-term use of the ECCE CBR model, the historical database will continually increase over time as more data is added to it, resulting in a high requirement for maintaining the quality of the case-base. In particular, changed resource costs, construction methods, design styles and economic conditions create outdated and inconsistent data, which should be carefully handled. Also, the size of the case-base can grow very fast with the continuous use of the CBR model (Smiti & Elouedi, 2018a). The efficiency of solving a new problem thus becomes slow, resulting in the compromised overall performance of the CBR model (Khan, et al., 2019b; Lupiani, et al., 2014b). Without proper handling, long-term use will impair the performance of the CBR model, the typical issues being low efficiency because of the continuously increasing size of the case-base. This inevitably raises the problem between the benefits of having more data and the deficiencies of having inappropriate data.

1.3 AIMS AND OBJECTIVES

To address the above research issues, this study proposes four research objectives:

1. *To provide a comprehensive literature on the previous studies on ECCE CBR model*

To have an in-depth understanding of the ECCE CBR model, it is necessary to conduct an extensive literature review. The literature review chapter aims to provide a comprehensive summary of the related work on construction cost estimation, application of CBR in ECCE, and case-based maintenance. The literature review on construction cost estimation aims to provide the necessary knowledge to understand the research question. The literature review on the application of CBR in ECCE seeks to provide a careful examination of each step of using in ECCE CBR models and summarizes the widely used weight determination and similarity function, as well as identifying the research gaps. The literature review on CBM aims to provide the knowledge for case-base maintenance during long-term use.

2. *To conduct an empirical study to compare the methods for calculating weight and similarity, as well as exploring the influence of sample size on accuracy of ECCE CBR.*

The existing ECCE CBR models exploit various parameters, including weight determination, similarity functions and case adaptation. To understand how to combine these parameters to achieve the optimal results in the CBR model, this thesis aims to conduct a comparative study of different weight determination methods, similarity functions and case adaptation values. This thesis attempts to provide a comprehensive understanding of the effect of different combinations of parameters in the ECCE CBR model.

Comparing the differences of accuracy results based on different sizes of case-base, this thesis aims to provide an empirical study to test the hypothesis that CBR has the advantage over ANN and regression for long-term use. By exploring the influence of the sample size on the CBR model's performance, this research aims to provide some insights in the contradictory results in previous studies. Additionally, comparison among different combinations of weight determination methods and similarity functions is also conducted. The research findings in this chapter can help

to provide a better understanding of the ECCE CBR model's performance. The results also assist in finding directions for improving the long-term use of CBR.

3. *To improve the robustness of the ECCE CBR model in long-term use.*

Since attribute weight determination significantly influences the reasoning process in CBR, it is necessary to optimize the weighting process from various perspectives. In the ECCE CBR model, attribute weights inevitably change due to updating and refining of the case-base. The existing research only focuses on the weight optimization in the initial establishment of the CBR model, while ignoring the reliability and consistency of the knowledge structure during the long-term use. The limitations in the existing studies that use equal weight for all cases in the case-base for attribute weighting, inevitably lead to the unstable knowledge structure of the case-base (Chan & Wong, 2007).

Therefore, this study aims to improve the robustness of the CBR model by introducing the robust attribute weighting method. By using a robust regression method, the attribute weighting in this study is more focused on the mainstream bulk of cases, resulting in being less sensitive to the changes in the case-base (Yao & Li, 2014). A robust regression,modal regression (MODLR), is used in weighting attributes. MODLR is developed based on the conditional mode of the response Y, and thus focuses on the main characteristics of the data. The mode takes the value with the highest probability of occurrence and thus focuses on the main features of the data (Yao & Li, 2014). Using MODLR in weight determination, the research aims to improve the robustness of the CBR model by reducing the influence of the changes in the case-base on weighting attributes.

4. *To develop a CBM strategy for ECCE CBR models to maintain its efficiency during long-term use.*

The performance of a CBR system can be evaluated from several dimensions. There is no guarantee that all of them can be maximized simultaneously. Several constraints will limit the performance of the CBR model when solving a real-world problem. These constraints include the limit of the case-base size and the limit of time. Without proper maintenance on case-base, the performance of the ECCE CBR model when solving the real-world problem will be inferior. How to improve the performance

of the CBR model when solving the real-world problem motivates this study to explore the case-base compression in the framework of ECCE.

Given the constraints in the real world, the typical issues of the low efficiency caused by the continuously increasing size of the case-base are addressed in this study. This research aims to develop a CBM method to effectively refine and update the case-base. An original CBM method based on the weighted coverage contribution is proposed to reduce the storage requirements and enhance the processing cost of the CBR model. Given one training case, the editing rules consider either deleting or keeping the case unchanged according to their coverage contribution in the case-base. The cases with the lowest coverage contribution will be eliminated. The performance of the ECCE CBR model was compared before and after using the proposed CBM approach.

1.4 SIGNIFICANCE

By solving the problems of ECCE CBR during long-term use, this research contributes to improving the robustness and efficiency of the ECCE CBR models. The research outcomes may support work related to cost estimation for decision-makers ranging from beginners to experts in both academia and industry. This research can be broken into four research questions, and the significance of answering each research question is illustrated as follows.

Firstly, this thesis conducts a comprehensive literature review on existing ECCE CBR studies. By providing a careful examination of each step of the CBR models in ECCE, the limitations in the previous research are identified and summarized. The literature review helps both researchers and practitioners achieve a better understanding of historied development and status quo of ECCE CBR. The gaps and limitations identified in current research provide the potential guidelines for improving the ECCE CBR model.

This thesis firstly conducts a comparative study on attribute weighting and similarity function for various sample sizes. This part of the research may provide a reason for the contradictory results of previous studies. Also, comparison among different weight determination methods and similarity functions helps to understand the differences among the different parameter settings. Analysing the influence of sample size on the accuracy of the CBR model may help to understand the changes in

long-term use. The research has shown its considerable practical significance in the construction industry. Given the confidentiality of the cost data, the data collection for small organizations remains difficult (Gardner, Gransberg, Jeong, et al., 2016). The answer to the second research question not only helps researchers to understand the effect of the different settings in CBR's model but also assists practitioners by minimizing the data collection effort.

Secondly, this thesis introduces a robust weight determination method. It enriches the existing ECCE CBR research in the literature by considering both the single-time performance and the stability of the long-term use. Besides, a robust weight determination method produces a stable knowledge structure and a more reliable result. By focussing on the main characteristics of the conditional distribution, the proposed method can minimize the changes in the knowledge structure in the case-base. This study assists the ECCE CBR system in maintaining a consistent knowledge structure despite continuously updating the case-base. It better prepares construction cost agencies and organisations to tackle the massive growth in the volume of the data and help practitioners have a consistent understanding of the knowledge stored in the case-base.

Finally, this thesis proposes a CBM strategy to improve the efficiency of ECCE by compressing case-base. To address the reduced efficiency caused by the constant increase in the case-base, this study develops a method based on a prototype selection using the weighted coverage contribution. The proposed method can reduce the storage requirements and improve the processing cost of the CBR model. A complete and efficient ECCE CBR model should not only have the function of retrieving, reusing, and updating knowledge in the case-base but also include the removal of useless and outdated cases (Lupiani, et al., 2014a). This chapter contributes to the research area of CBM in the ECCE CBR model by introducing a case selection method during long-term use.

Altogether, this thesis provides valuable theoretical knowledge for improving the ECCE CBR model for long-term use. This study explores the existing parameters in the ECCE CBR model, resulting in a better and comprehensive understanding of the existing ECCE CBR model. Besides, it may help improve the performance of the CBR system by maintaining a stable and consistent knowledge structure, as well as reducing the time cost and storage requirement during the long-term use.

1.5 THESIS OUTLINE

Chapter One provides a brief introduction and research background by illustrating the changes in the construction industry and the challenges ECCE is facing. The main research problem is proposed and broken down into four detailed research questions, each of which constitutes a subsequent chapter in this thesis. The aims and objectives are addressed to understand the motivation of this research, followed by the research significance, illustrating the potential use of this study.

Chapter Two provides a comprehensive literature review including construction cost estimation, application of CBR in ECCE, and CBM. The literature review on construction cost estimation consists of a brief overview, the inaccuracy in construction cost estimation, factors influencing construction cost estimation performance, significance of ECCE, challenges in ECCE, and ECCE methods. The application of CBR in ECCE carefully examines each step of the existing CBR model in ECCE. After briefly introducing the CBR and its advantage in ECCE, problem formulation, case retrieval, case reuse, case revision, and CBM are reviewed to provide an in-depth understanding of the existing research. Section CBM includes the definition of CBM, the criteria for evaluating case-base, influencing factor in CBM, and classification of CBM strategy, case-base reduction strategy, case-base partitioning strategy, and case-base optimization strategy.

Chapter Three introduces the research design. This chapter initially begins with the methodological considerations for solving the research problems. The overall framework of the research design is presented, together with the research hypotheses and their testing process. The process of data collection is then introduced. The procedure is presented to illustrate the overall schedule of the proposed study. The limitations are then summarized to show the potential weakness of the current research, followed by the direction for future studies.

Chapter Four explains the data analysis and pre-processing. This chapter begins with data cleaning, which deals with missing values of data and out-of-range data. The data transformation with respect to data scale and time is represented, followed by the description of the data after removing missing value and outliers. Then the final predictors for developing the model are determined in this study.

Chapter Five provides a comparative study of different weight determination methods, similarity functions and case adaptation values based on different sample

sizes. Three most widely used weight determination methods and the two most commonly used similarity functions are used in the CBR model. Given the sample size range in previous studies, different training sample sizes are used in this study. The Mean Average Percent Error (MAPE) and Root Mean Squared Error (RMSE) are used as measures of ECCE estimating performance.

Chapter Six introduces a robust weight determination method for the ECCE CBR model. A robust regression–MODLR is used in weight determination. After presenting the research background, the MODLR is briefly introduced, and the development of the CBR model using MODLR is illustrated. The comparative results among those using MODLR and others are presented, followed by the discussion and chapter summary.

Chapter Seven develops an original method based on CBM to improve the efficiency of CBR for ECCE. By using the concept of weighted coverage contribution, this chapter develops a case selection strategy for avoiding excessive storage and time complexity caused by the continuous increase in the case-base during the long-term use. This chapter initially begins with the introduction of the research background, followed by an explanation of the proposed CBM strategy. Then it illustrates how to combine the CBR model and the proposed CBM strategy. The comparative results of the CBR model before and after using the proposed approach are used to illustrate the effect of the proposed method, followed by the discussion and chapter summary.

Chapter Eight includes a brief overview of research objectives, conclusions on the research objectives, research contributions, limitations, and future recommendations. The research findings concerning research aim, research questions, and research objectives are reviewed to provide a comprehensive understanding of this research. The research contributions are divided into theoretical and practical perspectives. This chapter also outlines the limitations in this study as well as the recommendations for future ECCE CBR research.

1.6 CHAPTER SUMMARY

This chapter briefly introduces the research background and the research problem concerning the long-term use of the CBR model in ECCE. The challenges brought by the rapid data growth for ECCE and the significance of improving the CBR model's long-term use have been emphasised. The main research question of how to

improve the CBR model in ECCE for long-term use is identified and broken down into four detailed questions:

1. *What limitations exist in the current ECCE CBR studies with respect to long-term use?*
2. *What are the main differences among the existing parameter settings of ECCE CBR model?*
3. *How to maintain a stable knowledge structure of the CBR model during long-term use?*
4. *How to improve the efficiency of the ECCE CBR model for long-term use?*

Then the following research objectives are proposed:

1. *to provide a comprehensive literature on the previous studies on ECCE CBR models；*
2. *to conduct an empirical study to compare the methods for calculating weight and similarity, as well as exploring the influence of sample size on ECCE CBR models;*
3. *to improve the robustness of the ECCE CBR model by combining the CBR system and robust method;*
4. *to develop a method to enhance the efficiency of the ECCE CBR model and maintain its performance during long-term use.*

Then the significance of this research has been emphasised from theoretical and practical perspectives, illustrating the importance and potential application of this research. Except for providing a better understanding of parameter setting in the CBR model, this thesis contributes to improving the robustness and efficiency of the ECCE CBR model during long-term use. Finally, the thesis outline is summarised through a brief description of each chapter.

Chapter 2: Literature Review

2.1 CONSTRUCTION COST ESTIMATION

2.1.1 Overview

This chapter was restricted to studies that present in academic (peer-reviewed) journals. Other sources, such as the International Construction Measurement Standards (ICMS 1 & 2), as well as work done by professional societies in different regions on how costs can be captured and manipulated, have been excluded as outside the scope of this review. The rationale behind this is that peer-reviewed journal articles are the most valuable sources of information, given the quality of the peer review process they go through prior to publication (Darko & Chan, 2017).

The construction industry has a nature of being heterogeneous. There is no common definition of construction industry (Ofori, 2015). The Project Management Body of Knowledge Guide defined the construction cost as the " process of developing an approximation of the cost of resources needed to complete project work." (U.S.), 2017). The construction cost hereby refers to the cost of the resources needed to complete project activities. This work is normally carried out periodically throughout the project as needed.

The cost of a construction project differs due to the role the parties played in the project. From the owners' perspective, the construction cost includes both the initial capital cost and the subsequent operation and maintenance costs. The capital cost covers the monetary expense related to the initial construction of the building including land acquisition, feasibility studies, architectural and engineering design, construction, monitoring of construction, the financial cost during the construction, insurances and taxes during the construction, and equipment that is not included in construction and inspection (Hendrickson, et al., 2008). There are four main cost elements involved in initial capital cost: the project decision cost, bidding cost, design costs and construction cost (Liu & Xie, 2013).

The operation and maintenance costs include the project life cycle cost such as operating staff, labor and material for maintenance and repairs, periodic renovations, insurance and taxes, financing cost and utilities. From the owner's perspective, it is

important to complete the construction with the lowest overall cost that match the project's investment goal, while from the design and construction practicioners, the initial capital cost is the only cost they care about (Hendrickson, et al., 2008).

Construction cost estimation is the process of forecasting the monetary resources needed to complete a project with a defined scope (Stackpole, 2010). It can be deemed as the implementation stage of cost modelling (Skitmore & Marston, 2005). Construction cost estimation is a part of cost engineering. According to the American Association of Cost Engineers, cost engineering is defined as that area of engineering practice where engineering judgment and experience are utilized in the application of scientific principles and techniques to the problem of cost estimation, cost control and profitability.

Construction cost estimation is commonly provided by the cost engineer or estimator on the basis of available information. There are several approaches for construction cost estimation. Production function is one of the widely used approaches in the construction cost estimation. In construction process, the scale of the construction can be used to expressed the production function. The production can be calculated using the function related to labor, material and equipment. Empirical cost inference is also polular in construction cost estimation. Based on mathematic techniques, empirical estimation of cost intends to use a range of project features and attributes to estimate the construction cost. In this approach, mathematics is commonly used to estimate the value of parameters in an assumed function. Unit cost for bill of quantities is another commonly used construction cost estimation method. For each cost component in the bill of quantities, a unit price is assigned. The total cost is the sum amount of the bill of quantities multiplied by the corresponding unit costs. Compared with other methods, unit cost is simpler and more straightforward, but it requires more effort on preparation for the bill of quantities. Allocation of joint costs is used to develop a cost function of an operatioin. The core ideas of this method is that each expenditure item can be assigned to a specific feature of the operation. The allocation of joint costs is expected to be related to the category of basic costs in an allocation process. In some cases, however, this relationship cannot be identified or found.

With the development of information technology, several computer-aided cost estimation systems are now available. From the simple spreadsheet calculation to the

integrated systems such as building information modelling, the computer-aided software greatly facilitates the work of construction cost estimators. Particularly, the efficiency of the construction cost estimation is significantly improved by these computer-aided systems. The computer-aided construction system consists of several components. Database for cost items such as labor, equipment and material are necessary. These databases can be used to estimate any construction project. If these rates change, the results can be updated based on the new rates. Import and output utilities and are also important in the construction aided system for providing the effective estimation system (Hendrickson, et al., 2008). Successful project management means the project accomplishes its design, maintains its schedule, and remains in its budget (Kim, et al., 2004b). Compared with other products, the construction project is quite different because each building is unique. Despite the identical design of buildings, the construction costs still differ because of the inherent uncertainties in the construction industry (Xiao et al., 2018). Therefore, construction cost estimation must be conducted for each individual project and the process may differ due to the different project situations.

Cost estimating is a significant task critical to successful project management (Carr & Management, 1989; Dysert & Elliott, 2002; Shin, 2015). It is deemed to provide valuable financial information for construction decision making in different stages. Various planning activities, including the bidding preparation, cost monitoring and control of projects during the construction stage, and financial performance evaluation of the project, are significantly influenced by the result of construction cost estimation (Akintoye & Fitzgerald, 2000).

Construction cost estimation also plays a significant role in the project scheduling (Žujo et al., 2017). Numerous studies have made an effort to capture the qualitative and quantitative relationships between the construction cost and duration. This has led to several significant factors being explored as the significant causes of construction duration and cost overruns, such as unsuccessful early project planning, changes and adjustment in design or scope, and late or deferred payments (Famiyeh et al., 2017; Larsen et al., 2015a; Mulla & Waghmare, 2015; Shehu et al., 2014).

The accuracy of construction cost estimation differs due to the different project stages. The cost estimates during the project planning phase can be adjusted because project stakeholders and investigators may revise the project before they are willing to

invest in the project (Yaman & Tas, 2007). After the project is well defined and with the release of information in the project design, the construction cost estimation is updated with more precision. Similarly, accurate pre-tender cost estimation helps to improve the design processe as more information is released gradually (Li et al., 2005; Skitmore, 1987, 1990). In projects in which detailed information has been clearly defined and the schedule is well planned, the range of construction cost estimation can reach ±10 percent (Stackpole, 2010). However, sometimes the high degree of complexity will increase the range of ECCE to ±50 percent (Stackpole, 2010).

2.1.2 Accuracy

Inaccurate early estimation will result in the failure of project planning, monitoring and management (Oberlender & Trost, 2001b). Both underestimation and overestimation have negative influences on the overall project performance (Swei et al., 2017). Overestimation of construction costs may result in the owner's cancellation or termination of the project or the inefficient investment of money, while underestimating may lead to time delay and the reduced quality of projects (Larsen, et al., 2015b). Large underestimation of project cost may even result in the bankruptcy of developers or contractors.

Construction cost overrun is a common phenomenon that happens all over the world (Flyvbjerg et al., 2003; Rui et al., 2012). Several studies have explored the construction cost overrun for various types of buildings. Pohl and Mihaljek (1992) examine 1015 projects financed by the World Bank and find that 22% of projects experience cost overrun and 50% for time overrun (Pohl & Mihaljek, 1992). In one study conducted by Merrow (1998), nearly nine out of ten megaprojects have the overrun of 88% averagely. Project size is identified as the most significant factor influencing the cost overrun and the large projects intend to experience greater cost overrun (Merrow et al., 1988). Flyvbjerg et al. (2003) examine the cost performance by using 258 transport infrastructure projects in 20 nations. The results show the average cost overrun of rail projects is 45%, while for tunnels and bridges it is 34% and for roads 20% (Flyvbjerg, et al., 2003). Bordat et al. (2004) analysed the overall cost overrun rate of transportation projects in Indiana and found that more than half of projects financed by the Indiana Department of Transportation (INDOT) had cost overruns from 1996 to 2001 (Bordat et al., 2004). As for mining projects, almost one out of thirteen projects have cost overruns. Bertisen and Davis (2008) reviewed more

than 60 international mining projects and the results showed the average final cost was 14% higher than as estimated in the early stage. In their study, projects in small sizes had more estimation inaccuracy than those with large size (Bertisen & Davis, 2008b).

2.1.3 Influencing Factors

Several researchers explore the influencing factors on cost overrun from different perspectives. Project features such as project size are initially deemed as potential causes of construction cost overrun. In a study that involved 56 projects and 102 questionnaires, the result shows that project size is the most significant influencing factor for the accuracy of construction project estimation. Large projects are found to experience more cost overruns (Jahren et al., 1990). However, this result is contradictory with another study that cost overruns intend to be more predominant among smaller projects than the larger ones in road construction projects (Odeck, 2004). Projects with small size suffered from more biased estimation than those with large size.

Project type is another factor that is deemed to have an influence on construction cost estimation. Empirical evidence has been found that the public-private partnership (PPP) project has superiority with respect to construction cost and time performance when compared with traditional procurement in Australia (Raisbeck et al., 2010). For PPP projects, the influencing factors of cost overrun include the project cost, duration, length, specific maintenance and rehabilitation activities (Anastasopoulos et al., 2014). In a study using 122 projects including road, building, and drainage projects, the results show that cost overruns differ due to the project type. The cost overruns of building projects increased with contract amount while the cost overruns for drainage projects decreased with increasing contract prices (Senouci et al., 2016). Other factors including technical, psychological and political-economic factors are also deemed to have a potential effect on cost overruns (Flyvbjerg, 2007).

Some studies have found that construction overrun and time overrun frequently occur together (Larsen, et al., 2015b). Thus, numerous studies attempt to understand the construction cost by exploring the quantitative relationship between construction duration and cost. Some studies use a linear model (Fulkerson, 1961), while other studies exploit non-linear methods such as the quadratic (Deckro et al., 1995) and exponential (Žujo, et al., 2017). There is a representative model that captures the relationship between cost and duration in Australia well (Bromilow, 1969). It is widely

used in different countries (Chan, 1999; Kaka & Price, 1991; Ogunsemi & Jagboro, 2006; Thomas Ng et al., 2001; Yeong, 1994). Another potential reason for cost overrun is information asymmetries (Pindyck & Rubinfeld, 1998). The lack of practical knowledge and insufficient time to prepare documents are also deemed as main reasons (Akintoye & Fitzgerald, 2000).

The recent studies mainly focus on the project management factors of cost overrun and time overruns (Adam et al., 2017). Some studies deem that the project management team has the incentive to underestimate cost (Bertisen & Davis, 2008a; Flyvbjerg, 2007). Larsen et al.(2015a) find the most significant influential factor for cost overrun is the mistakes in consultant material, while insufficient project budgeting is the main reason for the time delay in public construction projects (Larsen, et al., 2015b). The underlying reason is that planners and promoters intend to deliberately misrepresent costs to increase the likelihood to win the competition and get the funding (Flyvbjerg, 2007).

2.1.4 Significance of ECCE

Based on the availability of the information and accuracy, cost estimation can be classified into three categories: (1) Order of magnitude estimation – in this stage, only minimum project information is available, the accuracy of cost estimation is relatively low; (2) Conceptual estimation or early stage estimation – in this stage, primitive project information is available and with the release of information on project design, the construction cost estimation accuracy is updated with more precision; (3) Detailed estimation – in this stage, the project design and specification have been completed and the highest level of accuracy can be reached (Lai & Lee, 2006). ECCE usually refers to the first and second stage when the basic project design information and technologies for the design are known.

ECCE is a significant task of construction project management and plays a vital role in projects' ultimate success (Cheng, et al., 2009). Several researchers have emphasised the significance of ECCE from different perspectives (Balali et al., 2018; Canesi & Marella, 2017; Dysert & Elliott, 2002; Yu & Skibniewski, 2009). Firstly, ECCE has a great influence on early-stage decision making. It is an important project management activity in achieving a desired economic benefit (Ji, et al., 2010a). The accurate ECCE helps the managers to choose adequate alternatives and to avoid misjudging solutions (Arafa & Alqedra, 2011a). Various decisions with respect to the

main structural systems, major construction methods, and most construction materials need to be made at early stage (Yu & Skibniewski, 2009). The flexibility to adjust the project scope, design, specification, and standards at the early stage is very high. For helping project managers make valuable budgeting decisions, ECCE usually needs to be conducted as early as possible (Carr & Management, 1989; Sonmez, 2011).

Secondly, ECCE is important for successful project management (Dysert & Elliott, 2002; Shin, 2015). Various management activities including team management, scheduling, and financial evaluation are significantly influenced by the result of ECCE (Akintoye & Fitzgerald, 2000; Cheng, et al., 2009; Larsen, et al., 2015b). The comparison between the estimated cost and the actual cost of the mile-stone point of the projects provides the baseline and guidance for the overall construction management (Motwani et al., 1995).

Thirdly, ECCE is critical for all the participants in the construction process. The influence of construction cost estimation may differ due to different perceptions from stakeholders (Hampton et al., 2012). For the owners and investigators of the project, the inaccurate early cost estimation will cause the wrong decision in the project planning, which results in insufficient financial return and failure of optimizing the capital budget. For the contractors, the bidding decision is made based on the result of ECCE and the inaccurate early construction cost may cause serious business failure (Wu et al., 2018).

2.1.5 Challenges in ECCE

Information limitation at the early stage is the widely accepted reason for the difficulties encountered in preliminary cost estimation (Arafa & Alqedra, 2011b; Cheung et al., 2012; Haavisto, 2015; Hong et al., 2011; Jin, Han, Hyun, & Kim, 2014). The level of project scope is quite coarse at the early stage. For the order of magnitude estimation, it is usually made even before the facility is designed. Besides, the owners and contractors may adjust the project scope, design, specification, and standards in the early stage. This results in the high flexibility of ECCE, increasing its difficulty. The accuracy of ECCE is usually low because the project is not always well-defined and the estimates extend over a very long time period.

Time constraints are deemed as another challenging factor for ECCE (Akintoye & Fitzgerald, 2000; Arafa & Alqedra, 2011a; Petroutsatou et al., 2011). ECCE is

typically prepared under stiff time constraints. Construction clients are usually keen to know the budget because the ECCE provides the guidance for the owners or the investigators to choose adequate alternatives (Cheung & Skitmore, 2006c). However, not all the companies and agencies can afford to overinvest their attention and effort in projects at the early stage. When facing limited time and resources, an estimator's effort can be more valuable in the later design estimating stages (Gardner Brendon et al., 2016). All the investment in ECCE could then be considered worthless when the project is evaluated as unfeasible for further development after a cost-benefit analysis. For example, less than one-fifth of structural steel buildings that reach the early stage are ever constructed (Moselhi & Siqueira, 1998).

Given subjective analysis has limitations such as human mistakes and changing results caused by different cost estimators, data-based ECCE is deemed to largely reduce the subjective influence of human error, as well as improve the efficiency and consistency of the result (Adeli & Wu, 1998b). However, the availability of historical project cost may be limited due to confidentiality issues (Hegazy & Ayed, 1998). Public information databases are costly to access and sometimes lacking relevancy to the target project, and many companies are unable or unwilling to invest time and effort in data collection (Gardner, Gransberg, & Jeong, 2016). Thus construction cost estimates are largely based on the experience of cost estimators (Cheung & Skitmore, 2006a, 2006b). Several studies emphasise the importance of expertise in achieving accurate cost estimation in the early stage. In a series of experiments to measure early stage estimation abilities of quantity surveyors, the experienced quantity surveyors give more accurate estimates (Skitmore, 1985). However, subjective analysis has limitations such as human errors and varying results based on the proficiency of the estimator (Adeli & Wu, 1998b). The reliability of the ECCE based on experts are questioned.

The accuracy of construction cost estimation improved due to the progress of the project design and construction (Gardner Brendon, et al., 2016). The accuracy of early construction cost could only reach ±25% of the final cost (Burke, 2013; Petroutsatou, et al., 2011), and sometimes even ranges from–30% to +50% of actual project cost (PMI 2008).

2.1.6 ECCE methods

There are various classifications of construction cost estimation methods due to the different standards. In terms of modelling purposes, models can be classified into two types: deterministic models and stochastic methods (Smith & Mason, 1997). Single rate methods are the simplest deterministic method in ECCE. Unit estimating method, floor area estimating method and cube estimating method are considered as most conventional single rate models methods (Cheung & Skitmore, 2006c; Dang et al., 2018; Skitmore & Marston, 2005). Another single rate method is the obsolete cube method by using the volume of the building as the single variable. James developed the Storey Enclosure Model (SEM) by using a total weighted enclosure area or storey enclosure area (Cheung & Skitmore, 2006c). However, SEM is seldom used in practice because of lacking confidence in the arbitrarily prescribed weightings (Ashworth & Perera, 2015; Fortune & Lees, 1989). When there are sufficient historical data available on a particular type of project, it is appropriate to estimate the new project by using the data from a previous similar project. Single rate method is only limited to certain types of building and/or to early stages of the project when the available project information is not enough (Akinsiku et al., 2011; Cheung & Skitmore, 2006b).

With the rapid development of data science, various estimation models based on data analysis have been proposed to improve the accuracy of estimating early construction costs (Martin Skitmore & Thomas Ng, 2003). Numerous studies have focused on cost prediction in the initial phase of a project using various data-based techniques, including statistical techniques (Lowe, Emsley, & Harding, 2006; Phaobunjong, 2002; Trost & Oberlender, 2003c); probabilistic analysis techniques and distribution analysis (Barraza et al., 2000; Nassar et al., 2005); Monte Carlo Simulation (MCS) (Juszczyk, 2017; Wing Chau, 1995); and artificial intelligence (AI) techniques such as ANN (Bode, 1998; Juszczyk, 2017; Kim et al., 2004a). Given subjective analysis has limitations such as human mistakes and changing results caused by different cost estimators, data-based ECCE is deemed to largely reduce the subjective influence of human error, as well as improve the efficiency and consistency of the result (Adeli & Wu, 1998b).

There are a large number of regression models that have been developed for ECCE. These approaches focus on the designing a model that suits their own unique data (Chang et al., 1997; Hwang & Management, 2009; Jafarzadeh, Ingham, Walsh,

et al., 2014; Lowe, Emsley, & Harding, 2006; Lowe, Emsley, Harding, et al., 2006; Rui et al., 2011; Sonmez, 2008; Sonmez, 2004; Trost & Oberlender, 2003b). Chang et al. (1997) proposed a fuzzy regression method for construction cost estimation of water plants (Chang, et al., 1997). Trost and Oberlender (2003) combine factor analysis and multivariate regression together for predicting the accuracy of ECCE (Trost & Oberlender, 2003b). Shin (2015) uses the boosting regression trees for estimating the preliminary cost of building projects (Shin, 2015).

Different from deterministic estimating methods, the probabilistic method could provide an uncertainty estimate (Skitmore, 2001). Several studies attempt to use different approaches such as MCS and uncertainty analysis under correlation to address the uncertainty in construction cost estimation (Briggs et al., 1999; Ökmen et al., 2010; Touran & Lopez, 2006; Touran & Management, 1993). However, the probabilistic models have limitations as the simulating process is time-consuming (Dang & Le-Hoai, 2018). The underlying assumption that probabilistic simulation is the independence between variables may be invalid in practice (Irfan et al., 2011).

To find the most widely used techniques and provide a comprehensive literature review on ECCE, this study searched the Scopus database using the keywords related to ECCE ((conceptual cost estimat* or preliminary cost estimat* or early cost estimat* or predesign cost estimat*) and ("construction" or "building" or "project" or "infrastructure")). The wildcard character * is used to capture variations of a word, such as estimation and estimating. The result indicates more than 288 publications were found, including journal articles, conference papers, business articles, book chapters, books. Figure 2.1 indicates the research trend from 1977 to 2019.

Figure 2.2 shows the percentage of the ECCE publications in different research fields. The majority of the publications closely related to the engineering discipline and computer science are found in the second largest research field. Other research fields including business, management and accounting, energy, environmental science, earth and planetary science also play a significant role in ECCE.

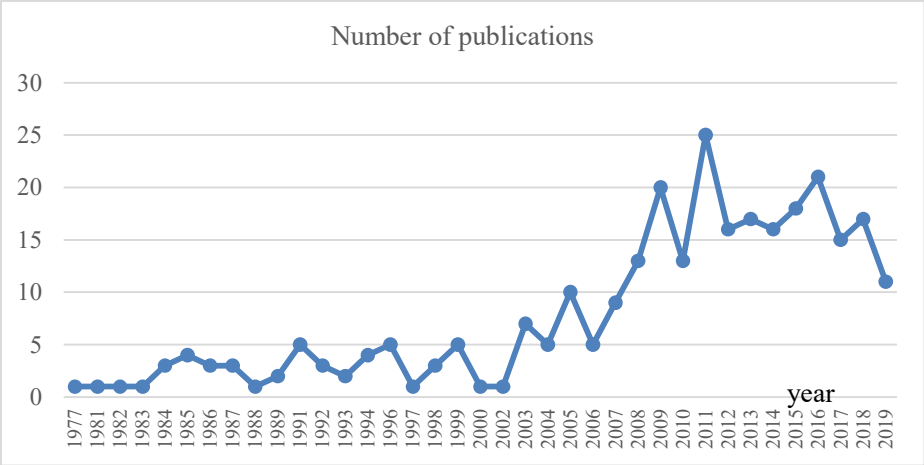


Figure 2.1 Annual publications relating to ECCE

Among all the publications related to ECCE, the three most popular methods are identified as regression, neural network and CBR (Chau, 2018; Kim, et al., 2004b). Therefore, the following sections will discuss the application of these three methods in ECCE separately.

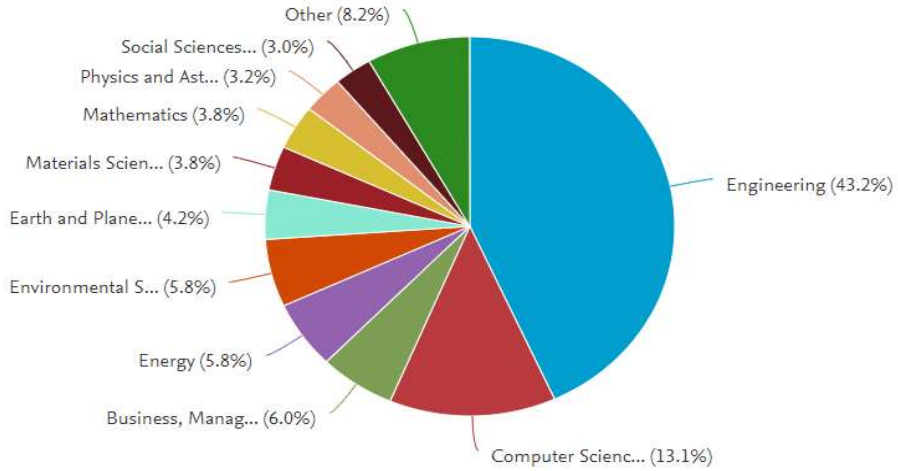


Figure 2.2 The research area classification

Regression

A regression model is defined as the formal means of expressing the two essential ingredients of a statistical relation and it serves three main purposes, including description, control and prediction (Neter et al., 1989). It has a solid mathematical foundation and thus it has the most application in the construction industry (Wilmot et al., 2005).

Regression has been developed since the 1970s (Bowen & Edwards, 1985; Khosrowshahi et al., 1996; Kim, et al., 2004b; Kouskoulas & Koehn, 1974; Trost & Oberlender, 2003a). It is a very powerful statistical approach for analysing and predicting the result based on the historical data (Skitmore & Ng, 2003). Various studies have used regression analysis for various types of buildings. Lowe et al. (2006) use the regression model to identify the influence of the building feature. The results show that gross internal floor area, function, duration, mechanical installations, and piling are the most significant factors for building cost (Lowe, Emsley, Harding, et al., 2006). Fragkakis et al. (2011) develop a model for estimating the predesign cost of bridge foundations. The whole model includes the foundation system selection, the material quantities estimation and the foundation cost estimation. Rui et al., (2011) develop five regression models to estimate the pipeline construction component cost and explore the influencing factor for different types of pipelines in different regions. Jafarzadeh et al., (2014) establish a statistical model for estimating seismic retrofit net construction cost of confined masonry buildings. The best predictors include total floor area, seismic weight indicator, floor and roof diaphragm type, and mortar quality.

Regression is also used to evaluate the cost estimation accuracy (Oberlender & Trost, 2001a). Oberlender and Trost (2001) identify four influencing factors of construction cost estimation performance and develop an estimate scoring system to evaluate how well the construction cost can be estimated (Oberlender & Trost, 2001b). Trost and Oberlender (2003) further conduct a study to help estimators and project managers have an objective assessment of the early estimates. By using 45 potential drivers, basic process design, team experience and cost information, time allowed for preparing the estimate, site requirements, and bidding and labor climate are identified as the five most significant factors (Trost & Oberlender, 2003b).

Furthermore, several robust regression methods have been developed to achieve high efficiency in estimating (Yu et al., 2017). Compared with the commonly used regression method, robust regressions have more advantages for handing the noisy data and outliers. The advances in the regression research area also benefit the application of regression in ECCE. To address the unusual y observations, M-estimates are designed in the normal equation with appropriate weight functions. M-estimates are limited for being sensitive to high leverage points on x (Huber, 2011). R-estimates attempt to minimize the sum of scores of the ranked residuals that have relatively high

efficiency but are limited in the low tolerance of breakdown points (Jaeckel, 1972). Several squares estimates are proposed including Least Median of Squares estimates and Least Trimmed Squares estimates to minimize the median or the trimmed sum of squared residuals (Rousseeuw, 1985; Siegel, 1982). To have a high breakdown point, S-estimates are introduced to minimize the variance of the residuals by decreasing the efficiency (Rousseeuw & Yohai, 1984). Later, Generalized S-estimates are developed and have a higher efficiency (Croux et al., 1994). MM-estimates, Mallows Generalized M-estimates and Schweppe Generalized M-estimates can be seen as variations of M-estimates. MM-estimates have the advantage of simultaneously attaining a high breakdown point and efficiencies (Yohai, 1987), while Mallows Generalized M-estimates and Schweppe Generalized M-estimates are limited because of their impaired efficiencies (Mallows, 1975) (Handschin et al., 1975). To address these limitations, Schweppe one-step Generalized M-estimates is proposed and can be calculated in one step (Coakley & Hettmansperger, 1993). An overall comparison of some of the mentioned robust methods has been conducted to show their strengths and weakness (Wilcox, 1996; You, 1999).

Continuous development in robust regression has resulted in a number of recent studies. Robust and efficient weighted least-square estimator is introduced by Gervini and Yohai (2002) and it attempts to achieve a high breakdown point and high efficiency (Gervini & Yohai, 2002). Another robust approach is developed based on the regularization of case-specific parameters for each response and it deems M-estimator as its special case (Lee et al., 2012; She & Owen, 2011). Later, a new modal linear regression (MODLR) based on the conditional mode of the response Y given the set of predictors x was proposed (Yao & Li, 2014). The mode takes the value with the highest probability of occurrence and thus provides a shorter prediction interval than other linear regressions (Yao & Li, 2014). When the data is not distributed as assumed, modal regression works quite robustly. MODLR is deemed to produce shorter predictive intervals than mean linear regression, median linear regression, and MM-estimators. It has received much research attention and is widely used in machine learning (Damir et al., 2007; Wang et al., 2017).

Despite its advantage of having a well-defined mathematical basis (Kim, et al., 2004b), regression is limited in ECCE because it cannot deal with a large amount of data (Chau, 2018). Regression analysis is designed to deal with the data containing

less than 1,000 samples and fewer than 50 variables (Neter, et al., 1989). MRA is also criticized because the data in the real world do not necessarily comply with its parametric assumptions (Adeli & Wu, 1998a; Bode, 2000; Yau & Yang, 1998). Besides, regression requires users to choose the relation between the predictors and responsors (e.g., linear, quadratics) (Sonmez, 2004). However, regression has an advantage over other models since it uses fewer parameters when the relations among the variables can be presented adequately (Sonmez, 2011).

ANN

ANN is another popular method that has drawn much attention in recent years in ECCE. The fundamental idea behind ECCE is to model the response as a nonlinear function of various linear combinations of predictors (Neter, et al., 1989). Several ANN models have been widely developed for ECCE for various types of constructions including building projects, highway projects, road tunnels, public school buildings. (Hegazy et al., 1998; Juszczyk et al., 2018; Kim, Choi, et al., 2005; Li et al., 1999; Peško et al., 2017; Wilmot, et al., 2005).

Residential building has been mostly studied in the ECCE ANN model. Günaydın and Doğan (2004) develop an ECCE ANN model for structural systems of buildings for both designers and project managers. By using cost and design data from thirty projects, this study indicates the applicability of ANN in structural systems of buildings and the average estimated accuracy reaches 93% (Günaydın & Doğan, 2004). Kim et al. (2004b, 2005a) use an ANN model to estimate the construction cost of the residential building using 12 predictor variables. By combining genetic algorithms (GA) and backpropagation, their study aims to optimize the process of determining the model's parameters. Kim et al., (2005) combine ANN and GA to estimate the early cost of residential building in Seoul, Korea (Kim, Seo, et al., 2005). Cheng et al. (2009) design an Evolutionary Fuzzy Neural Inference Model to improve the building cost estimation accuracy. Their study combines the advantages of GA, fuzzy logic and ANN together, to obtain the optimal solution (Cheng, et al., 2009).

Other types of projects include highway projects, road tunnels, public school buildings. Petroutsatou et al. (2012) establish an ANN model for ECCE of road tunnels. In this study, several factors including geological factors, geometrical factors and work quantities-related factors are determined and two types of neural networks, including the multilayer feed-forward network and the general regression neural network, are

compared (Petroutsatou, et al., 2011). Jafarzadeh et al. (2014) develop an advanced ANN model for estimating retrofit net construction costs by using 158 earthquake-prone public school buildings. Several components are defined and explored in the model including the number of hidden layers and neurons, learning rate and momentum by using sensitivity analysis (Jafarzadeh, Ingham, Wilkinson, et al., 2014). Hyari et al. (2016) attempt to estimate the cost of the engineering service of public construction projects by using the ANN model. This study predicts the percentage of engineering services in construction according to project type, engineering services category, project location, and project scope by using the data from the Governmental Tenders Department in Jordan (Hyari et al., 2015).

However, ANNs have limitations in explaining the results of the model, as well as in choosing the optimal parameters. It is a black box technique, in which several of the trial-and-error processes are involved. The model processing is very time-consuming, which makes the model less convincing (Chau, 2018; Creese & Li, 1995; Hegazy & MOSELHI, 1994; Li, 1995). Petroutsatou et al. (2012) explore the influence of the amount of data on the performance of the ANN model (Petroutsatou, et al., 2011). Besides, ANNs accept only numerical input attributes and this may result in the loss of information during the data transformation process (Pal et al., 2018).

In summary, ANN has the advantage of having the capacity to accommodate complex nonlinear behaviour and an ability to learn from numerical examples, but is criticized for being an impenetrable black box (Graves & Pedrycz, 2009) and only admitting numerical input data (Arditi; & Tokdemir, 1999) as well as being time consuming for obtaining optimal model networks (Sangyong & Jae Heon, 2014).

CBR

CBR is quite popular in the research field of ECCE as it provides an effective and comprehensive procedure for estimating project cost (Hu et al.). ECCE traditionally relies extensively on the previous knowledge and experience of practitioners, making CBR a perfect match for ECCE (An, et al., 2007). CBR is widely used for various types of projects including residential buildings, road, bridges, river facilities, military facilities, pump stations, craft and tunnels (Kim, et al., 2004b; Kim & Kim, 2010b; Lee et al., 2013a). Table 2.1 summarized the building types in the current ECCE CBR models.

Table 2.1 Distribution of project types in CBR cost estimation

Project type	Building features	No. of studies
Building	Not mentioned	6
	Multiple-family housing	3
	Apartment	2
	Residential building	2
	High-rise building	1
	Building structure system	1
	Large building	1
	Historical building	1
	Total	*17*
Road	Pavement maintenance project	2
	Highway road	1
	Public road	1
	Total	*4*
Bridge	Not mentioned	3
	Railway bridge	1
	Total	*4*
River facility	Not mentioned	2
Military facility	Military barrack	1
	Not mentioned	1
	Total	*2*
Pump station projects	Not mentioned	1
Craft	Not mentioned	1
Tunnel projects	Not mentioned	1
Construction projects	Not mentioned	3

Some studies compare the regression analysis, ANN and CBR. Kim et al., concludes that both CBR and ANN models are appropriate for estimating construction costs (Kim, et al., 2004b). Another study found CBR is more effective than ANN in ECCE for apartment building. The average mean error of CBR is only half of that in the ANN model (Kim, et al., 2005).

The majority of the studies focus on building and optimization of the case retrieval process. Doğan et al. compare the three weight determination methods including feature counting, gradient descent, and genetic algorithms by using 29 residential building projects. The results show that GA performed better than other

methods (Doğan et al., 2006a). Similarly, An et al. compare three weight determination methods including the analytic hierarchy process (AHP), feature counting and gradient descent. The result shows that AHP produces more accurate and explanatory attribute weights than the other models (An, et al., 2007). However, this weight determination is limited due to its reliance on questionnaires and its incompetency in updating with changes in the case-base. Doğan compares the three different weight determination methods based on decision-tree in CBR model. The binary-dtree method, info-top method and info-dtree method are compared by using cost data of residential building projects. The result indicates that information gain is the best weight determination methods. Their study also shows that the performance of CBR largely depends on the data associated with the parameters used in the model (Doğan et al., 2008). Kim and Kim (2010) further optimize the CBR model using the GA in weight determination and the results show that GA is superior to other conventional methods (Kim & Kim, 2010a). Koo et al. develop a hybrid CBR model for predicting construction cost and duration of multi-family projects simultaneously. It uses the GA as an optimizing method for improving the overall performance of the CBR model (Koo, Hong, Hyun, & Koo, 2010). Koo et al. improve the CBR model by introducing two optimized parameters: the minimum value for calculating attribute similarity and the range of attribute weights. By using the optimized method, their study explores the changes in the influence of project features on the owner's decision making. (Koo, Hong, Hyun, Park, et al., 2010). Ji et al. introduce the Euclidean distance concept in the case retrieval. The results show that Euclidean distance can improve the ECCE CBR's performance in terms of accuracy. This conclusion is deemed as a basis for further research on improving the CBR method (Ji, Park, & Lee, 2011).

Except for case retrieval, research focus also needs to be paid on case revision and adaptation. Hong and Hyun (2010) proposed a revision method in CBR mode to improve the accuracy of the model. When there are insufficient established cases in the case-base, the revision method can be effective in improving the model's accuracy (Ji et al., 2010c). Ji et al. introduce a case adaptation method that can help in demise and transformation of the data in case-base. Totally there are 129 military barrack projects in Korea used for training the model. The result was validated by using 13 test cases. Another 164 Korean public building projects are used for the applicability test (Ji et al., 2012b).

CBR is also used for estimating the cost of roads, bridges, military facilities and restoration costs of historical buildings. Kim uses combinations of features; criteria of similarities and retrieval ranks and applies GA to optimize the attribute weight. The verification results show that the mean absolute error rate is 11.9% and the standard deviation is 12.7% (Kim, 2011). Ji et al. apply the ECCE CBR model to military projects in Korea. Their study uses 422 construction projects at 16 military facilities and involves ten military engineers for system validation experiments. Similarly, GA is used in weight determination. Their study illustrates the effectiveness of the system in terms of estimation accuracy and user-friendliness is confirmed (Ji, Park, et al., 2011b). Chou develops a web-based CBR model for pavement maintenance. The attribute and case similarity can be calculated and displayed using browsers. Eigenvector and equal weighting methods are used in weight determination (Chou, 2009). Wang et al. proposed an ECCE CBR model for historical buildings. Two retrieval techniques including inductive indexing and nearest neighbour are used in the case retrieval process (Wang et al., 2008).

The more recent studies focus on the CBM of the CBR model and green building. To address the interrelation in the input attributes in the CBR model, Hong et al. use correlation analysis for selected cases. The attributes weight is determined by GA. Their results illustrate that CBR generates good results for early construction cost estimation (Hong, et al., 2011). Ji et al. produce a learning method for addressing the issues of insufficient data. By using the first cost-variance impact factor and data-expansion rate, their study attempted to improve the model's accuracy by generating training cases. This study shows how to generate the cases in the case-base when insufficient cases happens. It illustrates the significance of continuously updating of case-base and simplifying the updating process (Ji et al., 2018). Leśniak and Zima (2018) develop a CBR model including the sustainability factor for cost estimating. Their study considers the environmental impact of the building, materials used, and the impact of the facility on the surroundings by using 143 construction projects. (Leśniak & Zima, 2018). Tatiya et al. developed a CBR model for estimating the cost of building deconstruction, which introduces an additional option of design for deconstruction of green buildings (Tatiya et al., 2018). The next subsection carefully examines each step of the existing ECCE CBR models to provide a detailed analysis.

2.2 APPLICATION OF CBR IN ECCE

2.2.1 Overview of CBR

CBR is a problem-solving process variously described as involving addressing a current issue by recalling and reusing previous knowledge and experience (Kolodneer, 1991), solving new problems by adapting the solutions of previous cases that have been successfully solved (Riesbeck & Schank, 1989) and a process "to solve a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation" (Aamodt & Plaza, 1994).

The origin of CBR can be traced to the work of the dynamic memory model of Roger Schank and his team at Yale University in the late 1970s. Since then, CBR has developed in four main phases. The first is from 1977 to 1992, which was characterized by schema-oriented memory models (Richter & Weber, 2013) including CYRUS (Schank, 1982, 1983), MEDIATOR (Simpson, 1985), CHEF (Hammond, 1986) and HOPY (Ashley, 1991). Schema-oriented memory models go back at least to Bartlett's work on the process of remembering (Bartlett & Burt, 1933). The first German workshop in 1992 marked the beginning of the second phase (Richter & Weber, 2013), the main feature of which was the wide application of engineering techniques that enabled the more systematic development of CBR theory. The complex techniques used, decreased the time and the cost of putting CBR theory into practice. The induction and reasoning from the case-based system INRECA provides a good example in this phase (Bergmann, 2001). Extensive CBR workshops that included a European Workshop on CBR, International Conference on CBR and International Joint Conference on Artificial Intelligence Workshop, were also organized globally during this period. The third phase started in the second half of 1990, when data mining and machine learning appeared more often in CBR publications (Aha et al., 2005). After the mid-2000s, CBR research entered its current period, which is characterized by a combination of Web searching, context-aware systems and textual mining (Greene et al., 2008). Since 1995, a large number of conferences and seminars have been organized all over the world. To provide a brief understanding of its development process, Table 2.2 summarise the historical development and application of CBR from 1977 to 1995 .

Table 2.2 Historical development and application of CBR from 1977

1977	Schank and Abelson's *Scripts, Plans, Goals and Understanding: an Inquiry into Human Knowledge Structures* provides the history behind the creation of CBR (Kolodner, 1992; Shin & Han, 1999).
1982	Schank published *Dynamic Memory: A Theory of Learning in Computer and People,* which for the first time addressed the *CBR framework and theory (Schank, 1983)*.
1983	1. Roger Schank and his work team at Yale University introduced the dynamic memory framework, which has a reminding system of the earlier situations and situations in the process of problem solving and learning (Schank & Abelson, 1977). Schank's theory is the earliest prototype of CBR system.
	2. Kolodner developed the first CBR system called "CYRUS", a question-answering system with a rich knowledge of travelling and meetings of former US secretary of State, Cyrus Vance (An, et al., 2007; ChoongWan et al., 2011). It is widely recognised as the basis of the other CBR models.
	3. The first precedence reasoning framework in legal judgments was developed by Edwina Rissland and her group at the University of Massachusetts (Sangyong & Jae Heon, 2014). It is the basis of the HYPO system.
1986	The first case-based planner CHEF was developed by Hammond. CHEF creates its plans by recalling old plans that worked under similar circumstances and modifying them to fit new situation (Hammond, 1986). It is the first CBR tool to be successfully commercialized.
1988	1. The First Defence Advanced Research Projects Agency (DARPA) CBR workshop held on May 1988, Clearwater Beach, Florida, USA.
	2. The AAAI Workshop on CBR held on 21-26 August 1988 in Saint Paul, Minnesota, USA.
1989	1.The Second Defense Advanced Research Projects Agency (DARPA) CBR workshop held in 1989, Pensacola Beach, Florida, USA.
1991	1. The third Defense Advanced Research Projects Agency (DARPA) CBR workshop held in 1991, San Mateo, California, USA. The continuous DARPA CBR workshop from 1988 to 1991 formally marked the birth of the discipline of CBR.
	2. The CBR tool CBR-Express™ was developed by Inference Corp., which is a CBR environment tailored for developing help desk application.
	3. The first interpretive Case-based reasoner *HYPO* which works in the domain of law was developed by Ashley (Ng, 1996)
	4. The CBR was introduced in China and Japan. Chinese academic Zhizhong Shi introduced the memory network model and case retrieval method. Japanese Academic Kobayashi, Shigenobu and Nabeta, Shigeko discussed the problems for the CBR.
1992	1. The first German CBR workshop was organised. (Riesbeck & Schank, 1989).
	2. The CBR tool ReMind™ (Cognitive System Inc.) was developed with the support of the US Defense Advanced Research Projects Agency. The ReMind™ formally marked the development of CBR shifting from cognitive science to artificial intelligence.

1993	1. The first British CBR seminar *UK IEE Colloquium on CBR* held in London, 12 Feb 1993.
	2. The first *European Workshop on CBR* (EWCBR-93) held in Kaiserslautern, Germany. There were around 120 participants and more than 80 related CBR papers on scientific and application-oriented research.
	3. The CBR system of internal combustion engine oil product design system EOFDS was developed in China. It is the first CBR system developed in China.
1994	1. *The second European Workshop on CBR* (EWCBR-94) was held in Chantilly, France.
	2. Aamodt and Plaza published an article where CBR cycle was systematically introduced and four process in CBR cycle were defined (Aamodt & Plaza, 1994).
1995	1. The first International Conference on CBR (ICCBR) held in Sesimbra, Portugal, which illustrates the interest in CBR growing internationally.
	2. The first United Kingdom CBR Workshop on Industrial Applications of CBR was organised, which marks the attention of the industrial practice of CBR.
	3. International Joint Conference on Artificial Intelligence Workshop (IJCAI-95) held in Montréal, Canada.
	4. International Conference on Knowledge Discovery and Data Mining held in Montreal, Quebec, Canada.

2.2.2 Advantages of CBR in ECCE

ECCE is usually conducted before the construction is well designed, and thus must depend on the experience of cost estimators (Cheung & Skitmore, 2006a, 2006b). There are several studies that emphasise the importance of expertise in achieving accurate cost estimation in the early stage. In a series of experiments to measure early stage estimating abilities of quantity surveyors, the experienced quantity surveyors give more accurate estimates (Skitmore, 1985). The main reasons that experienced estimators give more accurate cost estimates are their capability of recalling previous projects and adjusting to new requirements (Skitmore, 1985, 1988). However, subjective analysis has limitations such as human errors and varying results based on the proficiency of the estimator (Adeli & Wu, 1998b).

The historical database containing previous construction projects is another option for preparing cost estimates. Many researchers have studied cost prediction in the initial phase of a project using various data-based techniques, including statistical analysis techniques such as MRA (Lowe, Emsley, & Harding, 2006; Phaobunjong, 2002; Trost & Oberlender, 2003c), probabilistic analysis techniques including

distribution analysis (Barraza, et al., 2000; Nassar, et al., 2005), Monte Carlo Simulation (Juszczyk, 2017; Wing Chau, 1995), and artificial intelligence techniques (Bode, 1998; Juszczyk, 2017; Kim, et al., 2004a). Compared with subjective analysis based on different cost estimators, data-based ECCE is deemed to largely reduce the subjective influence of human error, as well as improve the efficiency and consistency of the result (Adeli & Wu, 1998b). To prepare a good cost estimation, historical data should be organized in a consistent and compatible format with further applications. Any changes in the historical data may have substantial impacts on the predicted results of the new project. Therefore, historical cost data must be used cautiously and maintained continuously for ECCE.

CBR is quite popular in the research field of ECCE because of its capability of reusing the knowledge in historical cost data for a current case (Hu, et al.). Though CBR is a data-based method, it can solve a problem by reusing previous knowledge in case-base. It is a relatively convenient and efficient AI-based approach and it is deemed to have more advantages than MRA and ANN (Kim & Shim, 2013b). Besides, CBR is deemed as superior and powerful for long-term use and easy for practice (An, et al., 2007; Chen & Burrell, 2001). In Kim et al.'s (2005) analysis of 540 apartment buildings in South Korea, CBR mean error (3.68%) is only half of ANN models (Kim, et al., 2005).

Numerous studies attempt to address ECCE issues in various project types including building (Ji, Park, & Lee, 2011; Tatiya, et al., 2018; Wang, et al., 2008), road (Choi et al., 2014), bridge (Kim, 2011), river facility (Lee et al., 2013b), military facility (Ji, Park, Lee, et al., 2011b) and tunnel projects (Petroutsatou, et al., 2011).The research attention has been extensively focused on improving the accuracy by selecting the good features or optimizing the attribute weight in the retrieval process (Ji, et al., 2010a; Kim & Hong, 2012a; Marzouk & Ahmed, 2011). Despite numerous mathematical methods having been developed for ECCE, only a few generate acceptable results and CBR is widely considered as one of them (Leśniak & Zima, 2018).

CBR is made up of several steps including case representation, case retrieval and case retainment. Case representation is the conversion of the problem into a descriptive format for the CBR cycle (Richter & Weber, 2013); Retrieval and reuse are the processes of searching for previous cases in the case-base that best match the target

problem and using them for solving the new problem; revision and adaptation are the process of adapting the retrieved solution to solve the problem of the new case; and retention is the process of updating the case-base by storing the revised solution of the new problem (Byung Soo, 2011). To provide a detailed analysis on the ECCE CBR model, the following subsections carefully examine the problem formulation, case retrieval, case reuse and revision, case retainment.

2.2.3 Problem Formulation in ECCE

Case representation

Case representation is the process of deciding what to store in a case and how to describe and organize the case contents in an appropriate structure (Aamodt & Plaza, 1994). Typically, case representation comprises identifying the problem, identifying the solution and/or identifying the outcome (Watson, 1999). In ECCE CBR modelling, the problem is how to estimate the cost of a certain project and the solution is the calculated cost derived from the historical cases. The outcome is to know whether the estimated cost successfully represents the actual cost. The outcomes in the literature differ due to the case-base data used. For example, if the historical data of a project is consistent with the target project, then there is no need to consider the outcome. Otherwise, the outcome should be considered when establishing the model. In the current literature concerning ECCE CBR modelling, the data used for the training and testing samples are all from the same case-base. The common background of ECCE CBR applications is in early-stage estimation. Many papers contain the word “early” or “design phase” or “conceptual” or “preliminary” or “planning” in the title (Byung Soo, 2011; Choi et al., 2013; Chou, 2009; Dogan et al., 2006; Dogan, Arditi, & Guenaydin, 2008; Hong, et al., 2011; Jin et al., 2012; Jin, Han, Hyun, & Kim, 2014; Kim, Seo, et al., 2012; Kim & Kim, 2010b; Kim & Shim, 2013a; Koo, Hong, Hyun, Park, et al., 2010).

Case indexing

Case indexing is the process of assigning characterizing attributes to the cases (Shin & Han, 1999) and is important for modelling a successful CBR system (An, et al., 2007; Dikmen et al., 2007). The cases can be indexed by a prefixed or open vocabulary, and within a flat or hierarchical index structure (Aamodt & Plaza, 1994). Although there is no consensus as to what information should be included in a case, there are two basic principles involved when conducting case indexing: the

functionality and the ease of acquisition of the information of the case index (Kolodner, 2014).

The case memory in the ECCE CBR model generally is a flat structure of historical project information relevant to construction cost (Jin, et al., 2012). In ECCE CBR models, index selection has a significant influence on recalling the case with the highest similarity score (KARANCI, 2010). Conventional index selection methods include literature review, questionnaire, expert subjective judgement and statistical analysis. Literature reviews provide a comprehensive overview of existing case indices and an extensive research background offers a reliable pool of case indices (Dikmen, et al., 2007). However, the problem with case indexing by literature review is the dilemma caused by having too many cost impact factors (Hong, et al., 2011). In practice, therefore, the literature review needs to be combined with other methods (Chou, 2009; Dikmen, et al., 2007; Kim & Kim, 2010b).

Case storage

Case storage refers to how cases are collected and stored. Case can be stored in the form of experiences, or a series of similar cases may form a generalized case (Aamodt & Plaza, 1994). It involves a process of establishing the structure of the case-base in preparation for the next step in the CBR cycle as the availability of historical data is important for developing a successful model because of its intimate relationship with the CBR cycle (Kim, Lee, et al., 2012). There are two things that need to be considered for case storage; its structure and the number case to be stored in the case-base. The case storage structure should be consistent with the case representation and case indexing because it further reflects the idea of what is represented in the case and indices that characterize the case (Watson, 1999). In ECCE CBR models, the case is stored in a flat structure, which is similar to the case representation. In the general ECCE CBR situation, it is thought that the greater the number of cases, the more accurately the attributes and their weights can be determined (Ji, et al., 2010a). However, this is complicated by case storage also being critical in determining the relevance of the case-base in providing the data needed to find similar cases to the target case. A case-base containing too few relevant cases causes the ECCE CBR model to yield insufficiently accurate results (Dogan, et al., 2006; Dogan, Arditi, & Guenaydin, 2008; Lee, et al., 2013a). On the other side, a case-base containing too many cases will impair the efficiency of the CBR system. This begs the research

question of case-base's size. There is no consensus in the CBR community of the number of cases stored in the case-base. Some CBR models assume a large amount of cases while others are based on a more limited set of typical cases (Aamodt & Plaza, 1994). As summarized in Figure 2.3, the various case storage sizes in the reported studies range from 9 to 786, with more than half being below 200 and only 5 out of 33 above 500.

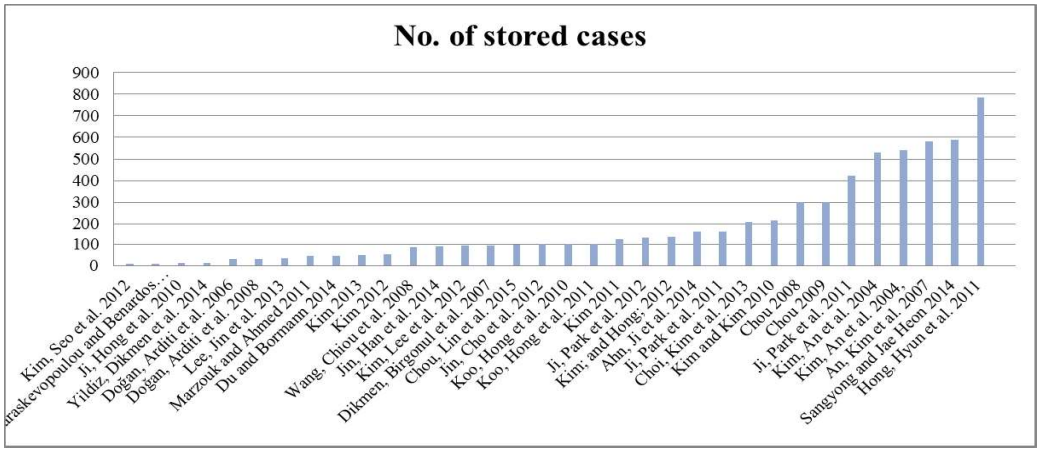


Figure 2.3 Number of stored cases in the case-base in precious studies

2.2.4 Case Retrieval in ECCE

Case retrieval refers to the process of determining the best matching previous solutions for the current problem (Aamodt & Plaza, 1994). The major focus of ECCE CBR is to optimize case similarity to identify one most similar case, while some studies use several criteria as a filter to search for more than one case (ChoongWan, et al., 2011; Kim, 2013). Even though there are different methods available, case similarity is mainly determined by attribute similarity and attribute weight (ChoongWan, et al., 2011).

Attribute similarity

Attribute similarity is the similar degree between the attributes (ChoongWan et al., 2010). The method to calculate the attribute similarity differs due to the type of attributes. The attribute can be classified as either numerical or nominal (Ji, Park, et al., 2011a). For attributes measured on the nominal scale, the most widely used approach is one of 'true/false', where 1=true and 0=false (Choong-Wan et al., 2010). Further, Ji et al. (2011) introduce a type index, which transforms nominal data to a

numerical scale and Kim et al. (2012) propose a similarity score method where a qualitative attribute is determined by its agreement with a character string. As Table 2.3 indicates, numerical attributes have various scores of similarity. Eqn 1 in Table 2.3 is the most widely used attribute similarity function. Eqns 1 and 2 are based on the relative minimum and maximum between the attribute values for a given problem case and those of the corresponding retrieved case (ChoongWan, et al., 2010; Sangyong & Jae Heon, 2014). These two methods are criticized because the calculation of attributes using these methods always produces a value of 1 or less and only highlights whether two attributes are different without being able to determine the extent of the difference (Sangyong & Jae Heon, 2014). They can only be used for selecting similar cases and have limitations in the revision phase. Thus, Eqn 7 is offered as an improvement, as it not only expresses the difference between cases but also makes it possible to verify the minimum and maximum relationship of the cases (Kim, 2013; Sangyong & Jae Heon, 2014). Eqns 3 and 6 are similar, because they are valid only when the minimum criterion for attribute similarity (MCAS) is less than its score. Eqn 4 is another popular similarity scoring method defined as true (matched=1) or false (not matched=0) for each search condition (Choi, et al., 2013; Kim, Seo, et al., 2012). Eqn 5 further refines the criterion for scoring attribute similarity (AS) by setting a ± percentage range for each search condition, as it can reduce the possibility of omitting cases similar to the target case (Kim & Hong, 2012a; Kim, Lee, et al., 2012).

Table 2.3 Attribute similarity measures

Eqn	Function	Frequency
1	$\mathrm{AS} = \frac{Min(AV_{new-case}, AV_{retrieved-case})}{Max(AV_{new-case}, AV_{retrieved-case})}$	9
2	$\mathrm{AS} = \begin{cases} 100 - \left(\frac{Min\lvert AV_{new-case}\rvert, \lvert AV_{retrieved-case}\rvert}{Max\lvert AV_{new-case}\rvert, \lvert AV_{retrieved-case}\rvert}\right) & if\ AS \gg \mathrm{MCAS} \\ 0 & if\ AS < \mathrm{MCAS} \end{cases}$	1
3	$\mathrm{AS} = \begin{cases} 100 - \left(\frac{\lvert AV_{new-case} - AV_{retrieved-case}\rvert}{AV_{new-case}}\right) & if\ AS \gg \mathrm{MCAS} \\ 0 & if\ AS < \mathrm{MCAS} \end{cases}$	1

4	$\mathrm{AS} = \begin{cases} 100, & if \ \frac{\lvert AV_{new-case} - AV_{retrieved-case} \rvert}{AV_{new-case}} \times 100 \ll x\,\% \\ 0, & if \frac{\lvert AV_{new-case} - AV_{retrieved-case} \rvert}{AV_{new-case}} \times 100 > x\,\% \end{cases}$	5
5	$\mathrm{AS} = \begin{cases} 100, if \ \frac{\lvert AV_{new-case} - AV_{retrieved-case} \rvert}{AV_{new-case}} \times 100 \ll y\,\% \\ a, y < if \ \frac{\lvert AV_{new-case} - AV_{retrieved-case} \rvert}{AV_{new-case}} \times 100 \ll z\,\% \\ 0, if \ \frac{\lvert AV_{new-case} - AV_{retrieved-case} \rvert}{AV_{new-case}} \times 100 > z\,\% \end{cases}$	2
6	$\mathrm{AS} = \begin{cases} \left(\frac{Min\lvert AV_{new-case} \rvert, \lvert AV_{retrieved-case} \rvert}{Max\lvert AV_{new-case} \rvert, \lvert AV_{retrieved-case} \rvert} \right), if \ AS \gg \mathrm{MCAS} \\ 0 \qquad if \ AS < \mathrm{MCAS} \end{cases}$	1
7	$\mathrm{AS} = \frac{1}{\lvert (AV_{new-case} - AV_{retrieved-case}) + 1 \rvert}$	1
8	$AS = \sqrt{\sum_{i=1}^{n} w_i \times \left(A_i(AV_{new-case}) - A_i(AV_{retrieved-case}) \right)^2}$	6

Where AS = Attribute Similarity, AV $_{new\text{-}case}$ = Attribute Value of new case, AV $_{retrieved\text{-}case}$ = Attribute Value of retrieved case. MCAS = minimum criterion for scoring the attribute similarity. $x\,\%, y\,\%, z\,\%$ =Different level value, $A_i(AV_{new-case})$= the value of the ith attribute of case x, w_i =the weight of the case's attributes.

Attribute weight determination

Attribute weight determination significantly influences the process of case retrieval (Changchien & Lin, 2005). As attribute weights become more accurate, the estimation accuracy of the CBR model is enhanced (Lee, et al., 2013a). The major trend of the ECCE CBR model focuses on calculating the optimal attribute weights in the retrieval phase of the CBR cycle (ChoongWan, et al., 2010; Dogan, et al., 2006; Dogan, Arditi, & Murat Gunaydin, 2008; Kim & Kim, 2010b; Kim, Lee, et al., 2012; Kim, 2012; Lee, et al., 2013a; Sangyong & Jae Heon, 2014). The methods adopted include the Genetic GA, AHP, MRA, Principle Component Analysis (PCA), Feature Counting (FC), Gradient Descent Method (GDM), Decision Trees and correlation coefficient (Table 2.4).

Table 2.4 Weight determination methods

No.	Weight determination methods	Frequency
1	GA	16

2	AHP	6
3	MRA	5
4	PCA	2
5	FC	2
6	GDM	1
7	Decision tree	1
8	Correlation coefficient	1

GA is the most widely used weight determination method because of its ability to deal with nonlinear relationships between the error rate and attribute similarity scores (Kim & Kim, 2010b). GA is superior when the size of the case-base is small and the output attribute is not binary (Dogan, et al., 2006). The most widely used target function is predictive accuracy (Du & Bormann, 2014b). The mean square error (ChoongWan, et al., 2010) and error ratio (Byung Soo, 2011) are also popular choices, as is "highest case similarity" (Dogan, et al., 2006) and "maximizing the predictive power of the CBR model" (Choi, et al., 2013). The optimal parametrics include: (1) the attribute weights (Kim & Kim, 2010b), (2) the retrieval criteria of similar cases (Choong-Wan, et al., 2010; Kim & Kim, 2010b), (3) the range of attribute weights, (4) the tolerance range of cross range between different methods (Choong-Wan, et al., 2010). Various optimal parametrics enable the ECCE CBR model with GA to be a good weight determination method while simultaneously hindering practical applications because it is uneconomic to do so (Chiu, 2002). An appropriate weight assignment method should be utilized to retrieve not only similar past cases but also other suitable cases (Ahn et al., 2014). Another problem is the inconsistency of the values of the attribute weight between the 'retrieve' phase and the 'revise' phase (Kim & Hong, 2012b). Thus, it is difficult to accurately estimate costs when the similarity between the retrieved case and the test case is low (Jin et al. 2012).

Other than GA, experts' opinions and surveys play an important role in attribute weight determination. While Kolodner (1992) suggests that experts should determine the weights, expert opinion and knowledge are difficult to measure (Kim, et al., 2004b; Xia et al., 2012). Besides, finding the right expert is not easy (Dogan, et al., 2006). Another problem is that expert intervention is a potential risk factor contributing to inaccuracy and uncertainty (Kim & Kim, 2010b; Morcous et al., 2002) and is likely to provide more convenient than reliable results. Therefore, Saaty's (2008) AHP, namely

the eigenvector method, is used together with surveys in weight determination to obtain a relatively objective result (Chou, 2008). AHP has the advantage in information being elicited by pairwise comparisons, with a series of mathematical manipulations being made to ensure that no inconsistencies exist between the different pairwise comparisons (Dogan, et al., 2006). It is used to convert experience into attribute values and has been shown to be superior compared with the use of equal weights and GDM (An, et al., 2007). This result is consistent with other research findings that AHP to be more reliable, accurate and explanatory than the gradient descent method (Kim, 2013). However, AHP, as an attribute weighting method, can be used together with questionnaire surveys (Chou, 2009).

Regression coefficients, correlation coefficients and other statistical measures are widely used due to their relatively simple computation process and ease of interpretation (Ahn, et al., 2014). Several studies confirm their superiority when compared with FC (Ji, et al., 2010a) and GDM (Kim, et al., 2005). Estimating weights using the conventional linear planning method, on the other hand, is limited because of the nonlinear relationship between the error rate and the attribute similarity scores and is incapable of capturing complex nonlinear relationships between case features and corresponding solutions (Kim & Kim, 2010b). By contrast, the feature counting method applies a weight of one to all the attributes on the understanding that there is no need to apply a higher value than one (ChoongWan, et al., 2011; Dogan, et al., 2006) to prevent bias against any factor (Dikmen, et al., 2007). Therefore, feature counting means the overall similarity of a historical project is represented simply by the number of matches (Kim, Seo, et al., 2012), which is very crude compared with other weight determination methods. The FC method is criticized because important attributes should have greater similarity (Riesbeck & Schank, 1989). An alternative is GDM - an optimization algorithm used for machine learning. Similar to GA, GDM has various functions whose objective is to maximize mode performance or to minimise the extent of similarity between the target and stored cases (An, et al., 2007). GDM is criticized, however, because it can become stuck in a local optimum (Xia, et al., 2012). It is also difficult to understand the procedures involved in determining the importance weights (An, et al., 2007). Finally, Doğan et al. introduce another approach named ID3 to determine the attribute weights. ID3 offers a number of variations including the binary-

tree method, info-top method and info-tree method (Dogan, Arditi, & Murat Gunaydin, 2008).

Case similarity

A critical issue with the ECCE CBR model is the measurement of similarity between two cases (Changchien & Lin, 2005). Several distance functions have been used in measuring case similarity. The distance calculation is a critical step in case retrieval (Madhusudan et al., 2004; Slonim & Schneider, 2001). Three distance calculation formulas, which have been widely used in existing ECCE CBR studies, and their usage frequency, are shown in Table 2.5.

The weighted sum of the attribute distance is the most popular. Its principle is to firstly determine the similarity of the target case to a case in the case-base for each case feature multiplied by a weighting factor; the weighted similarities are then summed to provide a measure of overall similarity (Wang et al., 2015). A typical algorithm for calculating nearest neighbour matching is the one used by Cognitive Systems' ReMind software reported in Kolodner (1993) and as shown in Eqn 1. Another popular distance measure is Euclidean distance (Eqn 2), calculated as the square root of the sum of the squares of the arithmetical differences between two corresponding objects (Pal & Shiu, 2004). This is the most basic algorithm for describing the relationship between two cases, with the most similar case being defined in terms of standard Euclidean distance (Mitchell, 1997). Furthermore, Mahalanobis distance is used to reduce unnecessarily influencing the covariance between variables (Du & Bormann, 2014b).

Table 2.5 Distance calculation formulas

Eqn	Distance	Formula	Frequency
1	Weighted sum of the attribute distance	$\mathrm{CS} = \frac{\sum_{i=1}^{n} w_i \times AS_i(x_a, x_b)}{\sum_{i=1}^{n} W_i}$	24
2	1-Euclidean distance	$\mathrm{CS} = 1 - \sqrt{\sum_{i=1}^{n} w_i \times (AS_i(x_a) - AS_i(x_b))^2}$	5

3	Mahalanobies distance	CS=1 $-\frac{\sqrt{(\overrightarrow{x_n}-\overrightarrow{x_l})\times W\times D^{-1}\times W^T\times(\overrightarrow{x_n}-\overrightarrow{x_l})^T}}{\max(\sqrt{(\overrightarrow{x_n}-\overrightarrow{x_l})\times W\times D^{-1}\times W^T\times(\overrightarrow{x_n}-\overrightarrow{x_l})^T})}$ Detailed information can be seen in (Du & Bormann, 2014b)	1
5	No description	/	4

where CS = Case Similarity, $AS_i(x_a)$ indicates the value of the ith attributes of case x,w_i = weight of the ith case's attributes, D is the covariance matrix, W can be calculated from W=$A \times S$; S is the TSIs indices obtained from Sobol's global sensitivity analysis; and A can be calculated from $P = X \times A$; where X is the standardized values of input variables; P is the transformed uncorrelated variables and A is a linear transformation matrix on original input data,

2.2.5 Case Reuse and Revision in ECCE

Case reuse involves dealing with the differences between the target case and those in the case-base (Aamodt & Plaza, 1994). After the retrieval phase, the CBR system should adapt the retrieved solution from the case-base to the needs of the target case (Watson, 1999). This usually involves some degree of adaptation of the retrieved case (Ji et al., 2012a). The significance of case revision lies in that it is difficult to provide a construction cost estimate with sufficient accuracy because of the low similarity of the retrieved case and target case (Jin, et al., 2012). Several studies have proposed a variety of revision methods including human intervention (Perera & Watson, 1998) such as adaptation by experienced estimators or experts and regression analysis (Jin, et al., 2012; Jin, Han, Hyun, & Kim, 2014; Kim & Hong, 2012a).

Watson and Marir (1994) mentioned that human collaboration should not be viewed as a weakness of CBR. Others considered that revision by humans is insufficiently reliable, inefficient and difficult to directly implement in the model (Jin, et al., 2012). Kim et al. (2012) propose a method in which estimators select or adjust the final estimate, where the practitioners' preference and modelling results are used together to increase the accuracy of estimation. However, this method is limited in terms of verification, because it depends on the estimator selecting or adjusting the final estimate value. Different estimators may have different preferences and cannot cope with big databases (Kim, Seo, et al., 2012). Moreover, Marzouk and Ahmed

(2011) have proved fuzzy technique's superiority in comparison with three other adaptation methods (null adaptation, weighted adaptation and Neuro-adaptation), while Ji et al., (2012) have also proposed an adaption method that decreases the need for adaptation and increases the capability of adaptation.

MRA is popular in the application of case adaptation as well. For instance, Ji et al. (2010a) and Jin et al. (2012) explore how to compensate the differences in attributes between the target case and the retrieved case by using MRA to amend the numerical variable values. A non-standardized factor can be used to calculate the effect of the independent variables on the dependent variable (Kim & Hong, 2012a). Jin et al. (2014) further improve the MRA method to model the deviations of both categorical and numerical variables using dummy variables. However, this method is also criticized due to its assumption of a simple and straightforward connection between features and solutions, which is not always justified (Du & Bormann, 2014b).

Case adaptation is the process of reducing the differences in the requirements between the new problem and the retrieved case (Craw et al., 2006). One of the most effective case adaptation approaches is to adjust the selection of promising candidate cases (Ji, et al., 2012b). Since using only one target case may not generate sufficient accurate results for the current problem, K-nearest neighbour (K-NN) method is widely used in case adaptation because of its convenience as the most basic instance-based method for approximating a real-valued or discrete-valued target function (Changchien & Lin, 2005; Shin & Han, 1999). It ranks the case's neighbours in the case-base and uses the labels of the K-most similar neighbour to predict the label of the new case (Liao et al., 2002). The weighted average value of the k most similar neighbours' solution is adapted for solving the current problem. However, K-NN has difficulty in deciding the value of k in a given situation since the determination of attribute weights has a significant influence on the efficiency and accuracy of the case retrieval (Barletta, 1991; Changchien & Lin, 2005).

2.2.6 Case Retainment in ECCE

Case retention is the process of updating the case-base by storing the revised solution of the new problem (Byung Soo, 2011). In case retainment, the question of what to retain and how to retain should be addressed. It can be deemed as the process of learning, triggered by the outcome of the evaluation and possible repair (Aamodt & Plaza, 1994). In practice, ECCE is usually conducted before the construction is

designed, and thus must depend on the case retained in the case-base previously. The retainment of the target case will be considered after its cost has been finalised. The decision whether to retain a case in the case-base can be made only when the project has been completed and the final cost has been calculated. Based on the evaluation of the solution provided by the CBR model and its final cost, the CBR system can be updated continuously. Despite the extensive work in CBR for ECCE (Ahn, et al., 2014; Ahn et al., 2017; Choi, et al., 2014; Chou et al., 2015; Ji, et al., 2018; Kim & Management, 2013; Lee, et al., 2013b; Leśniak & Zima, 2018; Tatiya, et al., 2018), there is a lack of research on case retainment.

As the final step in the CBR model, case retainment has a large influence on case-base and is closely related to the CBM. CBM is the process of refining the case-base to enhance the CBR system's performance. CBM not only involves the initial stage of case-base building, but also includes continuous case-base refinement and update (Leake & Wilson, 1998b). Although both case retainment and CBM involve case-base editing, the scope of CBM is much broader. In the ECCE CBR model, case retainment is used more frequently than CBM in existing ECCE studies. It includes several operations: deleting the cases to reduce the redundancy and inconsistency, clustering the cases to improve the reasoning power, editing the cases to repair incoherencies, etc. (Haouchine et al., 2007).

Despite their great importance in practice, little research attention has been given to case retainment or CBM. The situation in which the actual cost of the testing cases is already known when building the model has led to the study of case retainment and CBM being overlooked in existing ECCE studies. It would be simpler to only focus on certain steps, such as case retrieval, reuse and revision. This ignorance on the CBM not only hinders the development of a complete framework for the CBR model of ECCE, it also leads to difficulties in the enhancement of implementing CCE CBR research results into practice.

However, with the popularity of CBR applications in industrial environments, issues related to the building and maintenance of the case-base are becoming important (Khan et al., 2019a; Smiti & Elouedi, 2018b; Torrent-Fontbona et al., 2019). As data pre-processing is the most challenging and time-consuming for applying data mining techniques to real-world data, CBM has a significant influence on the CBR's performance (Bilal, et al., 2016). Recently, research has focused on CBM in ECCE. Ji

et al. (2018) produce a learning method to maintain and enhance the performance of ECCE CBR when there are insufficient cases in the case-base. The covariance effect in case-base has been addressed in another recent study (Ahn, et al., 2017). More research should focus on CBM to address the inconsistency between the existing research and practice.

2.3 CASE-BASE MAINTENANCE

2.3.1 Defining CBM

CBM is defined as the process of refining the case-base to enhance the CBR system's performance. CBM "implements policies for revising the organization or contents (representation, domain content, accounting information, or implementation) of the case-base in order to facilitate future reasoning for a particular set of performance objectives" (Leake & Wilson, 1998a). It involves various maintenance operations and has a very broad scope. For example, CBM may include editing a single case or multiple cases (e.g., adding or delete information of the cases), adjusting the way a case is represented (e.g. revising the case indexes,) revising the domain knowledge in the case-base (e.g., adding or deleting an entire case), editing the structure or contents (e.g., changing the structure of the case-base), or revising the implementation process of the case-base (e.g., building a filter for case-base retrieval or changing the method to evaluate the distance between the target case and the previous case). The main objective of CBM is to enhance the problem-solving ability of the CBR system from various perspectives (Wilson & Leake, 2001a).

In the CBR system, case-base is deemed as a fundamental component. Additional maintenance of case-base is necessary to deal with the problems that arise during long-term use, especially when the knowledge in case-base changes over time (Lupiani, et al., 2014a). CBM affects the case-base from different levels including representation level, the knowledge level, or the implementation level (Dietterich, 1986). As one of the exemplar-based learning approaches, CBR's performance is significantly influenced by the cases stored in the case-base. Therefore, the fundamental CBM strategy involves editing the case-base. In the case-base, insufficient data will impair the ability of the CBR model and cause potential negative effects on the results (Ji, et al., 2018). When the size of the case-base becomes large, there are more noises in the data and the retrieval process gets slow. This inevitably raises the question of how to select cases to avoid excessive storage and time

complexity (Khan, et al., 2019b). Thus research attention has been directed to CBM strategy improving the system's efficiency in retrieving solutions to the current problem, as well as the requirement of avoiding the interference of the noisy data in the CBR model (Wilson & Leake, 2001b). The CBM strategy could facilitate the accuracy of the CBR system by reducing the sensitivity to noise and the time for execution. Error reduction and redundancy reduction are the main focus of this area (Lopez De Mantaras et al., 2005). These techniques are designed based on the different requirements from the redundancy and noise level (Lupiani et al., 2016).

2.3.2 Criteria for Evaluating Case-base

The criteria for evaluating case-base are critical because how to maintain a CBR system is largely determined by the evaluating criteria (Leake & Wilson, 2000). Several concepts have been proposed to provide the basic knowledge for understanding and evaluating the case-base. The competence and the performance of case-base are two major dimensions for classifying these concepts (Smiti & Elouedi, 2011c). Competence is defined by the range of the problems whose solution can be successfully provided by the system. Smyth and McKenna (1995) first defined two basic core concepts in competence: coverage and reachability. The coverage of a case refers to the range of problems whose answer can be satisfactorily provided by this case. The reachability of a target problem refers to the range of cases whose solution can be used for the target problem. Case-bases with high coverage and low reachability are usually deemed as competent cases (Smiti & Elouedi, 2011a). Later the concept of competence is extended from individual cases to group cases by introducing the retrieval-space and the adaptation space. The group coverage is defined as a function that is proportional to the size of the group and inversely proportional to the density of case-base (Smyth, 1998). It can be measured by two indicators: the time needed to produce the solution for case targets and the accuracy of solving the target problem by using the retrieved solution (Smiti & Elouedi, 2011c). These concepts of case-base provide the various evaluation criteria for evaluating the case-base. Based on these concepts, the performance of a CBR system can be evaluated from three dimensions (Smyt & McKenna, 1999):

1. Efficiency of the CBR system (e.g., total problem-solving cost).
2. Competence of the CBR system (e.g., the scope of target problems can be satisfactorily solved).

3. Quality of the CBR system (e.g., the average accuracy of solutions provided by the system).

However, there is no guarantee that all the dimensions can be optimized at the same time. When CBR is used to solve a real-world problem, it is inevitable to encounter some constraints. These constraints, such as limited storage capacity of the case-base size, and the balance between long-term use and short-term use, are critical when using the ECCE CBR system in real-world problems (Leake & Wilson, 2000). The criteria for evaluating the system and constraints on CBR performance differ due to varying external circumstances.

2.3.3 Influencing factor in CBM

Several research efforts had been made to understand the influence of case-base properties including the number of cases stored in the case-base, the distribution of the cases indexes, and the density of cases (Barua et al., 2018; Ji, et al., 2018; Lieber, 1994; Smyth & Cunningham, 1996).

Case-base size is deemed as an obvious factor for measuring the competence of case-base. Despite the close relationship between the size of the case-base and the competence of the resulting system, the precise nature of their interaction is still not clear. Though it is simple to measure the number of cases in the case-base, the size of case-base cannot completely influence competence (Smyth & Keane, 1995). For example, some cases in the case-base can solve a wide range of target problems while others may only be useful when the unusual problem occurs. Obviously, the former class of case makes a greater contribution in competence property of case-base while the latter one may contribute more in accuracy property of case-base by solving the rare problem. Thus, the research question has been directed to optimizing the performance of the ECCE CBR system by increasing the size in the case-base (Ji, et al., 2018). Besides, the size of the case-base indirectly influences the efficiency of the CBR systems resulting in a great number of case-base reduction strategies (Jalali & Leake, 2014; Leake & Schack, 2015, 2016).

The density of cases in the case-base is also an important factor influencing the competence of case-base. The effect of the cases differs due to the density of the case-base. For example, an individual case in the low-density group may have less contribution to the competence of the case-case than those in the high-density group.

This is mainly because dense groups contain more redundant cases than sparse groups. Compared with the size of case-base, the density of the case is more complicated to measure. The density of cases in the case-base is measured by a function of case similarity (Smyth & McKenna, 1998). Smiti and Elouedi (2010) propose a CBM method based on density for reducing its size and preserving the maximum competence of the system. By clustering the large case-base into small clusters of cases, the amount of cases stored in the case-base is reduced, which can be easily maintained during its operations (Smiti & Elouedi, 2010). Later, they propose a density-based CBM method named Density Based Spatial Clustering of Application with Noise (DBSCAN). DBSCAN is designed to improve the CBR system's efficiency. In their study, DBSCAN is combined with Gaussian-Means (GM) algorithms for clustering and is named for the DBSCAN-GM model. Their research has a hypothesis that the large case-base with weighted features can be transformed into a small case-base by enhancing its quality (Smiti & Elouedi, 2014). Although DBSCAN-GM has obvious advantages of automatically discovering the number of clusters and noisy data, it is limited for objects being often doubtfully classified. They further improve the DBSCAN-GM by combining DBSCAN-GM and fuzzy set theory. The simulation experiments on several datasets have shown the effectiveness of this method (Smiti & Elouedi, 2016).

The distribution of cases is another significant factor (Yang & Zhu, 2001). The distribution of the dataset has a significant influence on how to set the CBR model. By using various parameters in case selection, the performance of the CBR model can be significantly enhanced (Kocaguneli et al., 2011). This conclusion is supported by a recent study that data distribution has an influence on calculating the optimal number for searching similar cases in the CBR model (Azzeh & Elsheikh, 2017). Azzeh and Elsheikh explore how to optimize the value of K from the data distribution. Results show that understanding of the data characteristic is advantageous in helping to search the optimal value of the model (Azzeh & Elsheikh, 2017). Furthermore, the distribution of solutions also matters when designing a CBM strategy. Despite the difference in the distribution of problem features, problems can still be solved well if all problems have very similar solutions, which means that space could be covered by a small number of cases or even several single well-chosen cases (Smyth & McKenna, 1998). Several distribution assumptions are widely used to represent real value random

variables, but the data in the real world sometimes fails to match the assumption. The changes in the case distribution may have a negative impact on the CBR's problem-solving ability (Ji, et al., 2018). For example, when problem distributions are non-uniform, the compactness and competence cannot be especially good indicators of CBR's performance (Wilson, 2001). When the features of the problem are not Gaussian distribution, the spread of the values will not be biased toward the center and the accuracy of the CBR system will be significantly impaired (Ji, et al., 2018). Therefore, the irregular cases should be retained and considered if cases are unevenly distributed. When the problems are from the densely packed region of the case-base, they have a bigger chance to be satisfactorily solved while those from a sparse region are more likely to remain unsolved or have solutions with more errors. Ji et al. proposed an learning method, which transforms the distribution of raw data from biased bell shape to uniform shape. Training samples can be continuously generated based on the most influential factor on the dependent value identified in the study. The results how that the performance of the ECCE CBR model can be improved by using specific parameters (Ji, et al., 2018).

2.3.4 Classification of CBM Strategies

Various CBM strategies have been proposed and researchers evaluate them based on different standards. Leake and Wilson (1998) proposed a CBM framework that classifies the CBM approaches into different processes based on their role played in determining when and how a CBR system should conduct CBM. This framework includes data collection, triggering and execution. Data collection is the process of gathering the data needed in the case-base and prepare the information for the next stage. Triggering is the process of determining whether maintenance operations should be conducted and selecting the maintenance operation based on the data collection. Execution is the process that the selected maintenance operation actually applies to the case-base (Leake & Wilson, 1998b).

Wilson categorizes these strategies based on the way maintenance strategies are executed into three dimensions: strategies targeting domain content, strategies targeting indices, strategies targeting maintenance policies. Strategies targeting domain content can be further classified into strategies adding and deleting cases, and strategies revising internal case content (Wilson, 2001).

Mantaras et al. (2005) explored the reduction techniques in CBM and classified them into two groups: to shrink storage requirements and reduce sensitivity to noise. (Lopez De Mantaras, et al., 2005). Similarly, Massie et al. (2006) classify them into two major areas: the noise control and the redundancy reduction (Massie et al., 2006). The noise control strategies endeavour to improve the accuracy of the CBR system by editing case-base, while redundancy reduction strategies attempt to improve the efficiency of the CBR system by reducing the size of a case-base.

Pan et al. (2007) classify CBM strategies based on three dimensions: case search direction, case order sensitivity and selection criteria. Based on case search direction, the CBM strategies can be classified into incremental strategy and decremental strategies. The incremental CBM strategies begin with an empty case-base and continuously add cases from the original case-base to form a new one until particular requirements are satisfied. Otherwise, it is called the decremental strategy. Based on case order sensitivity, the CBM strategies can be classified into order sensitive strategies and order insensitive strategies. The order sensitive CBM strategies measure the sequence of the case stored in case-base and use it as the selection criteria, while the order insensitive CBM strategies do not consider the influence of the order of case-base when selecting cases. Based on selection criteria, the CBM strategies can be classified into local criteria strategies and global criteria strategies. The local criteria strategies rely only on the local case parameters while global criteria strategies make decisions based on all the cases in the case-base. (Pan et al., 2007b).

Lupiani et al. (2016) classify the CBM strategies into four groups: nearest neighbour (NN) strategies, Instance-based strategies, DROP family, Competence and Complexity models (Lupiani, et al., 2016). Storage reducing is most widely used in NN strategies because it could significantly decrease the size of the storage needed by slightly sacrificing the learning rate and classification accuracy. However, the performance of NN strategies degrades rapidly with the level of attribute noise in training instances. Instance-based strategies are derived from the NN strategies (Aha et al., 1991b; Wilson & Martinez, 2000). Competence model and the Complexity Profiling Family conduct maintenance operations based on a comprehensive analysis of case-base. Several concepts including Coverage, Reachability and Relative Coverage are introduced in these studies (Smyth & McKenna, 1999; Zhu & Yang, 1999). Strategies can also be classified according to whether CBM strategies are

deterministic or not (Lupiani, et al., 2016). The deterministic CBM strategies always produce the same result from a given case-base while the non-deterministic CBM strategies may have different results each time that the strategy is operated.

To provide a comprehensive literature review on CBM, this research classifies CBM strategies in three groups: reduction approaches, partitioning approaches and optimization approaches. Reduction approaches refer to the methods that select a subset of cases from the original case-base to improve the system's performance. Partitioning approaches are those establishing an elaborate case-base structure and maintaining it continuously. Optimization approaches attempt to optimize the performance of case-base given several real-world constraints. In the following subsections, this thesis provides overviews of each type of strategy.

2.3.5 Case-base Reduction Strategies

Reduction strategies are the methods that enhance the CBR system's performance by removing cases. They can be classified based on their main purposes: enhancing the accuracy of the CBR model, or the efficiency of the model. Accuracy-based reduction methods aim to enhance the CBR system's accuracy by removing cases that have a detrimental effect on accuracy (Smiti & Elouedi, 2011a). The cases are classified into corrupt cases and non-corrupt cases. Corrupt cases are those with incorrect solutions. Therefore, the reduction strategy aims to minimize the effect of these cases on the CBR model by deleting them. Efficiency-based reduction methods aim to enhance the CBR system's performance by removing cases that have minimum effect on its contribution in case-base. The reduction method may have a negative influence on other aspects of the ECCE CBR system, thus care must be given before using these methods. They can be either incremental, beginning with an empty set and adding cases from the original case-base, or decremental where cases are removed from the original case-base.

Noise reduction strategies attempt to enhance competence by removing cases whose effect on accuracy is detrimental. Several noise reduction techniques are proposed including Edited Nearest Neighbour (ENN), Repeated Edited Nearest Neighbour (RENN), All k-NN and Blame Based Noise Reduction (BBNR) (Massie, et al., 2006; Tomek, 1976; Wilson, 1972). ENN is the best-known approach to address the noisy data. It removes the cases whose label cannot be correctly classified by using the solution of their k nearest-neighbours (Wilson, 1972). ENN intends to retain all the

internal cases and deletes the border cases. It is a decremental approach and it tends to sacrifice the competence of the case-base for consistency with the initial training set (Aha et al., 1991a; Wilson & Martinez, 2000). When the ENN algorithm is applied successively until no further cases can be removed, it is named RENN (Cummins & Bridge, 2011; Guan et al., 2009; Tomek, 1976). The All k-NN is similar to the iterative ENN, except that after each iteration the value of k is increased (Tomek, 1976). ENN and its variations can be deemed as noise removal techniques (Wilson & Martinez, 2000). They are widely used in noise reduction in various research fields (Kanj et al., 2016). Compared with traditional noise reduction methods such as statistics-based outlier mining technique and deviation-based outlier mining technique, ENN is a non-parametric method without any assumption on data set parameters or distribution parameters, which aims to maintain perfect consistency with the initial training set (Aha, et al., 1991a).

The Condensed Nearest Neighbour (CNN) rule is used to compress the size of the case-base (Hart, 1968). It compresses the case-base to a subset whose performance does not have too much difference with the original case-base. CNN scans all the cases searching for the cases whose label cannot be classified by using its k-nearest neighbours and then adds them to the new case-base. This searching process will be terminated when all the original cases are correctly labelled. This method facilitates the reduction in the size of case-base but is limited because it is sensitive to noise. Sometimes the noisy cases in the case-base may be deemed as important exceptions and provide an unsatisfying solution. Despite CNN and ENN sharing some common concepts, ENN intends to retain all the internal cases and deletes the border cases and the noisy cases. It is a decremental approach and it tends to sacrifice the competence of the case-base for consistency with the initial training set (Aha, et al., 1991a).

Reduced Nearest Neighbour (RNN) rule is then introduced to further improve the CNN by considering the CBR system's performance when selecting cases (Gates, 1972). RNN initially searches the cases in the original case-base whose removal does not cause any other cases to be misclassified and adds them to the new one. This searching process stops when no further reduction can be conducted. However, RNN is criticized for being time-consuming and expensive when the original case-base is large.

Ritter et al., proposed the Selective Nearest Neighbour Rule (SNN) to address the limitations in the previous method by setting a minimal consistent subset. SNN tends to sacrifice storage more than accuracy when there are noisy data in the case-base. However, it is more complicated and thus the learning process is longer than others.

Aha et al., introduce a series of Instance Base Learning (LBL) methods (Aha, et al., 1991b). LBL methods retain and use only selected instances to produce the predictions (Wilson, 1972). An instance can be deemed as a data structure with a vector of input attributes and output value. Thus, Instance-based strategies are deemed closely related to ENN. Several algorithms, including IB_1, IB_2, IB_3 are proposed in Instance-based strategies. The IB_1 algorithm is the simplest instance-based learning algorithm. Compared with IB_1, IB2 algorithm adds the function of saving only misclassified instances. The IB3 algorithm further extends the IB2 algorithm by using restrictive case selection criteria (Aha, 1992). Similar to NN strategies, Instance-based learning algorithms have to deal with the problem of determining which instances to retain for use in the learning process (Wilson & Martinez, 2000). It is inevitable to reduce the efficiency of learning, and leads to sensitivity to noise when too many instances are stored (Wilson & Martinez, 2000).

The Blame Based Noise Reduction (BBNR) algorithm further extended the competence model by considering how wrong a case's solution could be when solving the problems of other cases. The concept of liability is introduced to measure a case's contribution to misclassifications of other cases. For any case, if the cases in its coverage can be successfully labelled without using it, this case will be deleted when its liability set covers at least one case. This method measures the focuses on apportioning blame for misclassifications. It can maintain or even slightly enhance the generalisation accuracy (Delany & Cunningham, 2004).

Ni et al., proposed a CBM strategy based on outlier mining and case sieving methods (Ni et al., 2005). This method follows three steps: identifying outlier cases in the original case-base; deleting aggressive outlier cases; and sieving cases from non-outliers. The case-base without outlier cases is known as non-outlier case-base. The goodness of each case is measured and the one who has the maximum goodness value is added to a new case-base and deleted from the original case-base. This process will be terminated when the searching criteria is reached.

Wilson (2000) introduces the DROP families for editing the case-base (Wilson & Martinez, 2000). A key concept in these algorithms is the associate case. If one case P is selected in the neighbourhood of another case Q, the case Q is named as an associate of P. In DROP1, a case will be eliminated when the majority of the cases in the original case-base can be successfully predicted by CBR system without it. DROP2 considers the influence of the removal of a case on all the cases in the original case-base instead of considering only those cases selected in filtered case-base. DROP3 redesigns DROP1algorithm by adding a noise-filtering process before executing and helps to reduce the "overfitting" of the data. Compared with other methods such as CNN, SNN, ENN, RENN, ALL K-NN, the DROP algorithms show better performance, especially when dealing with the uniform class noise.

Massie et al.(2005) introduce a case-base editing based on the case-base complexity profiling (Massie et al., 2005). It calculates the local complexity based on the spatial distribution of cases within the case-base, which forms the complexity metric. The complexity metric provides the probability of finding another case in the nearest neighbourhood of a case with the same solution. This method estimates not only the ratio of redundant and noisy cases, but also the inaccurate cases (Massie, et al., 2006). Despite its usefulness in case discovery, this method is limited in evaluating the case-base competence.

2.3.6 Case-base Partitioning Strategies

The partitioning strategies partition the original case-base into subsets. Each subset of the case-base structure can be seen as one cluster generated from the clustering process. A representative case is created in each cluster and it takes a subset of the case features. Thus, the case feature with sufficient information is retained to provide a wide coverage of case-base. These strategies add or delete cases in each cluster simultaneously instead of editing the whole original case-base.

Shiu et al. uses a fuzzy decision-tree to update the knowledge between different case-base containers in the CBR system based on case coverage and reachability. The redundancy in the case-base can be significantly reduced by learning the fuzzy adaptation knowledge. This method follows four steps. The first step is to evaluate the feature weight of the case-base. The second step identifies different clusters in the case-base using the acquired feature-weights. The third step is to use fuzzy decision trees to mine the adaptation rules, followed by the final step of case selection (Shiu et

al., 2000). This approach works particularly well on the case-base, which has a lot of redundancy caused by the interaction among features. This type of redundancy can be reduced through learning the feature weights of the cases (Shiu et al., 2001). However, this method is limited for its great complexity in the rule's generation and the selection of the final cases.

Cao et al. proposed another partitioning strategy based on the fuzzy-rough approach (Cao et al., 2001). This strategy attempts to make the big case-base become small by using several particular adaptation rules. The overall complexity becomes low and the process of knowledge adapting is more efficient when compared to the maintenance results of using fuzzy ID3.

Yang and Wu (2000) proposed a density-based clustering approach for CBM. Each case is seen as an individual and the distance between different cases is measurable. The cases in one cluster share more common features than cases in different clusters. After transferring clusters to the new case-base, the contents of the new case-base become more concentrated and simpler to reuse and refine. The clustering result can be seen as a reference for a domain expert to adjust the case-base. Each cluster will be labelled with a name and a list of keywords that can briefly summarize the case-base. This method is simple and easy to use because it divides the large-scaled case-base into several small-scaled clusters of closely related cases. Then the number of cases in the case-base is small, any simple CBR method can be used (Yang & Wu, 2000).

Smiti and Elouedi (2010) introduce a COID method: Clustering, Outliers and Internal cases Detection (Smiti & Elouedi, 2010). Firstly, it clusters the original case-base into several small case-bases, which can be simple to maintain each one alone. Secondly, outlier detection is applied to identify the outlier cases and internal cases. This method tends to retain the cases that have an effect on the quality of each cluster of the case-base. It facilitates the maintenance of the case-base by ensuring each case-base is small and can be maintained individually. This method can be further extended by adding the measurements of feature weights to improve the CBR system's performance (Smiti & Elouedi, 2011b).

2.3.7 Case-base Optimization Strategies

Case-base optimization strategies refer to those optimizing the performance of case-base given several real-world constraints. When using the CBR system to solve the real-world problem, it is inevitable to encounter several application limitations (Leake & Wilson, 2000). The different organizations may have different requirements, resulting in the numerous studies on applying CBR with constraints. These constraints include the limited storage capability of the cases, the balance between long-term and short-term performance, the inconsistency between the data distribution and the availability of sources of cases (Leake & Wilson, 2000). Among those, suppressing the size of the case-base while maintaining the performance of the CBR system is most widely addressed. A lot of studies have been conducted to address this issue and some of them are summarised as follows.

As the simplest approach, random deletion strategy, removing cases randomly when a certain limit of the size of the case-base, is given (Zhu & Yang, 1999). This strategy sometimes works as well as other more expensive methods (Smiti & Elouedi, 2011a). However, it is criticized because of its limitation in the preservation of the competence of the case-base (Abdel-Aziz & Hüllermeier, 2015).

Another simple approach, delete cases, depends on the frequency of each case used in the retrieving process (Minton, 1990). However, it may possibly delete the important cases which may be very good for reuse. This method over-sacrifices the competence of the case-base. Both of these strategies suffer from the drawback that important cases can be possibly deleted by mistake. Utility deletion approaches are then introduced to address the previous limitations by deleting the case with negative utility (Smyth, 1998). Various concepts of case utility are introduced in several studies (Minton, 1990; Smyth, 1998; Smyth & Keane, 1995). Minton's utility metric is used for measuring its performance benefits. The utility of a case is calculated as follows:

$$\text{Utility} = [\text{Application\_Frequency} * \text{Average\_Savings}] - \text{Match\_Cost} \quad 2.1$$

Where Application_Frequency refers to the number of times the case has been retrieved, Average_Savings refers to the time reduced by retaining that case in the case-base. Match_Cost is the expenditure to compute similarity.

The utility deletion strategy deems the relation between the solution quality and the retrieving efficiency as a trade-off problem. The system efficiency is measured by

the total or average time to solve the target problems. In other words, reducing the retrieving time increases the system efficiency. The solution quality is determined by the average accuracy of the solutions or the quantity of good solutions provided by the system. These deletion policies suffer from the drawbacks of significantly impairing the competence of a CBR system and causing disastrous results. The possible deletion of important cases that may be very good for reuse will make certain target problems unsolvable.

Smyth and Keane propose a footprint deletion method based on the case classification. Cases can be classified into pivotal cases, auxiliary cases, spanning cases and support cases. A case is deemed as a pivotal case if the competence of the case-base is reduced after deleting it. A pivotal case is only reachable by itself. If the deletion of a case does not affect the competence of the case-base, this case is an auxiliary cases. Spanning cases are those whose coverage spaces span regions of the problem spaces. Some special spanning cases are known as support cases because they provide similar coverage to the others in a group (Smyth & Keane, 1995). The footprint deletion strategy deletes the cases in the order of auxiliary cases, support cases and pivotal cases (Lawanna & Daengdej, 2010). When the size of the case-base exceeds the limits, footprint deletion can be combined with utility deletion. The footprint-utility deletion strategy follows the rule: firstly, the footprint methods are used for selecting the cases. If there is only one case, then it is deleted, otherwise, the coverage and the reachability of the selected cases will be calculated. The cases with the lowest utility will be deleted from the case-base.

They also propose the competence guided editing strategy by using local case information to rank the cases for selection. Ranking the coverage and reachability can be deemed as a method for compressing the size of case-base (McKenna & Smyth, 2000; Smyth & McKenna, 1999). An authoring system is then designed where the case-base developers could manage the selection of adding or deleting cases from the case-base (McKenna & Smyth, 2001).

Iterative Case Filtering (ICF) is a method for filtering cases in the case-base based on coverage and reachability (Brighton & Mellish, 1999). The ICF algorithm exploits the lazy learning parallels and it deletes cases with size of reachable set larger than the coverage set. The rule is applied successively until no more cases can be removed. This strategy tends to retain boundary cases and remove central cases.

Yang and Zhu (2001) proposed a case-adding strategy that can generate a case-base with good coverage quality. Firstly, the neighbourhood of each case in the case-base is measured; secondly, adding the cases with the maximal benefit with respect to the existing neiborhood to the new case-base; thirdly, this process is continuously repeated until the stop criteria is satisfied. This method is provided with a theoretical analysis illustrating its capability to generate a well-defined range on coverage (Yang & Zhu, 2001).

Salamo and Golobardes (2003) introduce an Accuracy-Classification Case Memory (ACCM) algorithm based on the rough sets theory. ACCM aims to reduce the size of the case-base by using reachability and coverage separately. Another algorithm named Negative Accuracy-Classification Case Memory (NACCM) further extends ACCM by allowing a broader range of case selection than the ACCM technique. It aims to maintain the minimal size of the case-base. NACCM selects the cases near the outlier region and fewer cases are maintained, thus producing more reduction (Salamó & Golobardes, 2003).

Delaney and Cunningham propose the Conservative Redundancy Reduction (CRR) strategy, where cases with small coverage set are selected first. CRR strategy achieves a higher accuracy than those comparable but more aggressive strategies. It is always a trade-off problem between the level of compaction and competence preservation (Delany & Cunningham, 2004).

The latest studies focus on combining CBM with other methods. Smiti and Elouedi develop a novel soft CBM method by using soft competency model and fuzzy clustering technique. By analysing and revising the theoretical foundations of the current CBM approaches, their study attempts to enhance the competence and efficiency of the CBR system (Smiti & Elouedi, 2018a). Khan et al. introduce a hybrid CBM method to deal with the large scale of case-base. By equally utilizing the benefits of case addition and case deletion strategies, this method maintains the case-base in online and offline modes respectively (Khan, et al., 2019b). Nakhjiri et al. introduce Reputation-Based Maintenance (RBM) to enhance the classification accuracy of a CBR model while shrinking the size of its case-base. In RBM algorithm, the reputation for each case in the case-based is measured and it can be used to represent the related competence (Nakhjiri et al., 2019). Smiti and Elouedi develop a dynamic CBM method

based on machine learning techniques to address the slow speed brought by the continuous growth in the case-base (Smiti & Elouedi, 2019).

Previous studies assume that current case-base knowledge can be used as a proxy for future problems during the case retainment process. However, this assumption sometimes may not hold for sparse case-base during initial case-base growth, especially during dynamically changing domains. To address this issue, Leake and Schack presents a novel method named Expansion-Contraction Compression (ECC) when the assumption that current case-base knowledge can be used as a proxy for future problems during the case retainment process does not hold. Their study attempts to enhance the competence preservation when the representativeness assumption is only partially satisfied. ECCE attempts to broaden the range of the available cases by creating "ghost cases" (Leake & Schack, 2018). It has good performance when used in cross-domain problem-solving. In summary, all these strategies solve the trade-off problem from different perspectives (Massie, et al., 2006).

2.4 CHAPTER SUMMARY

This chapter has provided a comprehensive literature review of the research problem and its related areas. The background of the construction cost estimation, inaccuracy in construction cost estimation, influence factors in construction cost performance are reviewed together with the significance and challenges in ECCE. The application of CBR in ECCE carefully examines each step of the existing CBR model in ECCE. After briefly introducing the CBR and its advantage in ECCE, problem formulation, case retrieval, case reuse and case revision, and CBM, are reviewed to provide an in-depth understanding of the existing research. Section CBM includes the definition of CBM, the criteria for evaluating case-base, influencing factors in CBM, and classification of CBM strategy, case-base reduction strategy, case-base partitioning strategy and case-base optimization strategy.

Several findings are made based on the literature review. Firstly, MLA, ANN, and CBR are found to be the three most widely used methods for ECCE. Some studies show the contradictory results among these methods and thus, which of these methods performs the best remains a question. Although some research suggests the potential superiority of the CBR model for long-term use, there is no empirical evidence to support this assumption.

Secondly, despite some advantages brought by the increase of sample size, the current CBR models are limited when facing the challenges brought by informatization in the construction industry and rapid data growth. When the size of the case-base become large, noisy cases are inevitably increasing and impair the model's performance. As one of the exemplar-based learning approaches, CBR's performance is significantly influenced by the cases stored in the case-base and will be impaired by the noisy cases. Therefore, how to improve the robustness of the CBR model becomes a critical question.

Thirdly, the research attention on improving the CBR's performance for long-term use is found to be far from enough. The majority of ECCE CBR modes focus on certain steps such as case retrieval or case reuse and the research area in CBM is largely neglected. To tackle performance problems of a CBR system, it is necessary to update the existing case-base while maintaining problem solving competence. By combining the CBM strategy, the current CBR model of ECCE will be improved by reducing the negative influence of noisy cases and be more adapted for long-term use.

Several limitations are found in this chapter. Firstly, the existing research lacks a systematic understanding of the parameter setting in the CBR model. Various parameters, including weight determination, similarity functions, and case adaptation, are used in the previous ECCE CBR model, yet the question on how to combine these parameters to achieve the optimal results in the CBR model remains a question. There is no consensus on how to combine these parameters to achieve the optimal results in the CBR model. The literature review also finds that although several studies deem CBR advantageous for long-term use, there is no empirical study illustrating this advantage and how the performance of the CBR model changes with the increase in the number of cases in the case-base.

The literature review also finds some limitations on weight determination in the current ECCE CBR model. In the CBR model, the solution to a target case generates from the most similar previous cases. This process is determined by the similarity function, which is significantly influenced by the attribute weight. In the CBR system, each attribute can be seen as an index that contains a part of the knowledge stored in the case-base. Attribute weight reflects the influence of this knowledge component on case-base. Therefore, attribute weights can be deemed as indicators of the overall knowledge structure of the case-base. In the ECCE CBR model, attribute weights

inevitably change due to updating and refining of the case-base. However, the existing studies have limitations in minimizing the changes of the structure of case-base.

Additionally, the literature review finds some limitations on case-base maintenance (CBM) in the current ECCE CBR model. CBM is the process of refining the case-base to enhance CBR 's performance. Since case-base is a fundamental component in the CBR system, numerous studies emphasize that additional maintenance of case-base is necessary in the CBR system, especially when the knowledge in case-base changes over time (Lupiani, et al., 2014a). However, the existing ECCE CBR model extensively focuses on the initial establishment of the reasoning cycle, resulting in the ignorance of the case-base maintenance during long-term use. During the long-term use of the ECCE CBR model, the historical database will continually increase over time as more data is added to it, resulting in a high requirement for maintaining the ability of the case-base. In particular, the changed resource costs, construction methods, design styles and economic conditions create the outdated and inconsistent data, which should be carefully handled. Also, the size of the case-base can grow very quickly with the continuous use of the CBR model (Smiti & Elouedi, 2018a). The efficiency of solving a new problem thus becomes slow, resulting in the compromised overall performance of the CBR model (Khan, et al., 2019b; Lupiani, et al., 2014b). Without proper handling, this problem raised during long-term use will impair the performance of the CBR model: the typical issues being the low efficiency because of the continuously increasing size of the case-base. The research question of how to maintain the efficiency of the CBR system remains to be answered.

In summary, although various, the ECCE CBR applications have been developed, there is still a big gap between research and cost estimation practice. The popular trend of focusing on studying specific steps such as case retrieval, reuse, and revision has resulted in the other steps being ignored, which is hindering the development of a complete framework for CCE CBR. Based on the reviewed literature and research findings in this chapter, the proposed methodology and detailed research design is introduced in the next chapter.

Chapter 3: Research and Design

3.1 INTRODUCTION

The previous chapter provided a literature review on the related topics of the ECCE CBR model and the limitations in the current studies. This chapter next illustrates the research methodology to address the research questions and objectives in this study. It deals with developing and implementing the research plan for attaining the expected research outcomes. Several research components need to be considered during this process. As shown in Figure 3.1, this study considers the research approach, research methodology, research methods, model development, and data collection in the research design. This chapter begins by explaining the research approach used in this study. Section 3.2 provides the overall research framework by presenting research questions, research objectives, research methods, and expected research outcomes. Section 3.3 further explains the research methods used in this study. Section 3.4 shows the model development process. Section 3.6 discusses the data collection. Section 3.7 summarizes the timeline and the potential research limitations.

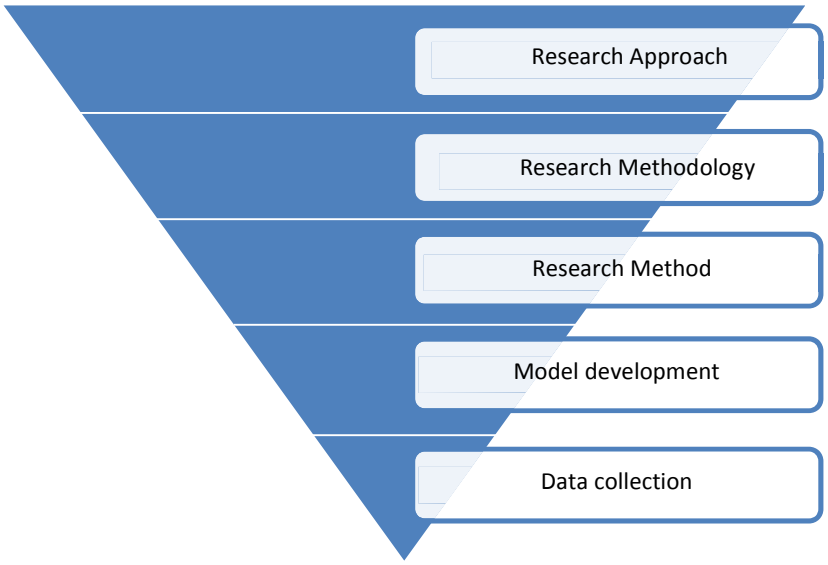


Figure 3.1 Structure of research plan

3.2 RESEARCH APPROACH

The research approach is a methodological link between the research steps chosen by the researcher for addressing the research questions. For the research to be effective, it has to comply with certain accepted criteria and must follow a prescribed procedure. Research philosophy, approach to theory development, methodological choice, research strategies, time horizon, techniques and procedures make up the completed research design (Mark et al., 2015). To provide a qualified research, it is necessary to consider these factors when designing the research.

There are five different research philosophies: positivism, critical realism, interpretivism, postmodernism, and pragmatism (Saunders, 2011). In this thesis, pragmatism and positivism are used. According to Saunders et al. (2011), pragmatism philosophy is based on the hypothesis that theory is only relevant where it supports action. Pragmatism emphasises the balance of objectivism and subjectivism, facts and values, accurate and rigorous knowledge, and different contextualised experiences. It evaluates and measures the role that the theories, concepts, ideas, hypotheses, and research findings play in attaining the expected practical outcomes in certain contexts. Pragmatism focuses on the practical effects and consequences of the ideas and knowledge.

In pragmatism, research begins with a practical research problem and attempts to provide a solution. Since practical outcomes are deemed more important than abstract consequences, identifying the research problem is critical in pragmatism. Thus, the research problem should combine the pragmatist emphasis on practical outcomes. When the research problems do not clearly illustrate which type of method should be proposed, the research could work with different kinds of knowledge and techniques. It is common to use multiple techniques in one study. Pragmatists deem that it is necessary to conducting research from various perspectives because no single approach can ever provide the entire picture of a research domain.

Positivist philosophy uses defined objectives and measures variables to derive conclusions. Positivist research uses a linear strategy of formulating hypotheses then attempts to disapprove these assumed relationships by concentrating on the null hypothesis. It is replicable and relies on deductive reasoning (Saunders, 2012). It allows the researchers to move from the theoretical position to a position based on empirical evidence (Cavana et al., 2001). The new position is based on the evidence

and assists in the identification of underlying theory that can be used to predict the behaviour of systems. In positivism, data gathering follows rigorous steps, and the quantitative data are analysed using statistical methods. To maintain objectivity during data collection and analysis, the researcher remains detached from the subjects. In this study, pragmatism and positivist philosophy are used together to identify the research problem, set the research objective, form the research hypothesis, analyse the results, and draw the conclusion. Table 3.1 has summarized the features of two types of philosophy.

Table 3.1 Features of positivist and pragmatism philosophy

Dimension	Pragmatism	Positivist
Assumptions	Concepts are only relevant where they support actions.	This world can be measured by science and 'mirrors' with privileged knowledge
AIM	To seek solutions to solve the practical problem.	To explore the rules that can be used to predict how and why things will happen
Stance of Researchers	The standard of being 'objectivist' or 'subjectivist' may be different based on different situation	Decisions should be made objectively.
Values	Researcher's values trigger the inquiry of the problem and drive the study.	The influence of researchers 'value is denied; value-free;
Research Plan	Rigorous, based on the research question	Rigorous, based on the research hypothesis
Research Methods	Mixed and multiple techniques, qualitative and quantitative methods	Experiments; surveys; interviews; secondary data analysis; statistical analysis
Goodness or quality of criteria	Emphasis on the practical solutions and outcomes	Reliability and objectivity; internal and external validity; benchmarks.

Three approaches are used for theory development in this study: deduction, induction and abduction. If research aims to test the already formed theory by collecting data, then it is known as deductive research; if research attempts to generate or build theory to explain a phenomenon, then it is an inductive approach. If research attempts to collect data to explain a phenomenon, draw conclusions, and identify patterns, to establish a new theory, to improve an existing theory, it is an abduction research. The features of three different theorys' development are shown in Table 3.2.

Table 3.2 Summary of deduction, induction, and abduction

Dimension	Deduction	Induction	Abduction
Logic	The deduction is in a deductive inference. The conclusions must be true when the premises are true.	The induction is in an induction inference. The known premises are used to generate untested conclusions.	The abduction is in an abdutive inference. The premises which are known are used to generate testable conclusions.
Generalisability	The deduction is from the general to the specific	The induction is from the specific to the general	The abduction is from the interactions between specifics and the generals
Use of Data	To evaluate hypotheses or propositions related to an existing theory	To explain a phenomenon, identify patterns and create a conceptual framwork	To explore a phenomenon, identify patterns and locate the findings in a theoretical field.
Theory	Verifying the theory.	Building and generating the theory.	Generating, building and modifying the theory.

Adapted from "Research methods for business students", by Mark Saunders, Philip Lewis, Adrian Thornhill, 2012, Understanding Research Philosophies and Approaches, p. 145

As shown in Table 3.3, positivist and pragmatism are used in research philosophy; induction and deduction are used as the approach for theory development. Mixed methods are used in methodology choice and strategies. A cross-sectional feature is adopted in the time horizon. CBR models, combined with GA, MRA, modal regression, and CBM, are used in the data collection and data analysis process.

Table 3.3 Selection of approach in this study

Research Elements	Selection of approach
Philosophy	Pragmatism and Positivist
Approach	Induction and deduction
Methodological choice	Mixed method
Strategies	Mixed methods research (literature review &Model development and validation)
Time horizon	Cross-sectional
Data collection and data analysis	Construction cost data
	CBR, GA, MRA, Modal regression. CBM

3.3 RESEARCH METHODOLOGY

Table 3.4 summarises the overall research framework in this thesis. The research framework is proposed to cooperate with research questions, research objectives, research methods, and expected research outcomes.

Table 3.4 Overall research framework

No.	Research question	Research objectives	Research method	Expected research outcomes
1	What limitations exist in the current ECCE CBR studies concerning long-term use?	To well address the gaps and limitations in current ECCE CBR research.	Literature review.	Identify the potential factors and propose the following research question.
2	Which methods are better for calculating weight and similarity based on different sample sizes?	To compare the methods for calculating weight and similarity and explore the influence of sample size on CBR.	CBR; MLR; GA; FC.	Understand the influence of the sample size, determining the optimal parameters in the CBR model for later use.
3	How to maintain a stable knowledge structure of the CBR model during long-term use?	To improve the robustness of the ECCE CBR model by using a robust weight determination method.	CBR; Robust regression.	The proposed method can improve the robustness of the CBR model.
4	How to improve the efficiency of the ECCE CBR model for long-term use?	To develop a CBM strategy for ECCE CBR models to maintain its efficiency during long-term use.	CBR; CBM strategy.	The proposed CBM strategy can improve the efficiency of the CBR model.

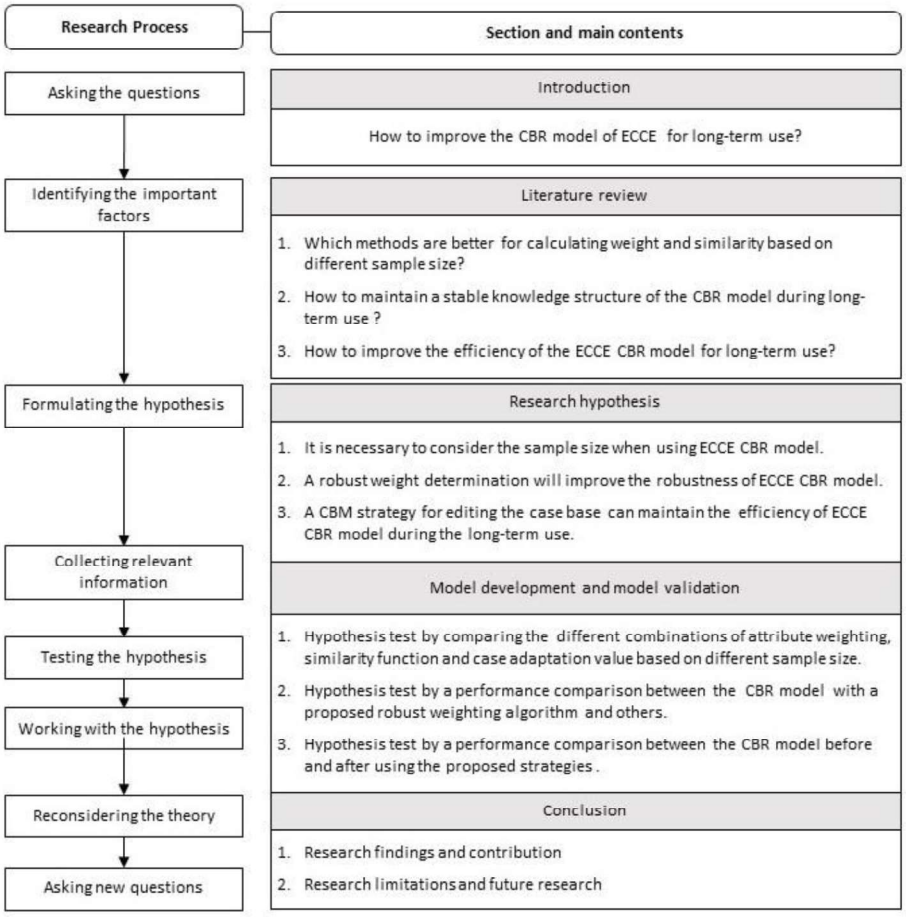


Figure 3.2 Research methodology (adapted from (Salkind & Rainwater, 2006))

Figure 3.2 illustrates the research methodology in this study (Salkind & Rainwater, 2006). Firstly, the main research question of how to improve the CBR model of ECCE for long-term use is proposed. Then, the critical factors are identified by conducting a literature review on ECCE, application of CBR in ECCE and CBM. Three research hypotheses are formulated to address the main research issue: it is necessary to consider the sample size when using ECCE CBR model; a robust weight determination would improve the robustness of the ECCE CBR model; and a CBM strategy for editing the case-base would improve the efficiency of ECCE during long-term use.

3.4 RESEARCH METHODS

3.4.1 CBR

CBR is a problem-reasoning process variously described as involving addressing a current issue by retrieving previous knowledge and experience (Kolodneer, 1991).

In the CBR system, new problems are solved by reusing and adapting the solutions of previous cases that have been successfully solved (Riesbeck & Schank, 1989). It is a process "to solve a new problem by remembering a previous similar situation and by reusing information and knowledge of that situation" (Aamodt & Plaza, 1994).

There are two parts involved in the CBR system: problem formulation and the reasoning cycle, as shown in Figure 3.3 (Richter & Weber, 2013). Problem formulation is made up of three stages, including case representation, case indexing, and case storage. Case representation is the process of describing the problem (Richter & Weber, 2013); case indexing is the process of assigning characterizing attributes to the cases (Chen & Burrell, 2001); and case storage is the process of establishing the structure of the case-base in preparation for the next step of the cycle (Watson, 1999). The CBR cycle consists of four "RE" processes of retrieval, reuse, revision, and retention (Aamodt & Plaza, 1994; Byung Soo, 2011). Retrieval is the process of searching for previous solutions from the case-base that enable the best solution to the new problem; reuse is the process of using the retrieved solutions to solve the target problem; revision is the process of adapting the retrieved solutions to solve the target problem; and retention is the process of updating the case-base by storing the new problem and its solution (Byung Soo, 2011).

Case representation

Case representation is the process of deciding what to store in a case and how to describe and organize the case contents in an appropriate structure (Aamodt & Plaza, 1994). Typically, case representation comprises identifying the problem, identifying the solution, and identifying the outcome (Watson, 1999). In ECCE CBR modelling, the problem is how to estimate the cost of a particular project, and the solution is the calculated cost derived from the historical cases. The outcome is to know whether the estimated cost successfully represents the actual cost, namely the error rate of the model. Two error measures, namely Mean Average Percent Error (MAPE) and the Root Mean Squared Error (RMSE) of the log values were calculated to evaluate the prediction performance as follows:

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{n}\frac{|actual_i - predicted_i|}{predicted_i} \times 100\% \quad (3.1)$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Log\,(|actual_i|) - Log(|predicted_i|))^2} \quad (3.2)$$

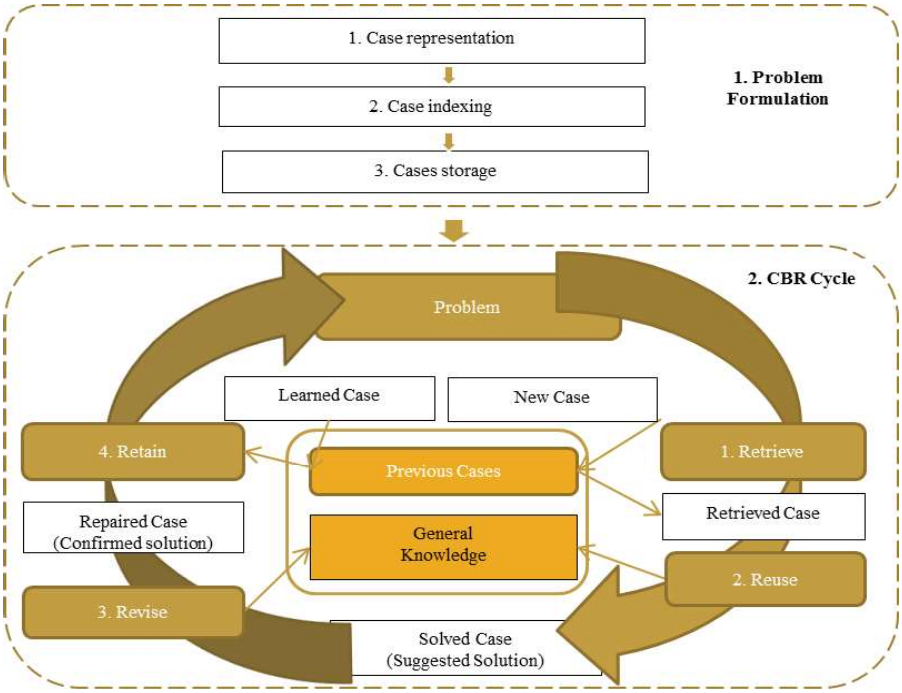


Figure 3.3 Tasks in the CBR Process

Case retrieval

In ECCE, case retrieval is the process of searching the cost of a previous project in the training set used as the cost of the new project. The core in case retrieval is the similarity function. As summarised in Chapter 2, the most widely used similarity functions are the weighted sum of the attribute distance and weighted Euclidean distance. Therefore, this thesis includes these to similarity functions. Equation 3.3 and Equation 3.4 illustrate the weighted sum of the attribute distance and weighted Euclidean distance respectively.

$$\mathrm{CS} = \frac{\sum_{i=1}^{n} w_i \times AS_i(x_a, x_b)}{\sum_{i=1}^{n} w_i} \qquad 3.3$$

$$CS = 1 - \sqrt{\sum_{i=1}^{n} w_i \times \left(A_i(AV_{new-case}) - A_i(AV_{retrieved-case})\right)^2} \qquad 3.4$$

where AV $_{new\text{-}case}$ represents the attribute value of the new case, AV $_{retrieved\text{-}case}$ represents the attribute value of the retrieved case.

Weight Determination

The determination of attribute weights significantly influences the performance of the CBR system (Changchien & Lin, 2005). The CBR's performance will be improved with the improvement of the evaluation of attribute weights (Lee, et al., 2013a). As summarised in Chapter 2, the three most widely used methods in previous

studies are the GA, MRA, and FC. Therefore, this study uses these three weight determination methods.

GA optimization

GA is a computational algorithm inspired by evolution. It is deemed as an optimization method and can be widely used in various research areas. In the GA algorithm, the possible solution to a specific problem can be labelled as a chromosome-like data structure. By applying recombination operators to these structures, the critical information is saved. Generally, the GA algorithm starts with a random population of chromosomes. Then these structures are evaluated, and reproductive opportunities are distributed in a manner that chromosomes with better solutions are offered more chances to reproduce than those with inferior solutions. (Whitley, 1994).

In this study, GA is used as another weight determination method in the ECCE CBR model. After computing the weight vector, they are used to label the attribute weight in the CBR model. More specifically, the attribute similarity and case similarity are computed by using these weights. Following the previous study (Ji, et al., 2018), the cost function that a case can be representing by appropriately weighting its attributes is measured as follows:

$$C_j = w_1 A_{1j} + w_2 A_{2j} + \cdots \ldots + w_i A_{ij} \quad 3.5$$

Where C_j, w_i, A_{ij} are the actual cost of the *j*th sample, the weight of the *i*th invariables, and the *i*th attribute value of *j*th sample, respectively. This formula can be represented using the following matrix:

$$\begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{1n} & \cdots & A_{1n} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = \begin{pmatrix} C_1 \\ \vdots \\ C_n \end{pmatrix} \quad 3.6$$

This research uses D_j to represent the distance between the actual cost of the jth sample and the sum of the product of the value of w_{ij} and attribute value A_{ij} , to optimize the value of w_i. To find the optimal value of w, this research conducts a minimization of the sum of the distance $\sum_{j=1}^{n} D_j$ by using GA. The fitness-function is defined as follows:

$$min \sum_{j=1}^{n} {D_j}^2, \quad 3.7$$

$$\text{when} \begin{pmatrix} D_1 \\ \vdots \\ D_n \end{pmatrix} = \begin{pmatrix} C_1 \\ \vdots \\ C_n \end{pmatrix} - \begin{pmatrix} A_{11} & \cdots & A_{1m} \\ \vdots & \ddots & \vdots \\ A_{1n} & \cdots & A_{1n} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} \quad 3.8$$

The MATLAB R2019 can be used to develop the GA model. The crossover rate and the mutation rate were set as 0.8 and 0.01.

MRA

MRA is a highly general and flexible data analysis method built on a defined mathematical basis (Kim, et al., 2004b). Basic MRA can be used whenever a quantitative variable, and the dependent variable Y, is to be studied as a function of the relationship (Cohen et al., 2013). There are several types of MRA, and ordinary least squares (OLS) are one of those widely used in determining the weight in the CBR model (Jin, et al., 2012; Jin, Han, Hyun, Kim, et al., 2014). OLS is a parameter estimating method, which minimizes the linear squares in a linear regression model. As one of the most powerful statistical purposes, OLS is convenient in practice (KARANCI, 2010; Kim, et al., 2004b). OLS assumes that the cost of a case can be represented by appropriately weighting its attributes and constant as follows:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.9}$$

where $\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{mn} \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_n \end{bmatrix}$

Where $\mathbf{Y}$, $\mathbf{X}$ are the dependent variable and independent variables respectively, $\boldsymbol{\beta}$ is the estimated value. $\mathbf{Y}$ can be deemed as the estimated total cost, and $\mathbf{X}$ can be deemed as measures of variables for estimating Y. For example, X_1 could be storeys and X_2 could be the total height of the building. $\boldsymbol{\epsilon}$ is the estimated constant, and $\boldsymbol{\beta}$ can be seen as parameters estimated by OLS. MATLAB R2019 can be used to develop the regression model.

Feature counting

The feature counting (FC) method deems all the features equal because of the assumption that no feature is more important than others (ChoongWan, et al., 2011; Dogan, et al., 2006). FC is considered to be effective for preventing bias against any factor (Dikmen, et al., 2007). Therefore, feature counting means the overall similarity of a historical project is only influenced by the quality of the cases themselves (Kim, Seo, et al., 2012).

Case adaptation

Case adaptation is the process of reducing the differences in the requirements between the new problem and the retrieved case (Craw, et al., 2006). One of the most

effective case adaptation approaches is to adjust the selection of the promising candidate cases (Ji, et al., 2012b). K-NN method is widely used as the case adaptation method. The K-NN ranks the case's neighbours in the case-base and uses the labels of the k most similar neighbour to predict the label of the new case (Liao, et al., 2002). A different number of neighbours can be used in the case adaptation.

3.4.2 Modal Linear Regression

Different from OLS, MODLR is another type of MRA method developed based on the conditional mode of the response Y. Mode is defined as the number that appears most often in a set of numbers. Therefore, in MODLR regression, the value with the highest probability of occurrence of Y is taken (Yao & Li, 2014).

Compared with other regression methods, MODLR generates a shorter prediction interval, and produces a more robust result when the data is not distributed as assumed. MODLR regression allows a skewed conditional distribution, and therefore the model pays greater attention to the main characteristics of the conditional distribution. It has received much research attention and is widely used in machine learning and artificial intelligence (Damir, et al., 2007; Wang, et al., 2017).

Basic concepts

There are several basic concepts in MODAL. As one of the MRA methods, MODAL shared some similar concepts with OLS. Given the independent and identically distributed observations (x_i, y_i), i=1, ⋯,n, to explore the relationship between the response y_i′s and the covariates x_i's, it is generally assumed the following linear regression model:

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i \qquad 3.10$$

where β is an unknown p × 1 vector, and the ε_i's are independent and identically distributed and independent of x_i with $E(\varepsilon_i \mid \boldsymbol{x}_i) = 0$. OLS method estimate $\boldsymbol{\beta}$ by minimizing the sum of squared residuals as follows:

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 \qquad 3.11$$

Introduction to MODLR

Suppose a probability density function f(y|$\boldsymbol{x}$) denotes the distribution of the response variable Y given the set of predictor x. MODLR assumes there is a unique

mode of f (y|$\boldsymbol{x}$), denoted by Mode (Y|$\boldsymbol{x}$) = argmax y (f(y|$\boldsymbol{x}$)). Therefore, Mode(Y|$\boldsymbol{x}$) is a linear function of $\boldsymbol{x}$, i.e.

$$\text{Mode}(Y|\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\beta}. \quad 3.12$$

In (5.3) the first element of x is the intercept term and assumed to be 1.The error is denoted as follows:

$$\epsilon = y_i - \boldsymbol{x}_i^T \boldsymbol{\beta} \quad 3.13$$

The conditional density of ϵ given x by g(ϵ|$\boldsymbol{x}$) can be deemed as the error distribution. To estimate the parameters, it is necessary to consider the condition of error distribution. When g(ϵ|$\boldsymbol{x}$) is symmetric about 0, the estimate of $\boldsymbol{\beta}$ in (Equation 3.12) will be equal to results obtained by conventional mean linear regression; when g(ϵ|x) is skewed, the results obtained by modal regression and classic mean linear regression will be different. For example, let ($\boldsymbol{x}$,Y) satisfy the following model assumption

$$\text{Y} = \text{m}(\boldsymbol{x}) + \sigma(\boldsymbol{x})\epsilon \quad 3.14$$

where ϵ has density $h(\cdot)$. Suppose $h(\cdot)$ is a skewed density with mean 0 and mode 1. If m($\boldsymbol{x}$) = $\boldsymbol{x}^{\text{T}}\boldsymbol{\beta}$ and σ(x) = $\mathbf{x}^{\text{T}}\boldsymbol{\alpha}$, then

$$\text{E}(Y|\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} \text{ and } \text{Mode}(Y|\boldsymbol{x}) = \boldsymbol{x}^T(\boldsymbol{\beta} + \boldsymbol{\alpha}) \quad 3.15$$

In this condition, both mean regression and modal regression can assume Y depends on x linearly; If m(**x**) = 0 and σ(**x**) = $\mathbf{x}^{\text{T}}\boldsymbol{\alpha}$, then

$$\text{E}(Y|x) = 0 \; and \; Mode(Y|x) = x^T(\alpha) \quad 3.16$$

Equation 3.14 to 3.16 shows that Y does not depend on **x** in terms of the conditional mean; while Y does depend linearly on **x** in terms of conditional mode. The difference between modal regression and OLS regression can be seen clearly in this instance.

To estimate the modal regression parameter β in (Equation 3.12), an objective function is developed based on the kernel density function as shown in (Equation 3.17).

$$Q_h(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n} \emptyset_h(y_i - x_i^T\beta) \quad 3.17$$

where $\emptyset_h(t) = h^{-1}\emptyset(t/h)$ and $\emptyset(t)$ is a kernel density function. Aftering choosing this kernel, the M-step of the MEM algorithm can be used for parameter estimating by maximizing the objective function in (Equation 3.17) by $\hat{\beta}$.

Modal EM algorithm

Similar to an EM algorithm, the MEM algorithm consists of an E-step and an M-step:Starting with $\beta^{(0)}$, repeat the following two steps until it converges(Li, Ray, and Lindsay, 2007; Yao, 2013):

E-Step: In this step, we calculate weights $\pi(j|\boldsymbol{\beta}^{(k)})$,j = 1,...,n as

$$\pi\left(j\middle|\boldsymbol{\beta}^{(k)}\right) = \frac{\emptyset_h\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}^{(k)}\right)}{\sum_{i=1}^{n}\emptyset_h\left(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}^{(k)}\right)} \propto \emptyset_h\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}^{(k)}\right) \quad 3.18$$

M-Step: In this step, we update $\boldsymbol{\beta}^{(k+1)}$

$$\boldsymbol{\beta}^{(k+1)} = arg\max_{\beta}\sum_{j=1}^{n}\left\{\pi\left(j\middle|\boldsymbol{\beta}^{(k)}\right)log\emptyset_h\left(y_j - \boldsymbol{x}_j^T\boldsymbol{\beta}\right)\right\} \quad 3.19$$

$$= (\boldsymbol{X}^T\boldsymbol{W}_k\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}_k\boldsymbol{y}, \quad 3.20$$

where $\mathbf{X} = (\boldsymbol{x}_1,...,\boldsymbol{x}_n)^T$, $\mathbf{W}_k$ is an $n \times n$ diagonal matrix with diagonal elements $\pi(j|\boldsymbol{\beta}^{(k)})$s, and $\boldsymbol{y} = (y_1,...,y_n)^T$.

The major difference between the OLS estimate and the MODLR lies in the E step.

E-Step is a critical step causing the major difference between MODAL and OLS. In the OLS regression, equal weights are used for each observation. However, in MODAL regression, the weights of observations depend on how close y_i is to the modal regression line. By using unequal weights of observations, MODAL can minimize the influence of observations far away from the modal regression line.

Since the normal kernel is exploited in Equation 3.17, the function optimized in the M-step is a weighted sum of log-likelihoods corresponding to OLS regression. As the starting point influences the MEM algorithm, it is difficult to guarantee that the algorithm will converge to the optimal global solution (Equation 3.20). Therefore, it is necessary to use different starting points to run the algorithm multiple times to find the best local optimal. For any choice of kernel for ϕ in (Equation 3.17), the proposed MEM has the ascending property. Each iteration of (Equation 3.19) and (Equation 3.20)

will monotonically non-decrease the objective function (Equation 3.17), i.e., $Q_h(\beta^{(k+1)}) \geqslant Q_h(\beta^{(k)})$, for all k.

Bandwidth Selection

Bandwidth selection is a critical step in the MODLR regression. Although theoretical results on the traditional M-estimators cannot be directly applied to the proposed MODLR estimator, the asymptotic optimal bandwidth h for estimating β_0 can be obtained by minimizing the asymptotic mean squared errors (MSE) (Yao & Li, 2014). Since the asymptotically optimal bandwidth formula contains the unknown variables, these unknown variables are replaced by their estimates.

Given the initial residual $\hat{\epsilon}_i = y_i - x_i^T\hat{\beta}$, $\hat{\beta}$ is estimated by the robust estimate proposed in the previous study (Huber, 1984), later, their mode $\hat{m}$ can be estimated by maximizing the kernel density estimator (Parzen, 1962). It is assumed that ϵ is independent of x and x,$\hat{\varepsilon}_i - \hat{m}$ approximately has density g(·), thus $g^{(v)}(0|x)$ can be estimated by using the following equation

$$\hat{g}^{(v)}(0|x) = \frac{1}{nh^{v+1}}\sum_{i=1}^{n} K^{(v)}\left\{\frac{\hat{\epsilon}_i - \hat{m}}{h}\right\}, v = 0,2,3 \qquad 3.21$$

where h is determined by using the method proposed in the previous study (Botev et al., 2010) and $K^{(v)}(\cdot)$ is the *v*th derivative of kernel density function K(·). Then *J, K, and L* can be estimated by using the following equation:

$$\hat{J} = n^{-1}\sum_{i=1}^{n} \hat{g}''(0|x_i)x_i x_i^T \qquad 3.22$$

$$\hat{K} = n^{-1}\sum_{i=1}^{n} \hat{g}'''(0|x_i)x_i \qquad 3.23$$

$$\text{and } \hat{L} = n^{-1}\sum_{i=1}^{n} \hat{g}(0|x_i)x_i x_i^T \qquad 3.24$$

Therefore, the asymptotic optimal bandwidth $\hat{h}_{opt}$ can be estimated by using the following equation:

$$\hat{h}_{opt} = \left[\frac{3v_2 tr(J^{-1}LJ^{-1}W)}{K^T J^{-1}WJ^{-1}K}\right]^{1/7} n^{-1/7} \qquad 3.25$$

where $v_2 = \int t^2 \emptyset^2(t)dt$ and tr $(J^{-1}LJ^{-1}W)$ is the trace of $J^{-1}LJ^{-1}W$ and W is a diagonal matrix, whose diagonal elements reflect the importance of the accuracy in estimating different coefficients.

When $W = (J^{-1}LJ^{-1})^{-1} = JL^{-1}J$, it can be deemed as proportional to the inverse of the asymptotic variance of $\hat{\beta}$. Therefore, asymptotic optimal bandwidth $\hat{h}_{opt}$ can be estimated by using the following equation

$$\hat{h}_{opt} = \left[\frac{3v_2(p+1)}{K^T L^{-1} K}\right]^{1/7} n^{-1/7}. \quad 3.26$$

The bandwidth selection can be iteratively updated by recalculating the residual $\hat{\epsilon}_i$ given by estimated value in MODAL. In MODLR model the weight of observation differs due to the distance between the y_i and the modal regression, while in OLS regression, each observation has equal weight (Yao & Li, 2014). The further distance between the point y_i and the modal regression line, the less weight of the point y_i. By using the MODLR regression in weight determination, the robustness of the CBR model can be improved, and the negative effect of noisy cases in the CBR model can be reduced.

3.4.3 CBM Strategy

With the long-term use of the ECCE CBR model, the historical database will continually increase over time as more data is added to it, resulting in a high requirement for updating and maintaining the ability of the case-base. Thus the CBM strategy can enhancing the CBR system's efficiency. The term 'case-base editing' is used when referring to the process of updating a given case-base through adding, deleting, and combining cases.

Basic concepts

Several basic concepts in the CBM strategy should be introduced (Pan et al., 2007a). A case-base is a set of cases collected in practice. It can be defined as a problem-solution pair. Namely, any given case c in a case-base can be seen as a pair $c = (x, s)$, where $s \in S$ is a solution to a problem description x, and S is a set of solutions. For any training case-base T denoted by $T = \{(x_i, s_i), i = 1,2,\ldots,N\}$, the objective is to select a subset, which is called a new case-base, denoted by $CB = \{(x_{i_j}, s_{i_j}), j = 1,2,\ldots,M\}$, that the performance of the model based on the new CB is better for future problems. For each problem x_1 in a case-base T, the corresponding solution s_1 can be seen as a function of the problem description, namely $s_1 = \pi(x_1)$. Therefore, c_1 can be

represented by (x_1,$\pi(x_1)$). Let N (x_1) be the set of problem x_s whose solution $\pi(x_s)$ is deemed close to $\pi(x_1)$. More formally:

$$N(x_1)=\{\{x_s | D(\pi(x_1),\pi(x_s)) \leq L\}\} \quad 3.27$$

Where L is an allowed similarity difference between the x_s to x_1. Essentially, N defines a coverage of x_1. $N(x_1)$ is the coverage or neighbourhood of x_1. It shows the problem solving capability of x_1 in the existing case-base. After evaluating the neighbourhoods of all cases in a case-base, a case is deemed good when its problem solving capability is high. To select good cases, the frequency of cases occurring should be taken into consideration. For example, if the cost of a project is more used as the solution of other projects, then this project is naturally deemed as a better project for minimizing the searching cost.

To evaluate the contribution of each case in the case-base, the coverage is defined as follows.

Given a case space X, let $x \in$ X be a case. For any given case-base x, denote N(x) the neighbourhood of x. For each x, its coverage can be determined given the value of L. P(x) is the frequency that x is considered as other cases' coverage in the case space X. Therefore, the coverage contribution (CC) of a case x in the case-base X, is defined by the total frequency that x is selected as a neighbourhood of remaining cases in the case-base. The ratio between CC of a case x and total neighbourhood of case-base is represented by the coverage contribution ratio (CCR). More formally, the coverage contribution of x in the case-base T is defined as:

$$M(x) = \sum_{x \in T} P(x) \quad 3.28$$

Then CCR of any cases x in any case-base X is defined as:

$$R(x) = \frac{P(x)}{\sum_{x \in T} P(x)} \quad 3.29$$

Since $x \in$ X, the case coverage is a real number between 0 and 1, and the sum of the coverage contribution ration of each case in case-base equals one. (Zhu & Yang, 1999).

Weighted coverage contribution

For a given case x_1, $N(x_1)$ differs due to the value of L. The larger value of L provides a broader range of neighbourhood while the smaller value of L provides a narrower range of similar cases. For any two different L_1 and L_2, the neighbourhood of x_1 is determined as follows:

$$N_1\ (x_1)=\{\{x_s|D(\pi(x_1),\pi(x_s))\le L_1\}\} \quad 3.30$$

$$N_2\ (x_1)=\left\{\left\{x_s\middle|D\left(\pi(x_1),\pi(x_s)\right)\le L_2\right\}\right\} \quad 3.31$$

If the value of L_1 is less than L_2, then the neighbourhood of x_1 determined by L_1 can be seen as a subset of the neighbourhood of x_1 determined by L_2, namely, if $L_1 <$ then $N_1\ (x_1) \subseteq N_2\ (x_1)$. Since L is an allowed difference between the x_s to x_1, it is naturally assumed that $N(x_1)$ calculated by the smaller value of L may be deemed more important. Therefore, a weight function given the value of L is necessary to measure the difference caused by different values of L. Since K-NN is used in the case adaption, the different value of K decides a different range of neighbourhood of x_1 . Given the influence of the neighbourhood range, the coverage contribution is weighted given the value of the K.

For any given integer K, define a set S=$\{1, \dots, K\}$. For each $s_i \in$ S, the value of the L is determined by choosing the value of s_i. Accordingly, the neighbourhood of x_1 can be defined as follows:

$$N_i\ (x_1)=\{\{x_s|D(\pi(x_1),\pi(x_s))\le f(s_i)\}\} \quad 3.32$$

The value of s_i determines the $N_i\ (x_1)$, thus the neighbourhood of x_1 can be classified into K sets ($S_1, \dots, S_k$) given the value of s_i. When s_i=1, the neighbour of x_1, denoted by $N_1\ (x_1)$, is the smallest and $N_1\ (x_1)$ can be seen as S_1; When s_i=2, the neighbour of x_1, denoted by $N_2(x_1)$, contains $N_1\ (x_1)$. Therefore, the difference of the $N_1(x_1)$ from $N_2(x_1)$, denoted by $N_2(x_1) - N_1\ (x_1)$, can be seen as S_2. Similarly, for any given $s_i\ (i \le 2), S_i = N_i(x_1) - N_{i-1}\ (x_1)$, i=2, …, K. For each element in S_i, coverage contribution should be evaluated differently by considering which set it belongs to. To better measure the coverage contribution, a weight function of S_i is proposed to represent the differences in each set. For any x_1, the coverage contribution weight (CCW) can be defined as follows:

$$CCW(x_1)=\begin{cases} k, \ if\ N(x_1) \in S_1 \\ \vdots \\ k-1, if\ N(x_1) \in S_i \\ \vdots \\ 1, if\ N(x_1) \in S_k \end{cases} \quad 3.33$$

Where $N(x_1)$ is the neighbourhood of x_1, $N(x_1)$ significantly influences the process of case adaptation. By using CCW, the differences in coverage contributions are considered in the case-base. Note that when K equals one, the coverage contribution of cases is deemed as equal. Table 3.5 provides the summary of the major notations in this chapter.

Table 3.5 Summary of the major notation.

Notation	Meaning
c	Case
x	Problem
s	Solution
T	Any training case-base
π	The function between problem description and solution
N(x)	Coverage or neighbourhood of x
P(x)	Frequency that x is considered as other cases' coverage
M(x)	CC of *x* in the case-base T
R (x)	CCR of *x* in the case-base T
CCW (x)	The weight of CC

Case-base editing methods

Case searching direction is a significant factor to consider when editing the case-base. CBM strategies can be classified into incremental strategies and decremental strategies. The incremental CBM strategies begin with an empty case-base and continuously add cases from the original case-base to form a new one until particular requirements are satisfied. On the contrary, the decremental CBM strategies begin with the original case-base and continuously delete cases from it until particular requirements are satisfied (Pan, et al., 2007a). Compared with the incremental CBM strategies, detrimental CBM strategies have a global view of the remaining case-base. Thus this research decides to use the detrimental strategies for editing the training case-base.

Given a case space X=$\{x_1, x_2, \dots, x_n\}$, let $x \in$ X be a case whose cost should be estimated. The proposed method to edit the case-base is shown as follows:

(1) Calculating the coverage contribution weight for any given K;

(2) For each training case x_i in $X_r = \{x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n\}$, calculated its coverage N (x_i);

(3) Calculating the weighted coverage contribution of each training case x_i in the case-base X_r;

(4) Ranking the training cases with respect to their weighted coverage contribution and determining the threshold of weighted coverage contribution.

(5) Selecting a subset D containing p instances with the lower weighted coverage contribution;

(6) Updating the case-base by deleting those in D;

3.5 MODEL DEVELOPMENT

As shown in Figure 3.4, this study consists of two parts. The first part includes establishing a classic CBR model. The classic CBR model includes reasoning cycle and case-base. The reasoning cycle includes problem formulation and the CBR cycle. Problem formulation is the process of identifying the problem, identifying the solution, and identifying the outcome, and it involves organizing case representation, identifying evaluation indicators, and planning data collection. CBR cycle consists of case retrieval, case adaptation, and case retainment. Case retrieval refers to the process of determining the best matching previous solutions for the current problem. In this study, calculations of attribute weights and retrieval of the similar case constitute the case retrieval process. In the classic model, the calculations of attribute weights include the three most widely used methods in previous studies: GA, MRA, and feature counting. Retrieval of similar cases includes the two most popular distance functions: the weighted sum of the attribute distance and weighted Euclidean distance. Case adaptation is the process of reducing the differences in the requirements between the new problem and the retrieved case. One of the most effective case adaptation approaches is K-NN by adjusting the selection of the promising candidate case. Therefore, the K-NN method is used in case adaptation. Two error measures, MAPE and RMSE, were used as evaluation indicators in this study.

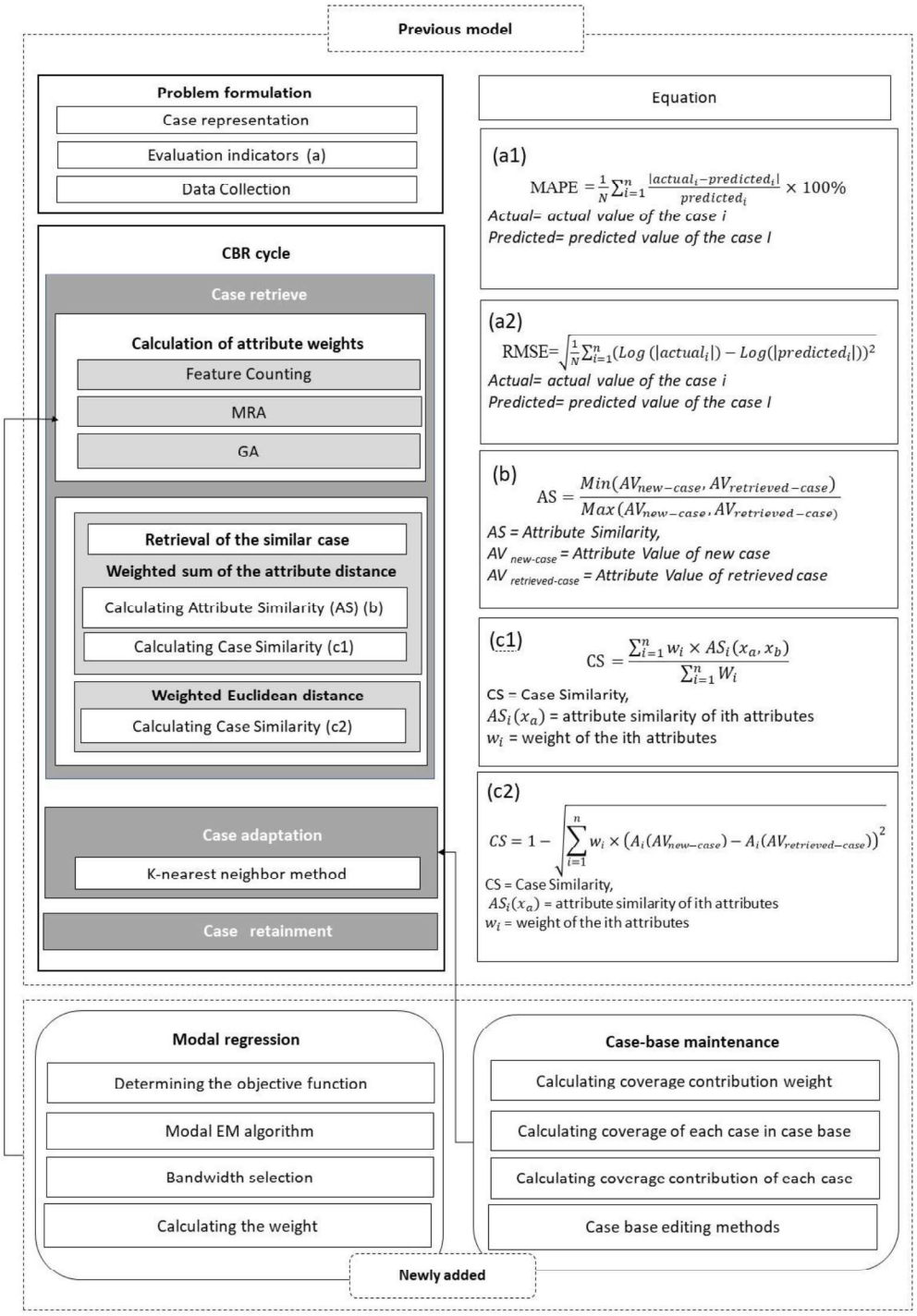


Figure 3.4 Model development

The ECCE CBR model includes the most widely used weight determination and retrieval method in previous studies. The first part of the research involves a comparative study of different combinations of parameter settings of the CBR model.

It aims to explore how sample size influences the performance of the CBR model and how to combine the parameters to get the optimal results. The comparative study provides a clear understanding of the advantages and disadvantages of each setting. The results assist in testing the hypothesis, that CBR has the advantage for long-term use when compared with other methods, as well as providing a reference for subsequent optimization of the long-term use of the ECCE CBR model.

The second part includes the optimizations in terms of the robustness and efficiency of the ECCE CBR model during long-term use. The first optimization is proposed to improve the robustness of the existing model. In the ECCE CBR model, attribute weights inevitably change due to updating and refining of the case-base. However, these changes should be minimized because of the consistency requirement of the knowledge structure in case-base. The stability of the attribute weights can be considered as the indicators of the CBR model's robustness. When the size of case-base gets large, it is inevitable to have a few cases which deviating from those mainstream bulk of cases (Chan & Wong, 2007). The existing weight determination methods have limitations in being sensitive to the outliers. A single outlier can have a large effect on the parameter estimates, thus will reduce the accuracy of the model. Therefore, improving the robustness of the ECCE CBR is the primary task during long-term use. To address this issue, a robust weight determination method, modal regression, is introduced in this study. By comparing this weight determination method with the existing methods, the superiority and effectiveness of the proposed method is validated.

The other optimization method is designed to improve the efficiency of the CBR model. Since case-base is a fundamental component in the CBR system, additional maintenance of case-base is necessary in the CBR system, especially when the knowledge in case-base changes over time. During the long-term use of the ECCE CBR model, the efficiency of solving a new problem becomes slow, resulting in the compromised overall performance of the CBR model. Therefore, an original CBM strategy is designed to improve the efficiency of CBR for ECCE. Several concepts including coverage, coverage contribution and weighted coverage contribution are introduced to evaluate the cases, followed by the illustration of the CBM strategy. By comparing the results of the CBR model before and after applying the proposed CBM strategy, the effect of the CBM method can be evaluated. All weight determination

methods including OLS, GA, MODAL are used to illustrate the effectiveness of the proposed method. By introducing the concept of weighted coverage contribution, this research attempts to compress the case-base to address the low efficiency of the retrieval process caused by the continuous growth in the size of case-base. The results can help the ECCE CBR model to avoid excessive storage and time complexity when dealing with the rapid growth of data in the construction industry.

3.6 DATA COLLECTION

3.6.1 Project Type

For the selection of predictor variables for ECCE, the basic classification is project type (e.g., buildings, roads, facilities and bridges), because the variables involved vary significantly for each. For example, the gross floor area is a typical major variable for buildings, while the length is a major variable for roads. The uncertainty surrounding the project phase of application due to lack of design development also means that some studies can only use early information such as landscape area (RunZhi, et al., 2012) and building type (Jin, Han, Hyun, & Kim, 2014), while others are able to utilize such detailed project information as the area of exterior finishes (Kim, Seo, et al., 2012), and type of overhang design (Doğan et al., 2006b). As shown in section 2.1.9, residential building is most widely studied in the CBR model because of its significance in the construction sector and strong market demand. Therefore, this study uses residential building for data collection.

3.6.2 Predictor Variables

Despite the extensive research on ECCE, there is no generally accepted standard set of predictors involved. Information and time constraints are two factors that need to be considered when determining the predictor's variables for ECCE. ECCE usually is deemed as the initial construction cost estimate completed for a construction project, as shown in Figure 3.5 (Gardner Brendon, et al., 2016). At the early stage, only the least amount of project information is available. With the release of details on project design, the construction cost estimation accuracy is updated with more precision.

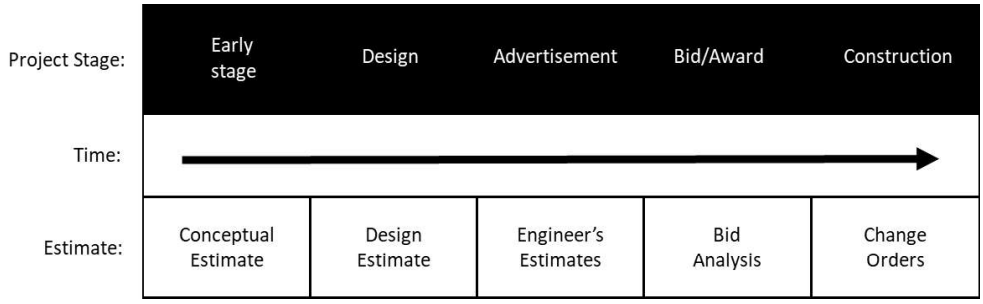


Figure 3.5 Construction cost estimating timeline (adapted from (Gardner Brendon, et al., 2016)

Data collection effort constraints also need to be considered when determining the predictors. (Akintoye & Fitzgerald, 2000; Arafa & Alqedra, 2011a; Petroutsatou, et al., 2011). ECCE provides the basis for the owners or the investigators to choose adequate alternatives and it requires high accuracy (Cheung & Skitmore, 2006c). However, the accuracy of an estimate and its preparation cost is not always proportional (Sanders et al., 1992). At some point, increased accuracy cannot justify the additional costs incurred. The sooner ECCE is developed, the more effort and cost can be saved for subsequent work.

Table 3.6 Summary of variables used in previous research for residential buildings

No	Reference	Predictors	Type of buildings
1	(Kim, et al., 2004b)	(1) the value of the gross floor area. (2) the value of storeys. (3) the value of the total unit number. (4) the value of duration. (5) the value of roof types. (6) the value of FDN types. (7) the value of the usage of the basement. (8) the value of finishing grades.	Residential building
2	(Kim, et al., 2005)	(1) the value of location. (2) the value of the area. (3) the value of storeys. (4) the value of roof types. (5) the value of the total unit. (6) the value of the unit per storey. (7) the value of the average area of the unit. (8) the value of the foundation type. (9) the value of the usage of the basement. (10) the value of finishing grades. (11) the value of duration.	Residential building
8	(Doğan et al., 2008)	(1) the value of the total area of the building. (2) the value of the ratio of the typical floor area to the total area of the building. (3) the value of the ratio of the footprint area to the total area of the building. (4) the value of the number of floors. (5) the value of the overhang design type. (6) the value of the foundation system. (7) the value of the floor structure type. (8) the value of the location of the core.	Residential building

3	(Koo, Hong, Hyun, & Koo, 2010)	(1) the value of the delivery method (DB or DBB). (2) the value of multi-family house type. (3) the value of households number. (4) the value of location. (5) the value of non-working days. (6) the value of the total floor area. (7) the value of no. of storeys above the ground. (8) the value of no. of storeys below the ground. (9) the value of the size of the household. (10) the value of land ratio.	Residential building
4	(KARANCI, 2010)	(1) the value of the year. (2) the value of duration. (3) the value of total construction area. (4) the value of the total site area. (5) the value of the total number of apartment blocks. (6) the value of the total number of apartments. (7) the value of percent area of social buildings in the total construction area. (8) the value of the earthquake region. (9) the value of the category of site topography. (10) the value of the type of insulation. (11) the value of the number of elevator stops. (12) the value of classification for degree day.	Residential building
5	(Hong, et al., 2011)	(1) the value of the structure type. (2) the value of the foundation type. (3) the value of the ground plan of the building type. (4) the value of façade type. (5) the value of household type. (6) the value of the ground plan of the household. (7) the value of floor height. (8) the value of the number of layers. (9) the value of pilotis size. (10) the value of building-to-land ratio. (11) the value of building ratio. (12) the value of the gross floor area. (13) the value of lot area. (14) the value of basement gross area. (15) the value of landscaping area. (16) the value of the gross floor area of the subsidiary facilities. (17) the value of the area of the underground parking lot. (18) the number of buildings (19) the number of households	Residential building
6	(Ji, Park, & Lee, 2011)	(1) the number of households. (2) the quantity of gross floor area. (3) the value of unit floor households. (4) the number of elevators. (5) the number of floors. (6) the quantity of pilotis with household scale. (7) the number of households of unit floor per elevator. (8) the value of height between storeys. (9) the value of depth of the pit. (10) the value of the roof type. (11) the value of hallway type. (12) the value of structure type (RC).	Residential building
7	(RunZhi, et al., 2012)	(1) the value of the site area. (2) the value of the underground area. (3) the value of the ground area. (4) the value of the building area. (5) the value of building coverage ratio. (6) the value of the floor area ratio. (7) the number of underground floors. (8) the number of floors. (9) the value of height. (10) the value of landscape area	Residential building
8	(Ji, et al., 2012a)	(1) the number of households. (2) the quantity of gross floor area. (3) the number of unit floor households;(4) the number of elevators; (5) the number of floors; (6) the number of pilotis with household scale; (7) the number of households of unit floor per elevator. (8) the value of the height between stories. (9) the value of the depth of the pit. (10) the value of the roof type. (11) the value of the hallway type. (12) the value of the structure type.	Residential building

9	(Ahn, et al., 2014)	(1) the number of households. (2) the quantity of gross floor area. (3) the quantity of unit floor households. (4) the number of elevators. (5) the number of floors. (6) the quantity of pilotis with household scale. (7) the number of households of unit floor per elevator. (8) the quantity of the height between storeys. (9) the value of depth of the pit. (10) the value of the roof type. (11) the value of hallway type.	Residential building
10	(Jin, Han, Hyun, & Kim, 2014)	(1) the value of the gross floor area. (2) the value of building coverage ratio. (3) the value of the floor area ratio. (4) the value of the number of households. (5) the number of floor households. (6) the number of floors. (7) the number of elevators. (8) the quantity of pilotis floors. (9) the value of the apartment type. (10) the value of hallway type. (11) the value of the foundation system. (12) the value of the roof type. (13) the value of the structure type.	Residential building
11	(Doğan, et al., 2008)	(1) the value of total building area. (2) the ratio of the typical floor area to the total building area. (3) the value of the ratio of the footprint area to the total area of the building. (4) the number of floors. (5) the value of the type of overhang design. (6) the value of the foundation system. (7) the value of floor structure. (8) the value of the core location.	Residential building
12	(Ji, et al., 2018)	(1) the number of households. (2) the value of building gross floor area. (3) the number of households per unit. (4) the number of elevators. (5) the number of floors. (6) the scale of pilotis with household. (7) the number of households of unit floor per elevator. (8) the quantity of height between storeys. (9) the quantity of depth of the pit. (10) the value of the roof type. (11) the value of hallway type. (12) the value of structure type (RC).	Residential building

Researchers usually only have a one-time opportunity to collect cost predictors (Smith & Mason, 1997). If the predictors selected in the study require a great amount of data collection and preparation time and effort, then they impair the usefulness of the model for early cost estimation (Gardner, et al., 2016). Therefore, the effectiveness and availability of the predictors both need to be considered. A comprehensive literature review of the predictors in previous studies was conducted to provide a clear understanding of the predictors in residential building, as shown in Table 3.6. To well understand the effectiveness of the predictors and maintain the consistency of data in different studies, the attribute weights are transferred by using the following equation:

$$wt_i = \frac{w_i}{\sum_{i=1}^{n} w_i} \quad 3.34$$

where wt_i , w_i are the transferred weight and original weight of the ith invariables in previous studies. Table 3.7 summarized the original attribute weight and transferred attribute weight in previous studies. Despite the extensive input attributes used in previous studies, there are only a handful of effective attributes of residential building (Ji, et al., 2018). These effective attributes contribute to more than 80% of the total attribute weight. The most influencing attributes of previous studies' effective predictors are gross floor area (total construction area), foundation type, no. of storeys below the ground.

Table 3.7 Weight and transferred weight of predictors

Reference	Predictors	Weight	Transferred Weight	Weight method
(Kim, et al., 2005)	the value of location,	[0.3296,	[0.3296,	MRA
	the value of area,	0.0029,	0.0029,	
	the value of roof type,	0.0426,	0.0426,	
	the value of total unit,	0.0071,	0.0071,	
	the value of average area of unit,	0.0243,	0.0243,	
	the value of foundation types,	0.3029,	0.3029,	
	the value of basement,	0.2489,	0.2489,	
	the value of duration	0.0414]	0.0414]	
(Doğan, et al., 2008)	the value of ratio of floor area to total area,	[1	[0.25,	Binary dtree;
	the value of overhang design,	1	0.25,	
	the value of core location,	1	0.25,	
	the value of foundation system	1]	0.25]	
(Doğan, et al., 2008)	the value of total area,	[0.387129,	[0.10671,	Info dtree,
	the value of ratio of floor area to total area,	0.451902,	0.12457,	
	the value of ratio of footprint area to total area,	0.439009,	0.12101,	
	the value of number of floors,	0.355676,	0.09804,	
	the value of overhang design,	0.509398,	0.14042,	
	the value of core location,	0.511249,	0.14093,	
	the value of floor type,	0.189805,	0.05232,	
	the value of foundation system	0.783560]	0.21599]	
(Doğan, et al., 2008)	the value of ratio of floor area to total area,	[0.204025,	[0.11115,	Info dtree
	the value of overhang design,	0.243221,	0.13251,	
	the value of core location,	0.604721,	0.32945,	
	the value of foundation system	0.783560]	0.42689]	

(KARAN CI, 2010)	the value of building Construction Cost Index, the value of total construction area, the value of total number of apartment blocks, the value of type of insulation	[0.264787, 0.406413, 0.108421, 0.220376]	[0.264787, 0.406413, 0.108421, 0.220376]	GDM
(Koo, Hong, Hyun, Park, et al., 2010)	the value of plottage area, the value of total floor area, the value of land ratio, the value of floor area ratio, the value of stories below the ground, the value of stories above the ground, the value of parking lots, the value of landscape area, the value of public open space, the value of facility function, the value of site location	[0.114, 0.5730, 0.0027, 0.2019, 0.9320, 0.2426, 0.5583, 0.5142, 0.0212, 0.4293, 0.2940]	[0.02871, 0.14766, 0.00070, 0.05203, 0.24017, 0.06252, 0.14387, 0.13251, 0.00546, 0.11063, 0.07576]	GA
(Ji, et al., 2012b)	the number of beds, the number of floors, the value of gross floor area, the value of unit floor area, the number of underground floors, pit, the value of quarter area ratio, the value of office area ratio, the value of pile foundation, the value of air conditioning	[0.080580, 0.017112; 0.875692; 0.004620, 0.000025, 0.027897, 0.000017, 0.000019, 0.000062, 0.000023]	[0.080580, 0.017112; 0.875692; 0.004620, 0.000025, 0.027897, 0.000017, 0.000019， 0.000062, 0.000023]	GA
(Ahn, et al., 2014)	the number of households, the value of gross floor area, the number of unit floor households, the number of elevators, the number of floors, the value of pilotis with household scale, the value of households of unit floor per elevator, the value of height between stories, the value of depth of the pit, the value of roof type, the value of hallway type	[67.52, 7433.60, 4.87, 0.97, 9.48, 3.88, 2.73, -0.05, 0.07, 0.43, -0.41]	[0.00898, 0.98810, 0.00065, 0.00013, 0.00126, 0.00052, 0.00036, -0.00001, 0.00001, 0.00006, -0.00005]	Attribute impact

(Ahn, et al., 2014)	the number of households, the value of gross floor area, the number of unit floor households, the number of elevators, the number of floors, the number of pilotis with household scale, the households of unit floor per elevator, the height between storeys; the depth of pit; the value of roof type, the value of hallway type	[0.00654, 0.93701, -0.1615, 0.13886, -0.04609, 0.11363, -0.0929, 0.01528, 0.07564, 0.05241]	[0.00654, 0.93701, -0.1615, 0.13886, -0.04609, 0.11363, -0.0929, 0.01528, 0.07564, 0.05241]	MRA
(Ahn, et al., 2014)	the number of households, the gross floor area, the number of unit floor households, the number of elevators, the number of floors, the value of pilotis with household scale, the value of households of unit floor per elevator; the value of height between storeys; the value of depth of pit; the value of roof type, the value of hallway type	[0.019, 0.361, 0.176, 0.004, 0.292, 0.007, 0.041, 0.002, 0.006, 0.007, 0.024]	[0.02023, 0.38445, 0.18743, 0.00426, 0.31097, 0.00745, 0.04366, 0.00213, 0.00639, 0.00745, 0.02556]	GA
(Jin, Han, Hyun, Kim, et al., 2014)	the value of gross floor area, the number of households, the number of elevators, the value of pilotis floors, the value of apartment type, the value of hallway type, the value of foundation system	[0.8396, 0.0764, 0.1071, 0.1220, 0.6730, 0.1809, 0.0222]	[0.41542, 0.03780, 0.05294, 0.06036, 0.33299, 0.08951, 0.01098]	MRA
(Ji, et al., 2018)	the value of gross floor area, the number of pilotis of a household scale	[0.9803, 0.0197]	[0.9803, 0.0197]	GA

It appeared that high influence attributes with low effort in data collection are the most preferred input variables for ECCE (Gardner, et al., 2016). A common feature is that the early stage variables reflecting primary design decisions have a bigger impact on the eventual building price than variables reflecting the more detailed later design decisions (e.g., Kirkham, 2014). The general principle, therefore, is that only

the variables for which information is available in the early design stages and that have a significant impact on cost are to be selected as predominant attributes (Doğan, et al., 2008). Therefore, to guarantee the prediction performance, three most influencing predictors including gross floor area (GFA), foundation type (FY) and no. of stories below the ground (NSBG) are identified. Considering the data availability and research focus, other factors including total above floor area (TAF), total below floor area (TBF), storey (S), no. of stories above the ground (NSAG), duration (D), commencement Date (CD), finish Date (FD), average storey height (ASH), and the total height of building (TH) are also considered. Totally, this study identified the following variables as predictors: GFA, TAFA, TBFA, S, NSAG, NSBG, D, ASH, THB and FT. Table 3.8 summarizes the predictors used in this study.

Table 3.8 Predictors for ECCE

NO.	Abbreviation	Predictors
1	TFA	Total floor area
2	TAFA	Total above floor area
3	TBFA	Total below floor area
4	S	Storey
5	NSAG	No. of stories above the ground
6	NSBG	No. of stories below the ground
7	D	Duration
8	CD	Commencement Date
9	FD	Finish Date
10	ASH	Average storey height
11	THB	Total height of building
12	FT	Foundation type

3.6.3 Sample Size

The sample size is a significant part of preparing cost estimates in practice. Any changes in the historical database may have substantial impacts on the predicted results of the new project. It is also a very fundamental and critical issue in academic research (Marshall, et al., 2013). Numerous qualitative and quantitative studies have explored its effect (Marshall, et al., 2013; Motrenko, et al., 2014; Wolf, et al., 2013). The performance of various ECCE models is greatly affected by sample size (Ji, et al., 2010a). Therefore, it is necessary to take sample size into account when establishing

the ECCE CBR model (Skitmore, 2001; Yeung & Skitmore, 2012; Yeung & Skitmore, 2005)).

Generally, the solution quality of the CBR model increases with case-base size (Haouchine et al., 2008). However, there is no agreement in existing studies of the relationship between sample size and the accuracy of ECCE methods. Previous studies of data-driven ECCE models make no mention of the sample size, even though this appears to plays an important role in ECCE.

To address the research questions of exploring the influence of sample size on CBR, it is necessary to consider the sample size used in previous studies. As shown in Figure 3.6 and Table 3.9, the sample sizes in previous studies range from 20 to 786, with more than half below 100, and only a few above 500. Since this study partly aims to test the hypothesis that CBR has the advantage for long-term use. The sample size should be large enough to provide reliable results. On the other hand, construction cost related to trade secrets are quite confidential in construction companies and agencies. Too many samples could result in inefficiencies. Taking into account the above factors, this study collects more than 1000 cases of residential building.

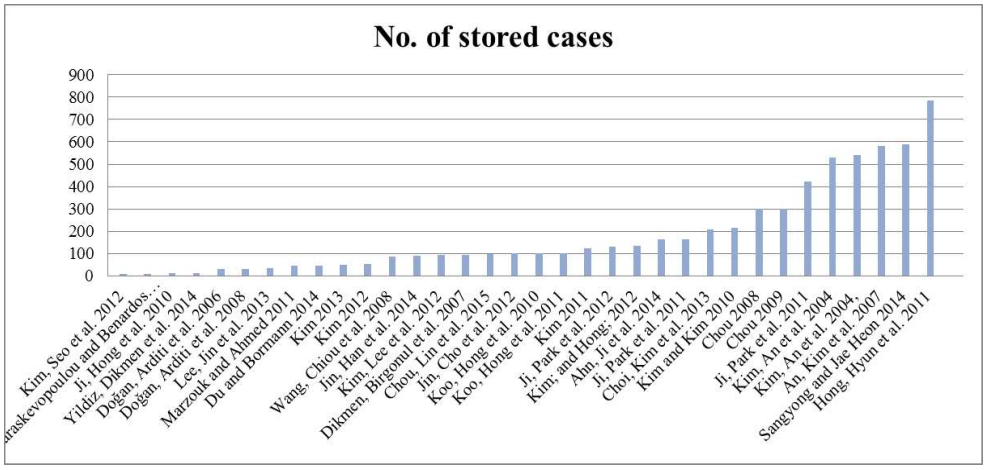


Figure 3.6 Number of stored cases in the case-base in previous studies

Table 3.9 Number of stored cases in the case-base in previous studies

No	Reference	No. of samples	Type of buildings
1	(Kim, Seo, et al., 2012)	8	Large building projects
2	(Ji, et al., 2010c)	9	Multifamily Housing
3	(Yildiz et al., 2014)	13	Construction projects

4	(Doğan, et al., 2006a)	29	Residential buildings
5	(Doğan, et al., 2008)	29	Residential buildings
6	(Jin, Han, Hyun, Kim, et al., 2014)	47	Crafts
7	(Kim & Management, 2013)	48	Highway project
8	(Jin, Han, Hyun, Kim, et al., 2014)	91	Apartment buildings
9	(Ahn, et al., 2017)	99	Multi-family housing
10	(Koo et al., 2011)	101	Multi-family housing projects
11	(Koo, Hong, Hyun, & Koo, 2010)	101	Multifamily Housing
12	(Lee, et al., 2013b)	121	Eco-type pavement trade cases
13	(Kim, 2011)	123	Railroad bridge
14	(Lee, et al., 2013b)	124	Eco-type structural trade cases
15	(Ji, et al., 2012b)	142	Military barrack projects
16	(Tatiya, et al., 2018)	143	Sports fields
17	(Ahn, et al., 2014)	163	Apartment building
18	(Ji, et al., 2012b)	164	Apartment
19	(Leśniak & Zima, 2018)	164	Apartment
20	(Ji, Park, & Lee, 2011)	164	Residential buildings
21	(Du & Bormann, 2014a)	207	Road projects
22	(Kim & Kim, 2010a)	216	Beam bridges
23	(Chou, et al., 2015)	275	Bridge
24	(Wang, et al., 2008)	293	Restoration projects
25	(Chou, 2009)	300	Pavement maintenance project
26	(Kim, 2011)	422	Military facility projects
27	(Kim, et al., 2004b)	530	Residential buildings
28	(An, et al., 2007)	580	Residential buildings

29	(Kim & Shim, 2013b)	590	High-rise building projects.
30	(Hong, et al., 2011)	786	Multifamily Housing

3.6.4 Collection Process

To keep the study aligned with the research philosophy and to maintain compatible logic, the data needs to be carefully collected and well organized. To guarantee the scale and quality of the data, the researcher contacted several construction consulting companies and agencies in China. About 1690 apartment cases from one construction cost consulting company in China were collected. This company is one of the largest cost consulting agencies covering projects in more than 20 provinces in China. It has more than ten years of history and provides a standard cost estimation service.

The data includes total floor area (TFA), total above floor area (TAFA), total below floor area (TBFA), storey (S), no. of stories above the ground (NSAG), no. of stories below the ground (NSBG), duration (D), commencement date (CD), finish date (FD), average storey height (ASH), the total height of the building (AHB), and foundation type (FT).

3.6.5 Cross-validation Algorithm

Cross-validation is a model validation technique used for evaluating how the results of the model will generalize to other data sets. It is widely used in the prediction model for estimating how accurate this model could be in practice. In a prediction problem, a model is established by using a known data set (training dataset) where the training process is conducted. An unknown dataset (validation dataset or testing set) set is used to test the model's performance. Cross-validation aims to evaluate the model's performance on a new dataset, which has not been used in estimating to avoid overfitting and selection bias.

The cross-validation method techniques can be classified into one round cross-validation technique and multiple round validation techniques. One round of cross-validation technique splits the sample dataset into complementary subsets. The model is developed based on one subset (training set) and evaluated on the other subset (called the validation set or testing set). Usually, one round of cross-validation technique selects approximately `P*N` samples to hold out as the test set. N is the total

number of the sample used in this study. P is the proportion of samples to hold out for the testing set. P must be a scalar between 0 and 1. The most widely used P is 20% or 25%, corresponding to one-fifth or one-fourth holdout sample.

Multiple rounds of cross-validation techniques use different partitions and the test results are averaged over the rounds to provide an overall estimate of the model's prediction performance. The k-fold cross-validation is widely used to reduce bias with respect to the random sampling of training and to test data samples (Hastie et al., 2009). Several studies have confirmed that K-fold cross validation optimizes the computation time (Han et al., 2011; Kohavi, 1995).

Basically, k-fold cross-validation uses part of the available data to fit the model, and a different part of testing it. In K-fold, the data is randomly split into K separate sets of equal size, as shown in Figure 3.7. The value of K may differ due to the research problem and aims (Hastie, et al., 2009). To avoid repetition with K in K-NN, the KF is used to represent the value of K in K-fold. The cross-validation is conducted in the following steps:

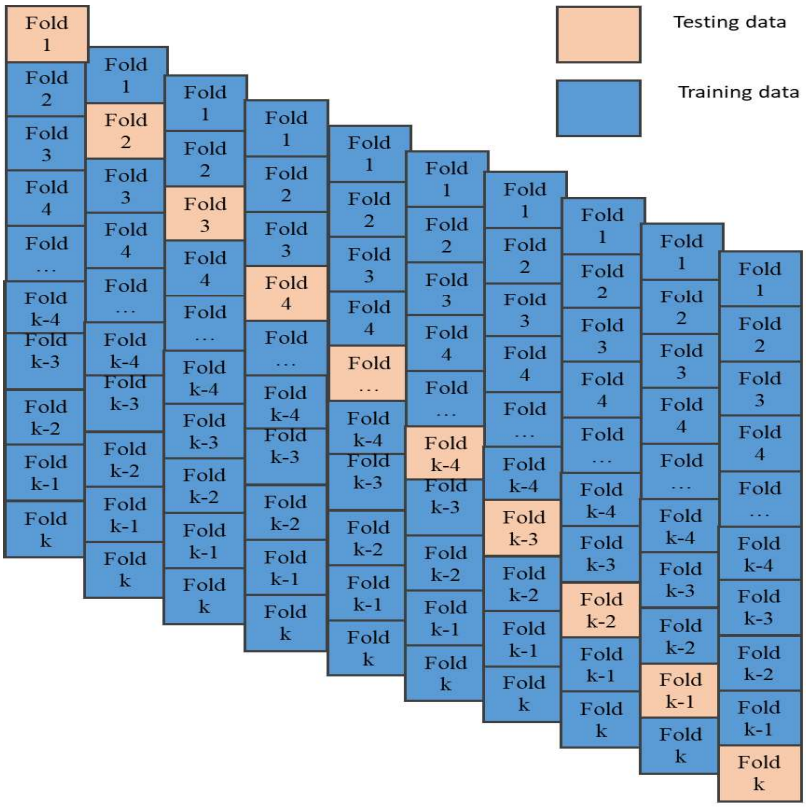


Figure 3.7 K-fold cross validation algorithm

(1) Split the training data T into KF sets (folds) with equal size T_{kf};

(2) For each KF = (1,2,… , kf-1, kf) fit the model $\widehat{f_{kf}}$ to the training set excluding the kth fold;

(3) Compute the predicted values $predicted_i^k = \widehat{f_{kf}}(x)$ for the observations in the kth fold, based on the training data that excluded this fold;

(4) The cross-validation of the error rate of MAPE and RMSE for a set T_k is calculated by the following equation:

$$\text{MAPE}_{kf} = \frac{1}{N}\sum_{i=1}^{n} \frac{\left|actual_i^{kf} - predicted_i^{kf}\right|}{predicted_i^{kf}} \times 100\% \quad 3.35$$

$$\text{RMSE}_{kf} = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(Log\left(\left|actual_i^{kf}\right|\right) - Log\left(\left|predicted_i^{kf}\right|\right))^2} \quad 3.36$$

(5) The overall cross-validation errors are:

$$\text{MAPE} = \frac{1}{kf}\sum_{i=1}^{kf} \text{MAPE}_{kf} \quad 3.37$$

$$\text{RMSE} = \frac{1}{kf}\sum_{i=1}^{kf} \text{RMSE}_{kf} \quad 3.38$$

3.7 CHAPTER SUMMARY

This chapter discovered the research design, considering the research approach, research methodology, research methods, model development, and data collection. The research philosophy, approach to theory development, methodological choice and strategies, timeline horizon, techniques, and procedures are presented in the research approach. Positivist and pragmatism are used in research philosophy. Induction and deduction are used as the approach for theory development. Mixed methods are used in methodology choice and strategy. A cross-sectional feature is adopted in the time horizon.

In the research framework, the main research question of how to improve the CBR model of ECCE for long-term use is proposed. The critical factors are identified to answer the research question by conducting a literature review on exploring the influence of sample size, improving the robustness, and the efficiency of the ECCE CBR model. Three research hypotheses are formulated to address the main research issue: that accuracy will be improved with the increase in the size of the case-base; a

robust weight determination will improve the robustness of the ECCE CBR model; and a CBM strategy for editing the case-base will improve the efficiency of ECCE.

Research methods, including CBR, GA, OLS, MODAL regression, and CBM strategy, are illustrated in the research methods. The case representation, case retrieval, weight determination, and case adaptation are illustrated in CBR. Two error measures, MAPE and RMSE, are used to evaluate the prediction performance, and three weight determination methods comprising GA, MRA, and feature counting are chosen for weight determination. A comparative study of different combinations of CBR settings based on different sample sizes is conducted. The basic concepts of linear regression are introduced, and modal linear regression is explained, followed by the illustration of the modal EM algorithm and bandwidth selection. After introducing modal linear regression, a CBM strategy is proposed for case-base editing. The basic concepts of CBM are illustrated, followed by the introduction of the weighted coverage contribution and case-base editing methods.

Model development represents the process of developing the classic CBR model and optimization model. The classic CBR model explores how sample size influences the performance of the CBR model and provides a clear understanding of the advantages and disadvantages of each setting. Two optimization algorithms with respect to robustness and efficiency are proposed to improve the long-term performance of the ECCE CBR model. The performance of the CBR model is evaluated before and after using the proposed methods. Data collection is stated from project type, predictor variables, and sample size and collection process, followed by the timeline and limitations. Altogether, this chapter provides the overall research design by considering the research approach, research methodology, research methods, model development, and data collection.

To address the proposed research questions well, the research tasks are broken into several parts: identifying the research problem, conducting the literature review, choosing the methodology, planning and conducting the data collection, developing and optimizing the model. Accordingly, this thesis includes six components: abstract, introduction, literature review, methodology, results and discussion, and conclusion. The abstract provides a comprehensive summary of this study. The introduction provides the research background and research problem, identifying the research aims, objectives, and significance. The literature review offers a clear understanding of the

existing research related to research topics. Methodology includes the methodological considerations for solving the research questions. It consists of the research approach, research design, research methods, and the development of the CBR model. Results and discussions consist of the research findings and analysis from the classic CBR model and the optimized CBR model. The conclusion involves a summary of the research contribution, research limitation, and recommendation for future research.

Chapter 4: Data Analysis

4.1 INTRODUCTION

This chapter presents the data pre-processing, data summary, and the selection of predictor variables. Data pre-processing deals with the data preparation for improving the data quality. Descriptive statistics provide the introduction of the data used in the model development. The selection of predictors' variables identified the final variables for subsequent model development.

The data collection process is usually loosely controlled, which may cause missing values, out-of-range values, and inconsistent values. Accordingly, data pre-processing involves operations for data cleaning, data transformation, and data reduction, as shown in Figure 4.1. Data pre-processing is a significant procedure before model development and data analysis. The widely used phrase "garbage in, garbage out" refers to the concept that flawed input data would lead to nonsense output. If data has not been carefully screened, the analysis may produce misleading results. Accordingly, the representation and quality of data is first and foremost before model development and result analysis. Data cleaning aims to address the missing value and out-of-range values of data. Data transformation in this study involves data standardization. Data reduction involves the operation of selecting and extracting the final features for the model development.

This chapter begins with data cleaning, which deals with missing values of data and out-of-range data. Section 4.2 introduces the data transformation concerning data scale and time. Section 4.4 represents the description of the data after removing missing value and outliers in the data. Section 4.5 illustrates how the final predictors for developing the model are determined in this study. Note that data pre-processing may influence the way in which outcomes of the final data processing can be interpreted. This influence should be carefully considered when the interpretation of the results.

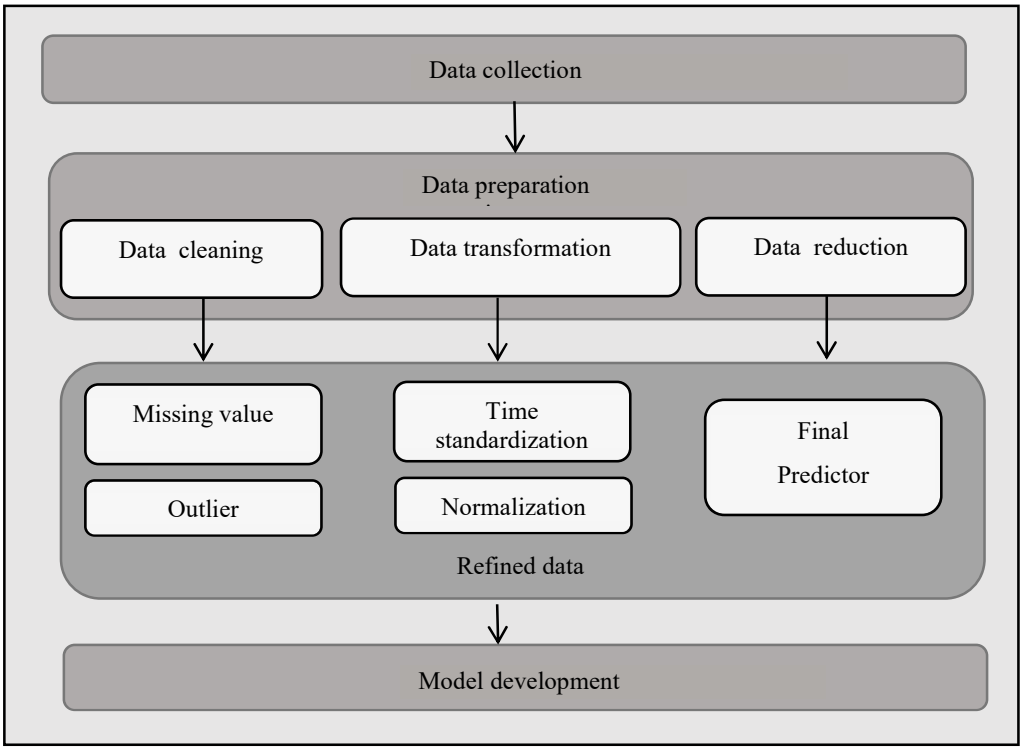


Figure 4.1 Data pre-processing

4.2 DATA DESCRIPTION

Cost data from a total of 1640 apartment building projects completed between 2006 to 2015 from six provinces (Guangzhou, Beijing, Shanghai, Chongqing, Henan, Shaanxi) in China were collected for model development. Figure 4.2 illustrates the distribution of the project locations. The data collected includes gross floor area (GFA), total above floor area (TAFA), total below floor area (TBFA), storey (S), no. of stories above the ground (NSAG), no. of stories below the ground (NSBG), duration (D), commencement Date (CD), finish Date (FD), average storey height (ASH), the total height of the building (THB), and foundation type (FT). Table 4.1 provides a descriptive summary of the raw data collected for the model development. To guarantee the quality of research, data pre-processing of data cleaning, data transformation, and data reduction is conducted before developing the model.

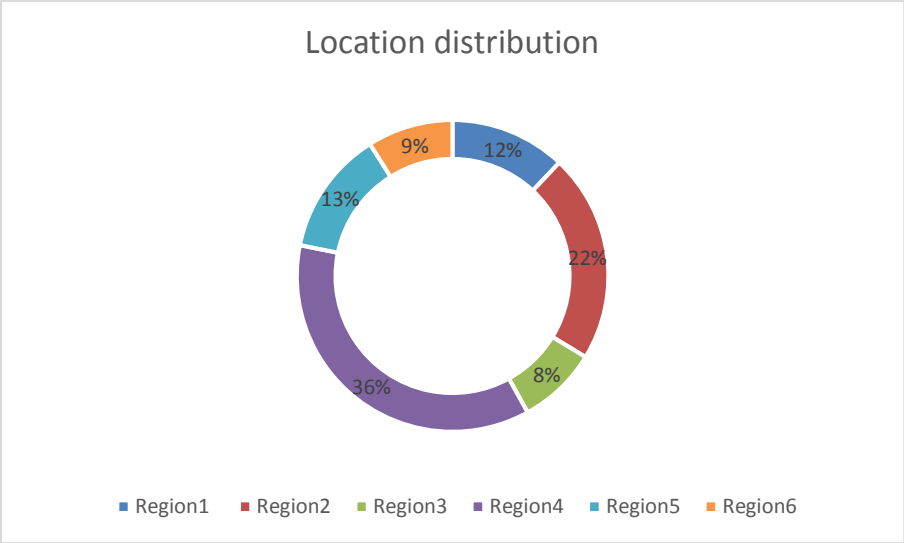


Figure 4.2 Location distribution of collected cases

Table 4.1 Information contained in the cases

1	Location type	Categorical	Region 1, 2,3,4,5,6
2	*Building type*	Categorical	Apartment
3	*Total floor area*	Numerical	186~ 92,472.8 m2, NA
4	*Total above floor area*	Numerical	186~82,181.95 m2 s NA
5	*Total below floor area*	Numerical	0~40038.30 m2 s, NA
6	*Storey*	Numerical	6~52 floors, NA
7	*No. of storeys above the ground*	Numerical	6~47 floors, NA
8	*No. of storeys below the ground*	Numerical	0~5 floors, NA
9	*Storey height*	Numerical	2.0~5.8 m, NA
10	*Basement height*	Numerical	1.0~6.0 m, NA
11	*Commencement date*	Date	2008-2015, NA
12	*Finish date*	Date	2009~2017, NA
13	*Duration*	Numerical	63~1515 days, NA
14	*Total height of building*	Numerical	13.4~138 m
15	Foundation type	Categorical	Concrete pile; Concrete pile and bolt support, NA
16	Total amount of contract	Numerical	CNY 624,365~151,599,908

NA represents the missing value.

4.3 DATA CLEANING

4.3.1 Missing Data

Missing data are valid omitted values on one or more variables that are not available for analysis. Generally, research analysis requires completed information for

each case. Missing data may have a practical and substantive influence on the research result and analysis. The practical implication of dealing with missing data may result in sample size reduction, while the substantive implication of that may result in biases in the research finding (Allison, 2001). Like all other studies, missing data should be adequately handled before analysis (Hair et al., 2017, p. 48). To ensure the quality of the data, it is necessary to conduct data cleaning. Data cleaning is a process of dealing with any missing values in the dataset. Missing data is the absence of one or more values of the variables in the data set (Bannon Jr, 2015). It might be caused by incomplete information stored in the database or data entry errors.

Since missing data has an influence on researching the findings, data-screening is developed for handling missing data. Data screening checks and finds the errors in variables, and then fixes or deletes the error values in the data file (Pallant, 2013). Data-screening can also be classified based on how they are being handled: case deletion, single imputation, and multiple imputation (Scheffer, 2002). Case deletion is widely used for handling missing value. The cases containing missing data are deleted. It can be either complete case only or pairwise-available case. Single imputation may use the value of group means, medians or modes to replace the missing value, or use regression imputation, stochastic regression imputation, or expectation maximisation algorithm imputation to predict the missing value, or use last value carried forward for longitudinal data to replace the missing values. Multiple imputations. Multiple imputation refers to predict missing value by multiple imputing. When the imputation model fails to converge, multiple imputation can be based on propensity scoring. Bayesian MI uses the non-informative algorithm to generate m separate datasets to estimate the posterior distribution for random extraction. End-users may avoid multiple imputations because multiple dataset computation is difficult and time-consuming, even for statisticians (Scheffer, 2002). The total percentage of missing values is approximately 8 % in this study. Since the percentage of missing values is small, removing these cases is deemed to have little influence on this study. Therefore, case deletion is used for handing missing values. The cases that have missing values in the attributes are deleted. Totally, there are 131 cases that have been deleted.

4.3.2 Outliers

After dealing with the missing value in the data set, the other issue of data cleaning is the outlier. Outliers are the extreme values in a data set that are located well

above or well below in comparison with other values, far away from the mean (Pallant, 2013). They are observations outside the bulk of the data and usually have a unique combination of characteristics (Hair et al., 2010). The outlier may have a practical and substantive impact on the data analysis. From the practical perspective, outliers may influence data analysis by impacting variance, while from the substantive perspective, outliers can be deemed as how representative they are of the population (Hair, et al., 2010). They are inconsistent with the majority of the data and are usually higher or lower than other observations, which affect the mean and the variance.

Outliers can be a valid value that represents the extreme value, or it can result from the failure of non-compliance on the part of the respondent, or due to the methodological error on the part of the researcher. Outliers can occur due to procedural errors, extraordinary events or observations, or the combination of all (Hair, et al., 2010). Whatever the reason for the occurrence of outliers, these can't be broadly classified as beneficial or problematic but rather looked into from the perspective of what insight they provide for the research. The beneficial outliers provide information about the population that might not be collected in the normal course, and the problematic outliers are not representative of the population and can distort results (Hair, et al., 2010). Classification of the observations as beneficial or problematic is determined by their effect on the data. Due to the impact of the outliers, it is important that the researcher should identify outliers.

There is no single and reliable method for identifying outliers. Although more than 50 different tests are used for this purpose, it is not rare that two or more tests applied to the same data set can yield different outliers (Sheskin, 2010). Some tests only identify a single outlier, whereas others can identify multiple outliers. It is a common practice for detecting outliers by determining an interval spanning over the mean plus/minus three standard deviations in practice (Howell et al., 1998; Leys et al., 2013). Therefore, this study uses this method to identify the outlier in the data set. In this method, outliers are defined as elements of more than three standard deviations from the mean.

Once identified, outliers can be addressed by several methodologies: trimming; winsorization; modified winsorization; semi-winsorization; and deleting. Trimming focuses on removing 10% of extreme data from both the tales, whereas, in winsorization, a fixed number of extreme values in the distribution are replaced with

the score that is in close proximity to the outliers in the tails (Sheskin, 2010). Modified Winsorization allows transforming only one data point in the distribution, and semi-winsorization allows converting the outlier/s, not to the nearest value but to a value representing the pre-determined number of standard deviations away from the mean (Reifman & Keyton, 2010).

The decision about retaining or handling outliers is based on the assertions of what information they provide about the population/sub-population. The researcher should decide about the retention or exclusion of the outlier not only on the characteristic of the outlier as beneficial or problematic but also on the type of the information they provide within the objective of the analysis (Hair, et al., 2010). Since retaining the outliers of the population can undermine the findings, in this study, identified outliers are deleted to avoid the harmful effects on model development. Totally, there are 161 cases with missing values, and 29 cases with outlier values are deleted. Table 4.2 provides a summary after the data cleaning.

Table 4.2 Case information after data cleaning

1	Location type	Categorical	Region 1, 2,3,4,5,6
2	*Building type*	Categorical	Apartment
3	*Total floor area*	Numerical	359.6~ 38,871 m^2
4	*Total above floor area*	Numerical	269~37,349m^2
5	*Total below floor area*	Numerical	0~8875.39m^2
6	*Storey*	Numerical	6~47 floors
7	*No. of storeys above the ground*	Numerical	6~47 floors,
8	*No. of storeys below the ground*	Numerical	0~4 floors,
9	*Storey height*	Numerical	2.0~5.8 m,
10	*Basement height*	Numerical	1.0~6.0 m,
11	*Commencement date*	Date	2008-2015,
12	*Finish date*	Date	2009~2017
13	*Duration*	Numerical	92~1405 days,
14	*Total height of building*	Numerical	16.8~138 m
15	foundation type	Categorical	Concrete pile; Concrete pile and Bolt support
16	Total amount of contract	Numerical	CNY706,940~64,541,256.13

4.4 DATA TRANSFORMATION

4.4.1 Time Standardization

To minimize the diverse characteristics and differences in the collected data, this study firstly conducts the time standardization. Accordingly, the time and regional differences in the data need to be eliminated. This study converted the cost data of all project information to an identical point of time (2015) by using the *price indices of construction and installation* from the National Bureau of Statistics of China. (Ji, et al., 2010a). This index evaluates the changes in the purchase prices of major construction materials, chemical materials, labor, equipment, and tools. The cost index applied to the conversion was official statistical data prepared to estimate the price fluctuation of input resources by 100 times scale as the price of the construction cost of a project at a certain point in time. Regional differences are illuminated by using regional *Price Indices of Construction and Instalment*. The detailed source of the *Price Indices of Construction and Instalment* can be obtained from the website of the National Bureau of Statistics. Table 4.3 summarizes the price indices of construction and installation from 2010 to 2015.

Table 4.3 Price indices of construction and instalment

	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
2006	99.7	100.1	101.5	100.7	101.1	103.5
2007	104.1	104.6	106.3	103.8	106	105.6
2008	111.8	112.1	112.1	112.2	113.7	113.3
2009	94.3	94.8	94.6	95.3	97	99.1
2010	104	106.1	104.9	104.3	102.7	105.3
2011	109.7	110.6	110.1	108	107.8	107.9
2012	99	98.7	101.4	101.9	102.1	103.4
2013	97.3	99.8	99.8	101.9	100.5	102.3
2014	98.5	100.3	100.1	102	100.4	101.2
2015	94.4	94.9	96.5	98.4	97.5	98.4

The price in last year equals 100.

The cost data of all the cases were converted to the year of 2015 cost level. For example, 2006 data in Region 1 were converted into 2007 data by multiplying 2006 cost data by the value (104.1/100) calculated by dividing 104.1 (the index value for 2007) by 100 (the absolute index in 2006). Appendix A provides the detailed conversion of the index to 2015. Table 4.4 summarizes the converted index.

Table 4.4 The converted index

Year	Transferred year	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
2006	2015	1.1215	1.2230	1.2727	1.3030	1.3000	1.4190
2007	2015	1.0773	1.1692	1.1973	1.2553	1.2265	1.3438
2008	2015	0.9636	1.0430	1.0680	1.1188	1.0787	1.1860
2009	2015	1.0219	1.1002	1.1290	1.1740	1.1120	1.1968
2010	2015	0.9826	1.0370	1.0763	1.1256	1.0828	1.1366
2011	2015	0.8957	0.9376	0.9775	1.0422	1.0045	1.0533
2012	2015	0.9047	0.9499	0.9640	1.0227	0.9838	1.0187
2013	2015	0.9298	0.9518	0.9660	1.0037	0.9789	0.9958
2014	2015	0.9440	0.9490	0.9650	0.9840	0.9750	0.9840

4.4.2 Normalization

To handle categorical variables, regression analysis is used to convert it into dummy variables (Keller, 2015). For example, the foundation type variable having two categories (e.g., Concrete pile foundation and Mixed pile foundation) can be converted into the dummy variable, as shown in Table 4.5.

Table 4.5 Converting the foundation type into dummy variables

Foundation type	Dummy variables
Concrete pile foundation	0
Mixed pile foundation (Concrete pile foundation and Bolt support)	1

After transforming the categorical variable into dummy variables, data normalization should be conducted. Different attributes may have different units. Therefore, this study minimizes the inconsistent units of the measurement in the attributes by using Equation 4.1.

$$\mathbf{X}_{\mathrm{n}_i} = \frac{\mathbf{X}_{\mathrm{0}_i}}{\max\left(\mathrm{X}_{\mathrm{0}_i}\right) - \min\left(\mathrm{X}_{\mathrm{0}_i}\right)} \tag{4.1}$$

Where $\mathbf{X}_{\mathrm{n}_i}$ represents the normalized value of the *ith* attribute; $\mathbf{X}_{\mathrm{0}_i}$ represents the original attribute value.

4.5 SELECTION OF PREDICTOR VARIABLES

Pearson correlation (PC) analysis is used for the final selection of variables (Ahn, et al., 2014; Hong, et al., 2011). The Pearson correlation matrix of the variables is shown in Table 4.6. Foundation type is excluded from the input variables because of its low correlation with total cost. To reduce multicollinearity, total floor area, storeys and average storey height are excluded. Finally, this research identified six variables as significant (at the 1% level): total above floor area, total below floor area, no. of stories above the ground, no. of stories below the ground, duration, total height of building.

Table 4.6 Pearson correlation analysis of the input attributes

Correlations		TC	FT	TFA	TAFA	TBFA	S	NSAG	NSBG	ASH	D	THB
TC	P C	1.00	-0.04	.942	.932	.493	.774	.767	.386	-.151	.318	.770
	Sig.		0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FT	P C	-0.04	1.00	-.089	-.095	0.00	-.129	-.133	-0.01	-0.01	-.093	-.136
	Sig.	0.14		0.00	0.00	0.89	0.00	0.00	0.66	0.76	0.00	0.00
TF	P C	.942	-.089	1.00	.994	.495	.783	.779	.354	-.162	.346	.785
	Sig.	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
TAF	P C	.932	-.095	.994	1.00	.406	.787	.789	.288	-.159	.345	.795
	Sig.	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00	0.00
TBF	P C	.493	0.00	.495	.406	1.00	.313	.263	.688	-.101	.169	.264
	Sig.	0.00	0.89	0.00	0.00		0.00	0.00	0.00	0.00	0.00	0.00
S	P C	.774	-.129	.783	.787	.313	1.00	.996	.430	-.306	.416	.986
	Sig.	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00	0.00
SA	P C	.767	-.133	.779	.789	.263	.996	1.00	.357	-.303	.408	.990
	Sig.	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00	0.00
BS	P C	.386	-0.01	.354	.288	.688	.430	.357	1.00	-.164	.240	.351
	Sig.	0.00	0.66	0.00	0.00	0.00	0.00	0.00		0.00	0.00	0.00
ASH	P C	-.151	-0.01	-.162	-.159	-.101	-.306	-.303	-.164	1.00	-.204	-.239
	Sig.	0.00	0.76	0.00	0.00	0.00	0.00	0.00	0.00		0.00	0.00
D	P C	.318	-.093	.346	.345	.169	.416	.408	.240	-.204	1.00	.402
	Sig.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00		0.00
TH	P C	.770	-.136	.785	.795	.264	.986	.990	.351	-.239	.402	1.00
	Sig.	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

Correlation is significant at the 0.01 level (2-tailed).

4.6 CHAPTER SUMMARY

Data pre-processing involves operations for data cleaning, data transformation and data reduction, as shown in Figure 4.1. It is a significant procedure before model development. This chapter begins with data cleaning, which deals with missing values of data and out-of-range data. Section 4.2 introduces the data transformation with respect to data scale and time. Section 4.4 represents the description of the data after removing missing values and outliers in the data. Section 4.5 illustrates how the final predictors for developing the model are determined in this study.

After the data pre-processing, six variables including total above floor area, total below floor area, no. of stories above the ground, no. of stories below the ground, duration and total height of building are identified as the final predictors in this study. Totally 1450 cases were retained after the data pre-processing. Note that data pre-processing may influence the way in which outcomes of the final data processing can be interpreted. This influence should be carefully considered when interpreting the results.

Chapter 5: Comparative Study of Existing CBR Models

5.1 INTRODUCTION

In this chapter, the most widely used weight determination methods and similarity functions are compared with respect to different sample sizes. This chapter aims explicitly to shed light on the questions: (1) What influence does sample size have on the efficiency of CBR? (2) What changes in accuracy occur with increasing sample size? and (3) Does CBR have an advantage in terms of long-term use in ECCE?

This chapter explores the influence of sample size on the accuracy of the CBR. From this, the differences in the different parameter settings of various sample sizes are explored as well as their general trends as size continues to increase. This chapter provides a comparison among different weight determination methods and similarity functions. Three weight determinations (MRA(W1), GA(W2), FC(W3) and two similarity functions (weighted sum of the attribute distance (S1) and 1-Euclidean distance (S2)), and three case adaptation values (5,3,1) are used in this chapter. Leave one cross validation is used for validating the model.

Totally, 1450 Chinese apartment building projects were retained in the dataset in this study. Given the sample size range in previous studies, the training sample sizes of 50, 100, 200, 400,600, 800, and 1000 are used. The training samples were randomly selected from the database, with a 20% project hold-out sample. This process is repeated ten times. The MAPE and RMSE are used as measures of ECCE CBR models' accuracy and are computed for each of the ten trials. The comparison of settings in CBR, justification the sample size and the data-oriented results are discussed in the discussion.

5.2 RESULTS

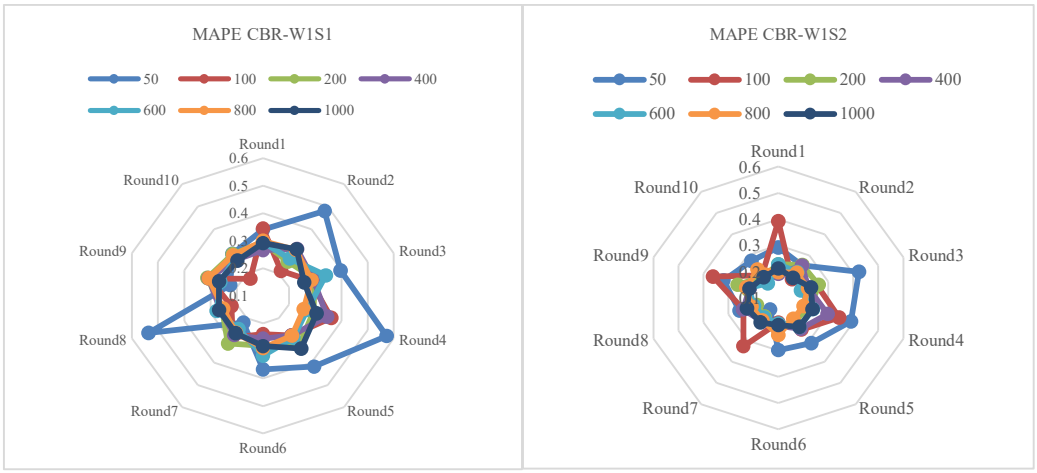
5.2.1 MAPE and RMSE

Table 5.1, for example, gives the results of the 50-size sample (K=5). MAPE and RMSE vary greatly during the 10 trials. For CBR-W1S1, the MAPE ranges from 18.39% and 61.57%; and RMSE ranges from 20.56% to 70.39%. For CBR-W2S1, the MAPE ranges from 18.96% and 49.58%; and RMSE ranges from 21.7% to 49.49%. For CBR-

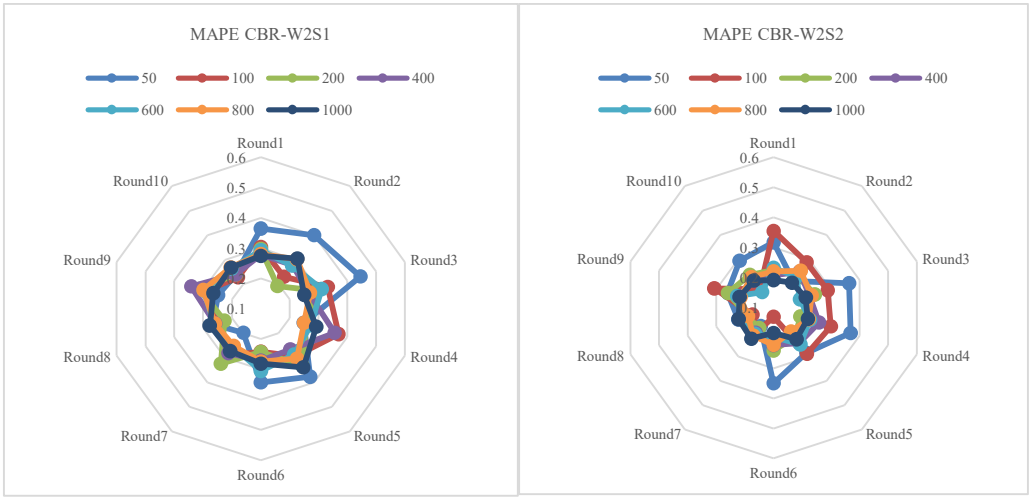
W3S1, the MAPE ranges from 26.55% and 56.28%; and RMSE ranges from 27.65 % to 64.12%. For CBR-W1S2, the MAPE ranges from 17.44 % and 47.08%; and RMSE ranges from 20.99% to 38.32%. For CBR-W2S2, the MAPE ranges from 17.25% and 36.81 %; and RMSE ranges from 20.27% to 38.32%. For CBR-W3S2, the MAPE ranges from 19.31% and 65.63%; and RMSE ranges from 21.42 % to 65.63 %. The results of remaining parameter settings can be seen in Appendix B-1 to B-20.

Table 5.1 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2，CBR-W3S2 (50-size sample and K=5)

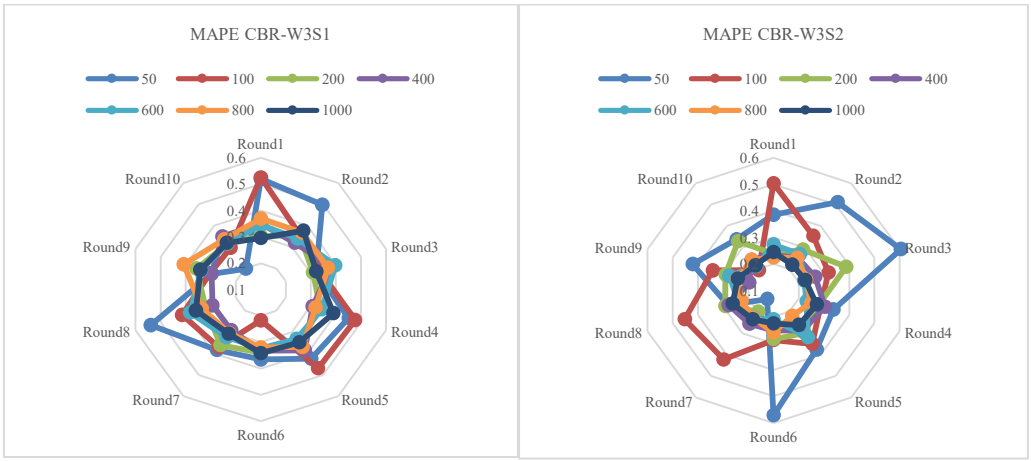
	MAPE					
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	25.68%	23.27%	34.54%	27.66%	23.40%	32.05%
Round2	61.57%	49.58%	56.28%	47.08%	36.81%	68.16%
Round3	42.67%	43.44%	39.57%	31.87%	28.33%	38.73%
Round4	52.83%	32.40%	55.25%	47.47%	19.70%	49.01%
Round5	42.44%	45.81%	60.41%	35.59%	28.76%	49.19%
Round6	41.78%	41.70%	48.81%	46.68%	29.78%	48.10%
Round7	35.49%	26.84%	44.68%	26.07%	25.02%	40.72%
Round8	39.60%	26.97%	38.99%	35.08%	25.55%	37.85%
Round9	28.09%	31.68%	43.91%	35.11%	29.40%	42.58%
Round10	18.39%	18.96%	26.55%	17.44%	17.25%	19.31%
Average	38.85%	34.06%	44.90%	35.01%	26.40%	42.57%
	RMSE					
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	27.51%	25.92%	35.57%	28.52%	24.91%	34.63%
Round2	70.39%	49.49%	64.12%	49.55%	38.32%	65.63%
Round3	39.96%	41.01%	39.33%	32.98%	31.13%	37.47%
Round4	46.52%	30.86%	47.58%	43.25%	23.36%	44.66%
Round5	41.91%	45.61%	62.95%	39.38%	31.83%	48.94%
Round6	45.20%	40.96%	53.29%	48.41%	31.98%	51.48%
Round7	35.66%	28.44%	42.80%	27.64%	26.84%	39.14%
Round8	63.43%	30.81%	66.71%	53.12%	32.60%	51.91%
Round9	31.85%	35.81%	46.40%	39.42%	33.91%	45.29%
Round10	20.56%	21.70%	27.65%	20.99%	20.27%	21.42%
Average	42.30%	35.06%	48.64%	38.33%	29.52%	44.06%



(a) (b)

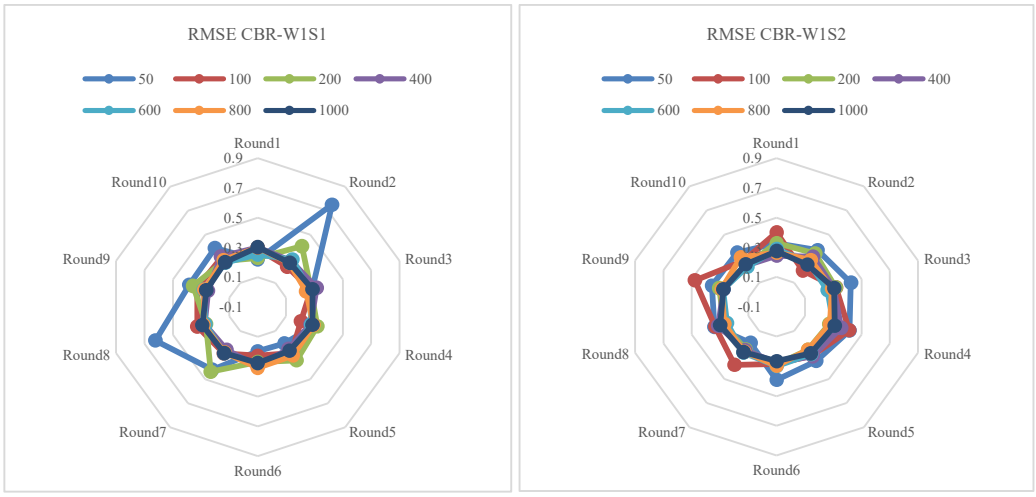


(c) (d)



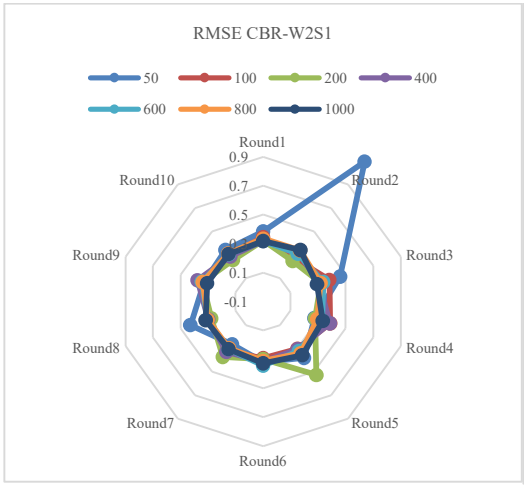
(e) (f)

Figure 5.1 Comparative results of MAPE (K=5)

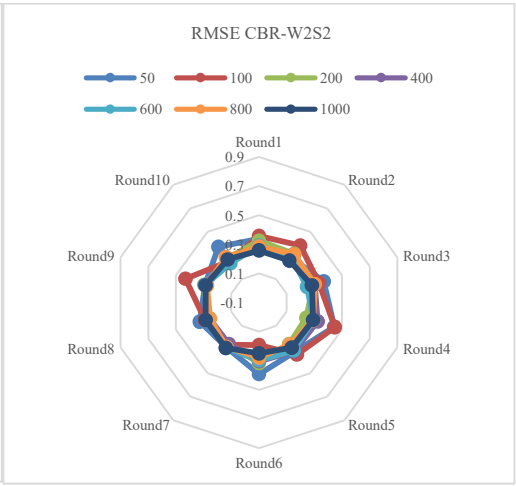


(a)

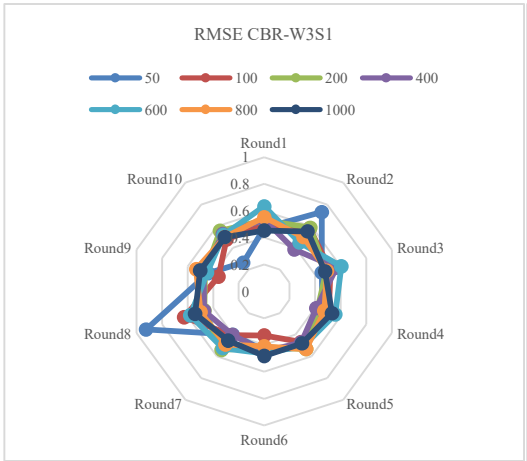
(b)



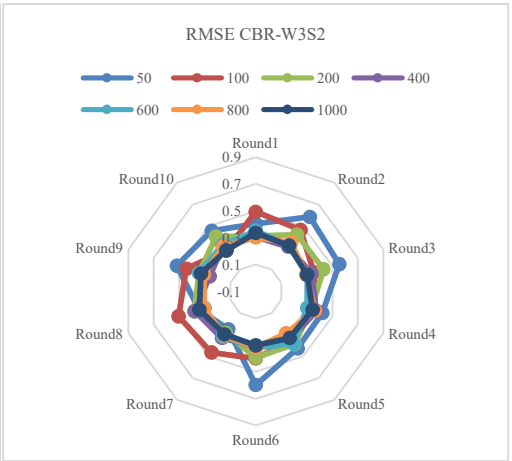
(c)



(d)



(e)



(f)

Figure 5.2 Comparative results of RMSE (K=5)

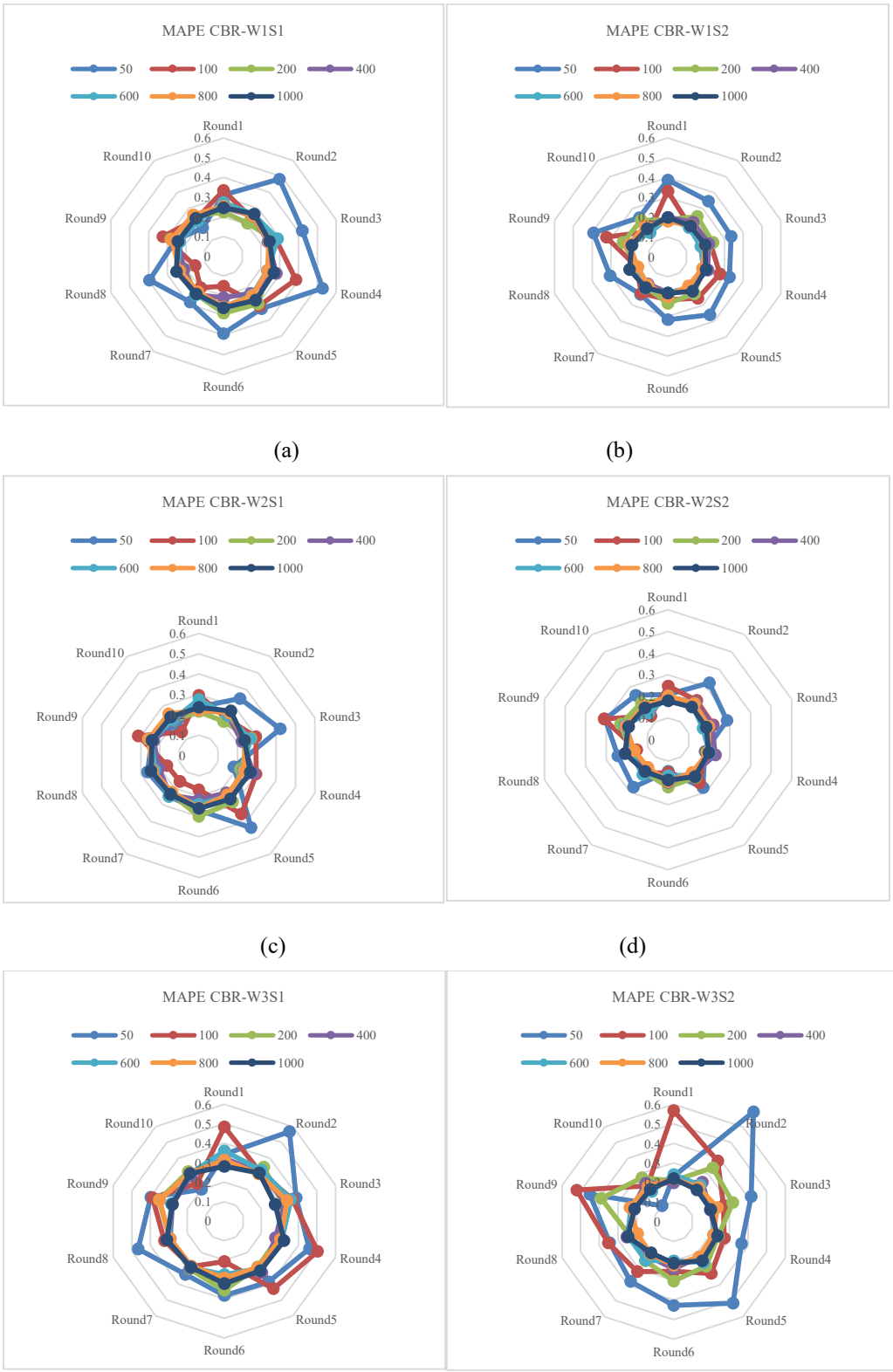


Figure 5.3 Comparative results of MAPE (K=3)

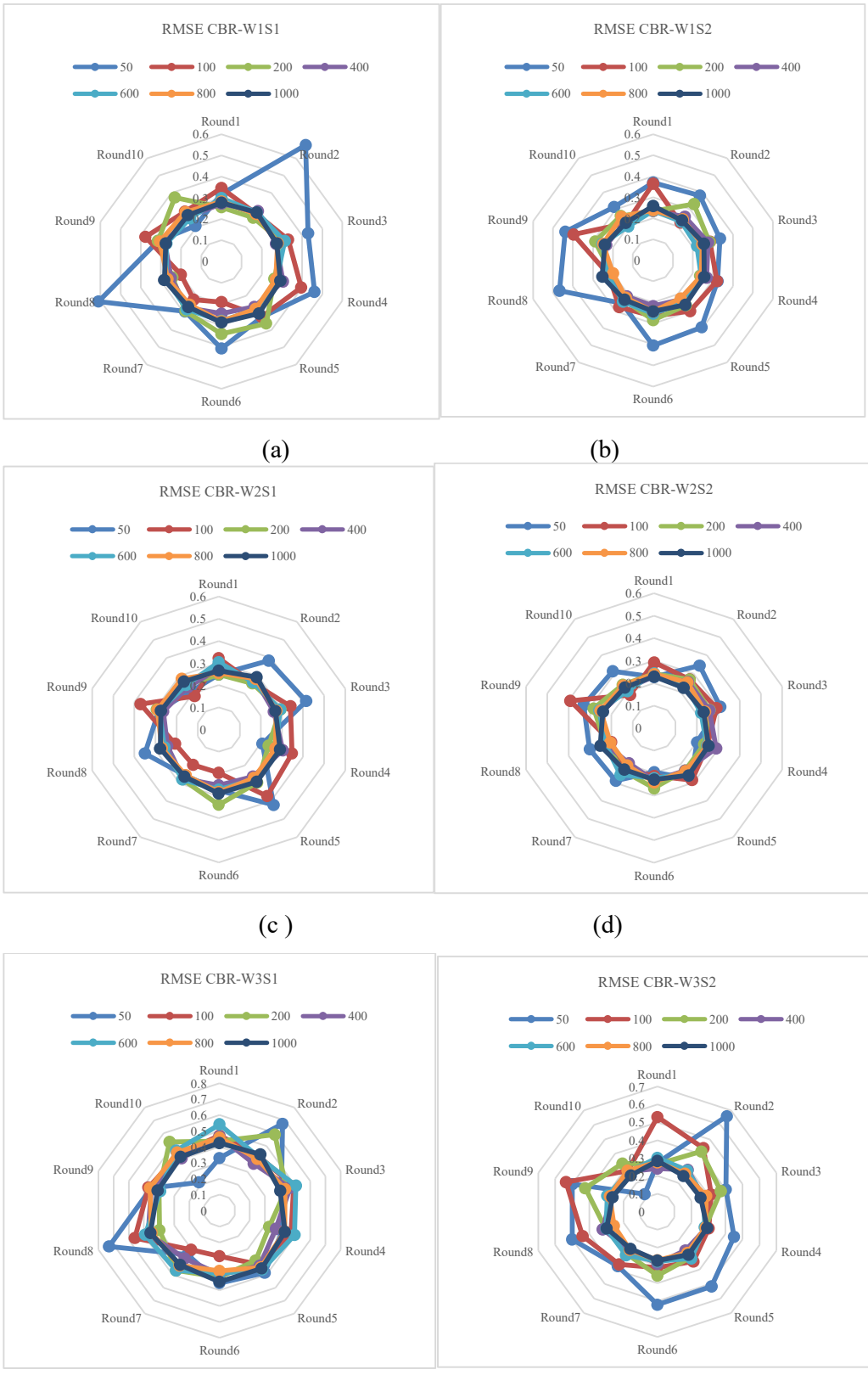


Figure 5.4 Comparative results of RMSE (K=3)

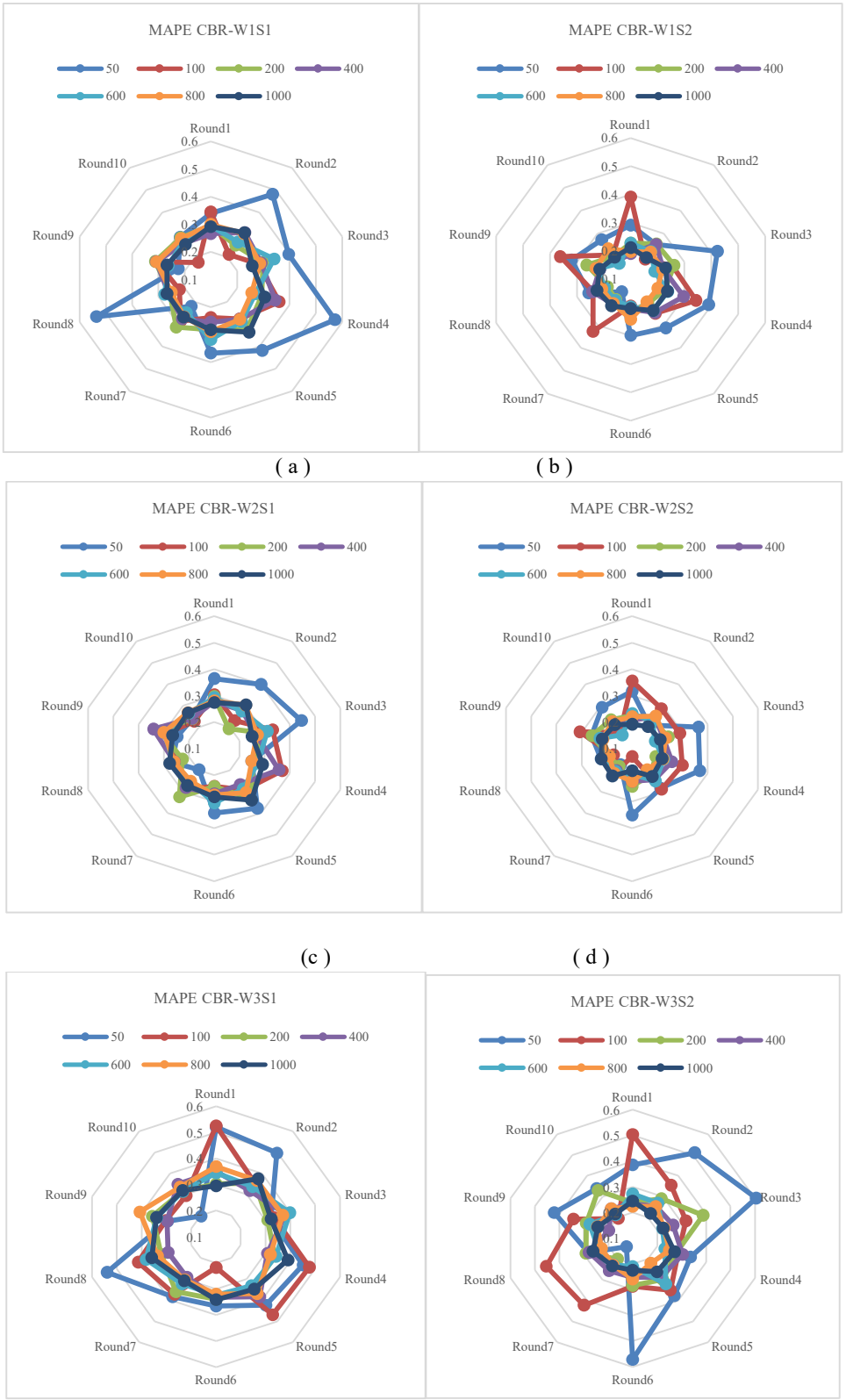
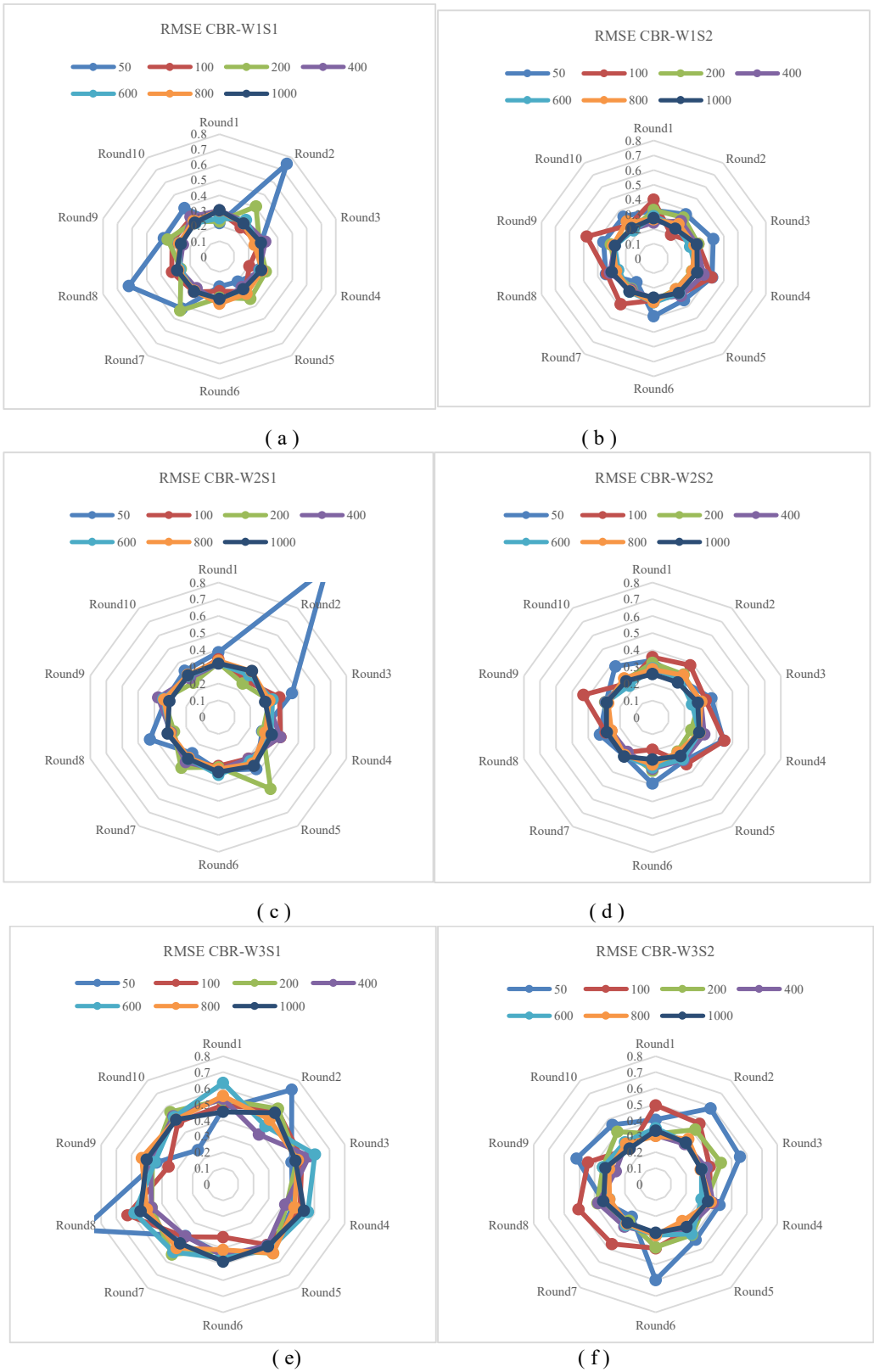


Figure 5.5 Comparative results of MAPE (K=1)



(a) (b)

(c) (d)

(e) (f)

Figure 5.6 Comparative results of RMSE (K=1)

Figure 5.1 to Figure 5.6 illustrate the comparative results of parameter setting based on different sample sizes. In each figure, there are six subfigures representing the MAPE/ RMSE of a specific model setting. The different colour lines represent the different sample sizes ranging from 50 to 1000. For example, in Figure 5.1 (a), the blue lines represent the MAPE of 50 training samples and the dark red lines represent the results of 100 training samples when case adaptation value is five.

Despite the changes in the weighting method and similarity function, the results of the 50 sample sizes have the largest fluctuations during the multiple calculations. With the increase in the sample size, the variances are getting less, resulting in more stable performance. Generally, CBR-W2S2 has the least mean MAPE, followed by CBR-W1S2. CBR-W3S1 has the weakest performance among all the CBR models. Despite the various parameter settings, this ranking remains the same.

5.2.2 Average Performance

Table 5.2 to 5.7 illustrates the changes in mean MAPE and RMSE with sample size. Figure 5.7 to Figure 5.12 shows the changing trend with the increase in the size of the case-base. Figure 5.7, for example, shows how mean MAPE decreases with the increase in the size of case-base (K=5). The remaining Figures 5.8 to 5.12 are interpreted in the same way.

From Table 5.2 to 5.7, it can be seen that CBR's predictive performance improves significantly, and both the MAPE and RMSE of CBR decrease dramatically, with an increased sample size. For example, in Table 5.2, when case adaptation is set as 1, the mean MAPE of CBR-W1S1 decreases from 38.85 % to 24.30%; the mean MAPE of CBR-W2S1 decreases from 34.06% to 23.73%; the mean MAPE of CBR-W3S1 decreases from 44.90% to 28.88%; the mean MAPE of CBR-W1S2 decreases from 35.01% to 19.02 %; the mean MAPE of CBR-W2S2 decreases from 24.90 % to 19.09 %; the mean MAPE of CBR-W3S2 decreases from 42.57 % to 21.72%. Similarly, the mean RMSE of CBR-W1S1 decreases from 42.30 % to 27.03%; the mean RMSE of CBR-W2S1 decreases from 35.06% to 26.53%; the mean RMSE of CBR-W3S1 decreases from 48.64% to 40.83%; the mean RMSE of CBR-W1S2 decreases from 38.33% to 23.83%; the mean RMSE of CBR-W2S2 decreases from 29.52 % to 23.04%; the mean RMSE of CBR-W3S2 decreases from 44.06 % to 26.78%. The remaining Tables 5.3 to 5.7 are interpreted in the same way.

Table 5.2 Changes in mean MAPE with sample size (K=5)

Sample size	50	100	200	400	600	800	1000
CBR-W1S1	38.85%	28.06%	24.34%	22.80%	23.17%	24.04%	24.30%
CBR-W2S1	34.06%	26.92%	23.11%	22.28%	22.80%	23.34%	23.73%
CBR-W3S1	44.90%	35.85%	31.39%	29.70%	30.20%	29.81%	28.88%
CBR-W1S2	35.01%	24.17%	21.23%	19.31%	18.39%	18.63%	19.02%
CBR-W2S2	24.90%	20.65%	20.38%	19.73%	18.76%	18.96%	19.09%
CBR-W3S2	42.57%	38.06%	28.08%	24.06%	23.24%	22.10%	21.72%

Table 5.3 Changes in mean RMSE with sample size (K=5)

Sample size	50	100	200	400	600	800	1000
CBR-W1S1	42.30%	30.86%	28.97%	25.74%	26.25%	26.86%	27.03%
CBR-W2S1	35.06%	30.47%	26.14%	25.31%	25.80%	26.33%	26.53%
CBR-W3S1	48.64%	42.40%	42.50%	41.09%	44.42%	41.60%	40.83%
CBR-W1S2	38.33%	29.05%	25.57%	23.48%	23.28%	23.23%	23.83%
CBR-W2S2	29.52%	25.78%	24.06%	23.01%	22.98%	22.96%	23.04%
CBR-W3S2	44.06%	41.29%	33.02%	28.80%	28.97%	27.48%	26.78%

Table 5.4 Changes in mean MAPE with sample size (K=3)

Sample size	50	100	200	400	600	800	1000
CBR-W1S1	35.81%	26.06%	25.07%	24.02%	25.02%	24.85%	25.31%
CBR-W2S1	28.38%	24.13%	24.26%	23.39%	24.37%	24.24%	24.92%
CBR-W3S1	39.29%	35.43%	31.43%	30.18%	31.07%	30.73%	30.05%
CBR-W1S2	32.53%	23.71%	21.41%	19.64%	18.25%	18.66%	19.51%
CBR-W2S2	24.90%	20.65%	20.38%	19.73%	18.76%	18.96%	19.09%
CBR-W3S2	39.20%	34.87%	28.41%	22.69%	22.64%	21.85%	21.75%

Table 5.5 Changes in mean RMSE with sample size (K=3)

Sample size	50	100	200	400	600	800	1000
CBR-W1S1	40.42%	29.56%	29.96%	27.18%	28.05%	27.79%	28.12%
CBR-W2S1	30.55%	28.01%	27.51%	26.51%	27.45%	27.30%	27.80%
CBR-W3S1	45.11%	43.28%	43.97%	41.99%	46.27%	43.32%	42.95%
CBR-W1S2	36.78%	28.68%	26.23%	23.97%	23.51%	23.61%	24.56%
CBR-W2S2	27.96%	26.02%	24.91%	23.80%	23.68%	23.56%	23.72%
CBR-W3S2	43.17%	38.63%	33.62%	27.81%	29.01%	27.48%	27.20%

Table 5.6 Changes in mean MAPE with sample size (K=1)

Sample size	50	100	200	400	600	800	1000
CBR-W1S1	38.45%	27.01%	29.12%	28.54%	28.34%	28.22%	28.42%
CBR-W2S1	31.32%	28.25%	27.20%	28.40%	28.24%	27.61%	28.24%
CBR-W3S1	39.96%	37.63%	33.40%	33.04%	34.38%	34.95%	33.99%
CBR-W1S2	29.72%	27.56%	22.30%	22.58%	20.91%	21.51%	21.68%
CBR-W2S2	28.47%	25.17%	22.26%	22.56%	21.37%	21.96%	21.20%
CBR-W3S2	45.63%	41.21%	36.95%	32.61%	32.69%	31.40%	31.83%

Table 5.7 Changes in mean RMSE with sample size (K=1)

Sample size	50	100	200	400	600	800	1000
CBR-W1S1	50.92%	31.35%	33.61%	32.79%	32.42%	32.24%	32.27%
CBR-W2S1	42.74%	32.48%	32.33%	32.42%	32.18%	31.86%	32.06%
CBR-W3S1	50.96%	47.61%	49.96%	46.87%	52.79%	49.79%	49.70%
CBR-W1S2	35.27%	33.96%	28.85%	29.04%	27.53%	28.17%	28.05%
CBR-W2S2	34.12%	32.80%	28.32%	29.02%	27.92%	28.21%	27.40%
CBR-W3S2	45.63%	41.21%	36.95%	32.61%	32.69%	31.40%	31.83%

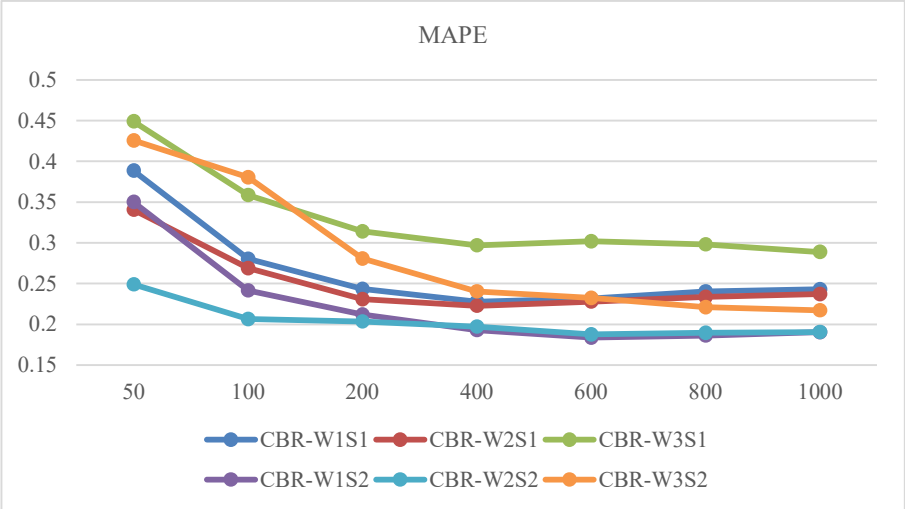


Figure 5.7 Changes in mean MAPE with sample size (K=5)

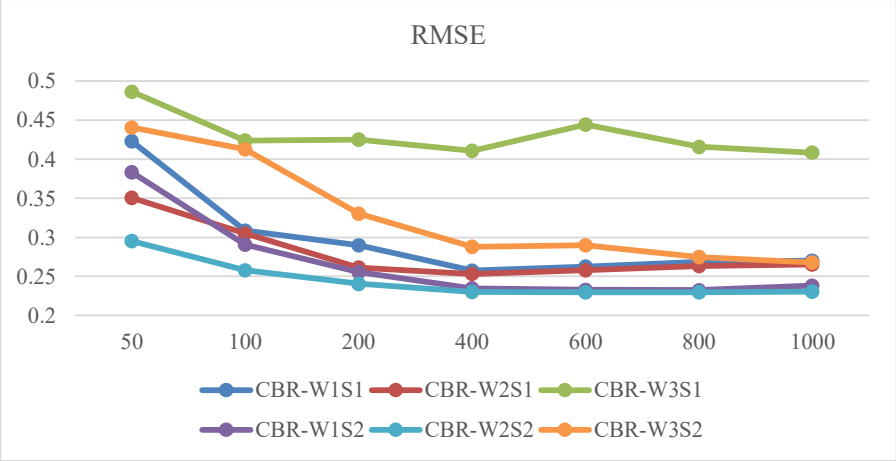


Figure 5.8 Changes in mean RMSE with sample size (K=5)

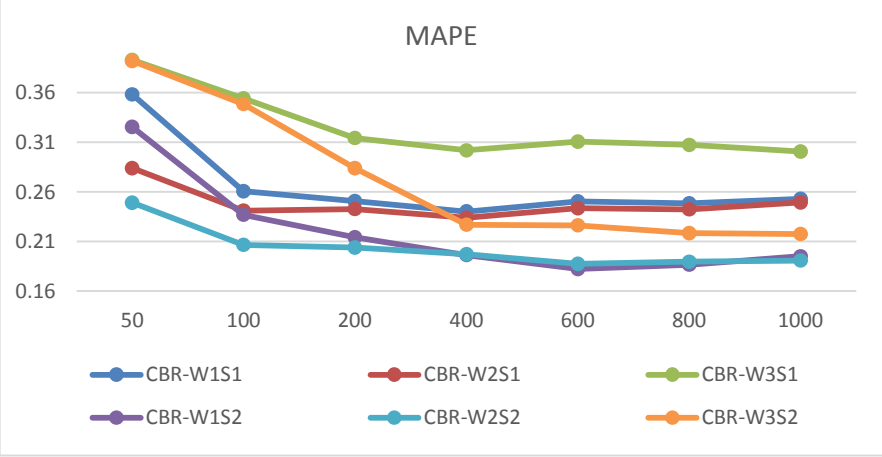


Figure 5.9 Changes in mean MAPE with sample size (K=3)

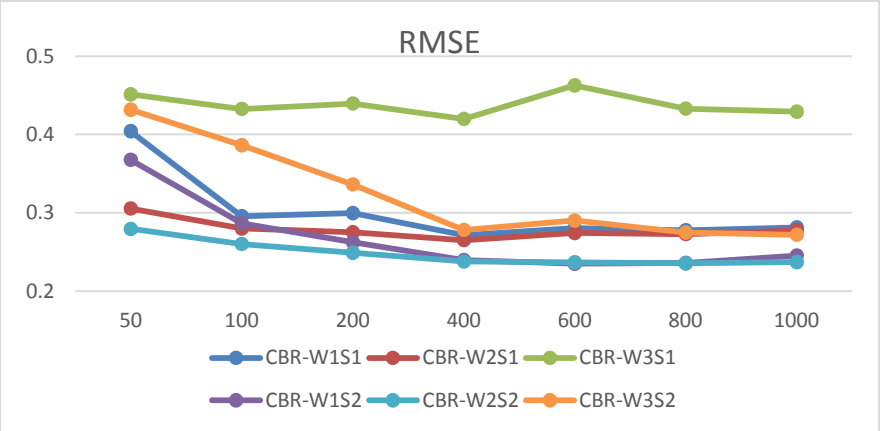


Figure 5.10 Changes in mean RMSE with sample size (K=3)

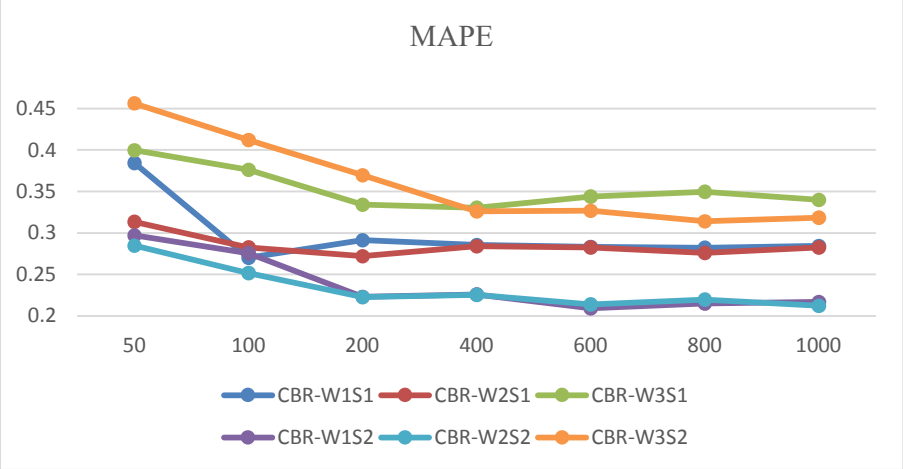


Figure 5.11 Changes in mean MAPE with sample size (K=1)

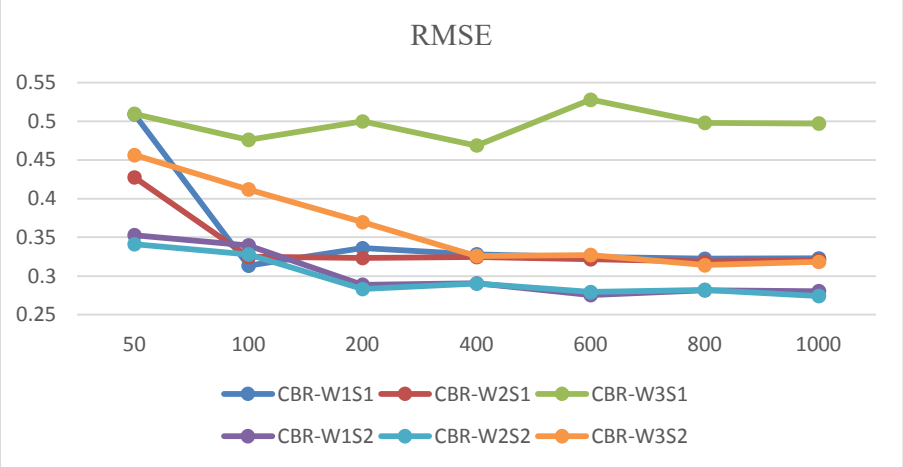


Figure 5.12 Changes in mean RMSE with sample size (K=1)

5.3 DISCUSSION

5.3.1 Comparison of Settings in CBR

The influence of weight determination

Despite different sample sizes, CBR-W2S2 produces the least error rates. The results of each weight determination based on different sample sizes are represented in Figure 5.1 to 5.6. For example, from Figure 5.1, when the sample sizes are small (50/100/200) , the CBR-W2S2 has the smallest MAPE and RMSE. When sample size reaches 400, there is no significant difference between CBR-W1S2's and CBR-W2 S2's MAPE and RMSE.

When the sample size is small (<400), the weight determination method GA has more advantages than MRA in terms of accuracy. This illustrates that GA is more suitable to be used in weight determination when dealing with a smaller sample size. After the size of the case-base become large (>400), there is no significant difference between models using MRA and GA. FC has the weakest performance in terms of accuracy, especially when combining Similarity1 function. CBR-W3S1 and CBR-W3S2 have the largest MAPE and RMSE. This illustrates the inferiority of feature counting as the weight determination method in ECCE CBR. It also illustrates the significance to weight attributes differently. Compared with equally weighting, considering differences in the contribution of attribute provides more reasonable results.

Similarity function

Totally, there are two similarity functions used in this chapter. In summary, Similarly 2 is better than Similarity 1 despite different weight determination methods. The results of each similarity function based on different sample sizes are represented in Table 5.2 to Table 5.7. Overall, CBR-W1S2 and CBR-W2S2 have the smaller MAPE and RMSE than CBR-W1S1 and CBR-W1S1. When Similarity 2 and FC are used together, the CBR-W3S1 model has the weakest performance in terms of accuracy. The results show that the Similarity 2 function has better performance than Similarity 1 on average.

Case adaptation value

K-NN is widely used as the case adaptation method. The K-NN ranks the case's neighbours in the case-base and uses the labels of the K most similar neighbour to predict the label of the new case. A different number of neighbours can be used in the

case adaptation. In this study, the value of K is set to five, three, and one. The results show that the CBR model has the best performance when K is set to 5 averagely. Despite the changes in the sample size and different weight determinations, the results show that adjusting case adaption value could improve the performance of the CBR model. For example, the mean MAPE of CBR-W1S1 for 100-, 200-, 400-, 600-, 800-, 1000-sample size can be decreased from 27.01%, 29.12%, 28.54%, 28.34%, 28.22% and 28.42% to 28.06%, 24.34%, 22.80%, 23.17%, 24.04%, 24.30% when the value of case adaptation is adjusted from 1 to 5. However, the results show that the optimal case adaptation value differs in each situation. For different sample sizes, the optimal case adaptation value may differ. When the sample size is 50, the optimal case adaptation value is 3, while in other situations, the optimal case adaptation value is 5. The finding shows us that multiple case adaptation values should be tried in the CBR study to have the best performance.

5.3.2 Justification of Sample Size in ECCE

Existing studies do not consider the influence of sample size when developing the ECCE CBR models, nor the implications of increasing it to as large as 1000. Therefore, using the different sample sizes, this research demonstrates the need to consider the influence of sample size used in ECCE studies. The results show that the performance of CBR improves with the increase of the sample size.

What is also becoming apparent is that expanding the database to contain more cases containing a small number of highly influential predictor variables (six in this case) may be a good use of resources. Therefore, even though ECCE lacks detailed target project information at the early stage, increasing the sample size may be sufficient to produce reliable results by objective methods.

Another finding is that the marginal contribution of the increasing size of case-base decreases. For example, when K is set to 1, the mean MAPE of CBR-W1S1 decreases from 38.45% to 28.54% by increasing the size of case-base from 50 to 400; while the performance of the model CBR-W1S1 remains the same after the size of case-base increases. This trend keeps the same in the remaining CBR models despite the differences in the case adaptation values. For example, when case adaptation value is set to 5, the mean MAPE of CBR-W1S2 decreases from 32.53% to 19.64 % by increasing the size of case-base from 50 to 400; while the mean MAPE of CBR-W1S1 remains the same after the size of case-base increasing (> 400).

5.3.3 Explanation of Data-oriented Results

The results show that existing studies are heavily data-based, given the difficulty of data collection for ECCE. When the sample size is small, MAPE and RMSE vary significantly. For example, as shown in Table 5.1, the MAPE of the CBR-W1S1 is from 18.39% to 61.57% in 10 trials, although their sample sizes are both 50. When the sample size is 50, the best case adaptation value is 3, while in other situations, the best case adaptation value is 5. Since there are various settings in the CBR model, the performance of the CBR method significantly relies on the quality of the case-base. This may provide a justified reason for the contradictory results in the previous comparative studies on CBR, ANN, and MRA.

This study also illustrates the necessity to consider the influence of sample size when establishing an ECCE CBR model. The optimal parameter settings may differ due to the different sample size. GA is more suitable to be used in weight determination when dealing with a smaller sample size. When the sample size is small (<400), the weight determination method GA has more advantages than MRA in terms of accuracy. After the size of the case-base becomes large (>400), there is no significant difference between models using MRA and GA. Besides, the optimal case adaptation value may differ based on different sample size. When the sample size is 50, the optimal case adaptation value is 3, while in other situations, the optimal case adaptation value is 5.

Several limitations are also found in the existing CBR model. The current research focuses on optimizing the accuracy of the CBR system for a single time running, while ignoring the long-term use of the CBR model. Except for accuracy, the robustness of the case-base is also necessary to be considered to reduce the CBR's dependence on the data. Besides, after the size of the case-base reaches a certain amount, an increase in the size of the case-base has a limited impact on improving the accuracy of the CBR system.

Therefore, improving the long-term use of the case-base, should not only consider the benefits of the increase in the size of case-base, but also address the stability and efficiency problem brought by continuous growth in the number of cases stored in the case-base. When the size of case-base gets large, it is inevitable to have a few cases which deviate from those mainstream bulk of cases. This begs the problem of robustness of the CBR system. Besides, with the increase in the number of cases stored in the case-base, the case-base can grow very fast in the sense that it can slow the speed of the query execution time concerning case-research phase. This inevitably raises the

question of how to select cases for avoiding excessive storage and time complexity, and possibly to maintain the system's performance. Greater attention should be paid on these issues of the long-term use of the ECCE CBR system.

5.4 CHAPTER SUMMARY

This chapter provides a comparative study of the different model settings of CBR based on a collected ECCE database of 1450 completed Chinese apartment building projects. Given the sample size range in previous studies and the size of the collected database, seven sample sizes of 50, 100, 200, 400, 600, 800, and 1000 contracts, cover the range of sample sizes used in previous studies. Three weight determination methods (MRA, GA, FC) and two similarity functions are compared by random selections from the database, with a 20% project hold-out sample. MAPE and RMSE are used as measures of ECCE forecasting performance.

The results in this chapter help to provide a better understanding of the settings in the CBR models. CBR-W2S2 and CBR-W1S2 are the best CBR models for all sample sizes. However, GA has more advantages when dealing with smaller sample sizes. After the size of the case-base becomes large (>400), there is no significant difference between models using MRA and GA. FC has the weakest performance in terms of accuracy, uniquely when combining Similarity1 function. Besides, the Similarity 2 function has better performance than Similarity1 on average. Also, adjusting case adaptation values in case adaption may improve the performance of the CBR model. However, for different sample sizes, the best case adaptation value may be different. Therefore, multiple case adaptation values should be tried in the CBR study to find the best performance.

Several limitations are also found in the existing CBR model. The results show that ECCE CBR models are heavily data-based. Different sample size and training sample will cause huge differences in the results. Besides, the current research focuses on optimizing the accuracy of the CBR system for a single time running while ignoring the maintenance of CBR model during its operation. Except for accuracy, the robustness and efficiency of the case-base are also necessary to be considered during the application of ECCE CBR system.

As a core component of CBR, the case-base is the knowledge and experience container. The knowledge structure in the case-base plays a key role in the ECCE CBR

model. Therefore, improving the robustness of the knowledge structure becomes essential.

Since the marginal contribution of the increasing volume of the case-base decreases, additional maintenance of it is necessary to deal with the problems that arise during long-term use, especially when the knowledge in case-base changes over time. The question of how to select cases for avoiding excessive storage and time complexity should be explored.

Chapter 6: Improving the ECCE CBR Model by using MODAL

6.1 INTRODUCTION

The previous chapter explored the influence of sample size on the accuracy of the three most studied ECCE methods. The results confirmed that the CBR model has an advantage in long-term use for ECCE. To further improve the ECCE CBR model, this chapter introduces a robust weight determination method, since attribute weights significantly influence on the efficiency and accuracy of case retrieval. The estimation accuracy improves when attribute weights become more robust (Changchien & Lin, 2005; Lee, et al., 2013a).

When the size of the case-base gets large, it is inevitable to have a few cases deviating from the mainstream bulk of cases (Chan & Wong, 2007). Existing weight determination methods have limitations in being sensitive to outliers. A single outlier can have a large influence on the parameter estimates, reducing the accuracy of attribute weight (Yao & Li, 2014). To minimize the influence of noisy data in the case-base, the ECCE CBR should increase the robustness for long-term use. Here robustness means the consistency in results (Schall et al., 2005). Given the significance of the weight determination in case retrieval, improving the robustness of the weight determination is the primary task when dealing with large volumes of data (Aamodt & Plaza, 1994).

The stable structure of the case-base is critical because it affects the consistency of knowledge updating in the CBR model. This is particularly important for ECCE because of the consistency of knowledge on related project features help practitioners have a consistent understanding of ECCE.

Therefore, this chapter uses the MODLR to increase the robustness of the weight determination. Based on the principle of MODAL introduced in section 3.4.2, this chapter presents the results of the MODAL-CBR model. A case study was used to illustrate the process of weight determination based on MODLR regression in the CBR model for ECCE. K-folder cross-validation is used for model validation. K-fold cross-validation uses part of the available data to fit the model, and a different part for testing

it. The data is split into K separate sets of equal size. Note that when the value of K is 1450, K-folder cross-validation equals leave one cross-validation. The larger the value of K in K-fold, the greater the proportion of duplicate cases in the training set in each calculation. The reason for using multiple K-folder cross-validation is to show the changes in the weighting with respect to the changes in the training set.

Given the influence of the number of cases in the test dataset on model development, the value of K-folder is set to be 10, 20, 40, 80, 160, 320 and 1450. The measure of the estimation accuracy includes the MAPE and RMSE. The results are compared with those using ordinary least squares (OLS) regression and GA in weight determination, followed by the discussion on the overall robustness of the weight determination.

6.2 RESULTS

6.2.1 Weight Calculation

Table 6.1 to Table 6.3 represent attribute weights calculated by OLS, GA and MODLR when the values of KF in K-fold are 10, 20, 40, 80, 160, 320 and 1450. For example, when K -fold is 10, the mean value of weights calculated by OLS are 0.644, 0.206, 0.035, 0.056, 0.048 and 0.011; the mean value of weights calculated by GA are 0.625, 0.277, 0.044, 0.022, 0.015, 0.017; the mean value of weights calculated by MODAL are 0.494, 0.186, 0.102, 0.015, 0.023, and 0.180. Similarly, when K -fold is 10, the variances of weights calculated by OLS are 8.49E-05, 1.22E-04, 1.06E-04, 1.60E-05, 4.54E-06, 2.36E-05; the variances of weights calculated by GA are 4.15E-04, 2.96E-04, 5.23E-04, 1.96E-04 ,4.69E-05, 2.67E-04; the variances of weights calculated by MODAL are 8.32E-05, 2.68E-05, 4.70E-05, 1.49E-05, 4.16E-06, 4.87E-05.

In OLS, the most significant attribute is the total above-floor area, followed by total below-floor area, no. of storeys below the ground, duration, no. of storeys above the ground and total height of building. Similarly, in GA, the most significant attribute is total above-floor area, followed by total below-floor area, no. of storeys above the ground, no. of storeys below the ground, total height of the building and duration. In MODAL, the most significant attribute is total above-floor area, followed by total below-floor area, total height of building, no. of storeys above the ground, duration and no. of storeys below the ground.

Table 6.1 Attribute weights calculated by OLS

Mean						
K-fold	Weight1	Weight2	Weight3	Weight4	Weight5	Weight6
10	0.6437	0.2055	0.0353	0.0560	0.0482	0.0113
20	0.6444	0.2055	0.0352	0.0562	0.0482	0.0105
40	0.6454	0.2054	0.0350	0.0564	0.0483	0.0095
80	0.6459	0.2055	0.0350	0.0564	0.0483	0.0088
160	0.6462	0.2056	0.0350	0.0565	0.0484	0.0083
320	0.6463	0.2056	0.0350	0.0565	0.0484	0.0083
1450	0.6466	0.2056	0.0350	0.0565	0.0484	0.0079
Average	0.6455	0.2055	0.0351	0.0564	0.0483	0.0092
Variance						
K-fold	Weight1	Weight2	Weight3	Weight4	Weight5	Weight6
10	8.49E-05	1.22E-04	1.06E-04	1.60E-05	4.54E-06	2.36E-05
20	1.16E-04	5.65E-05	1.10E-04	1.80E-05	1.96E-06	6.48E-05
40	2.63E-05	3.30E-05	5.36E-05	1.15E-05	1.06E-06	2.48E-05
80	1.91E-05	2.03E-05	3.11E-05	3.93E-06	5.74E-07	1.44E-05
160	8.26E-06	9.66E-06	1.45E-05	1.84E-06	3.07E-07	7.10E-06
320	1.15E-05	7.62E-06	1.46E-05	1.78E-06	2.71E-07	8.15E-06
1450	1.07E-06	1.03E-06	1.69E-06	2.10E-07	3.23E-08	8.81E-07

Table 6.2 Attribute weights calculated by GA

Mean						
K-fold	Weight1	Weight2	Weight3	Weight4	Weight5	Weight6
10	0.6252	0.2768	0.0440	0.0216	0.0154	0.0169
20	0.6343	0.2627	0.0387	0.0277	0.0164	0.0202
40	0.6379	0.2632	0.0304	0.0283	0.0158	0.0243
80	0.6372	0.2664	0.0346	0.0243	0.0159	0.0215
160	0.6384	0.2636	0.0341	0.0269	0.0156	0.0213
320	0.6387	0.2635	0.0307	0.0277	0.0156	0.0237
1450	0.6377	0.2639	0.0333	0.0273	0.0155	0.0223
Average	0.6356	0.2657	0.0351	0.0263	0.0158	0.0215
Variance						
K-fold	Weight1	Weight2	Weight3	Weight4	Weight5	Weight6
10	4.1483E-04	2.9627E-04	5.2310E-04	1.9612E-04	4.6859E-05	2.6664E-04
20	7.6623E-04	5.8544E-04	3.9157E-04	2.1340E-04	3.9835E-05	4.3158E-04
40	5.2765E-04	4.9299E-04	4.9029E-04	2.0390E-04	6.2274E-05	2.8568E-04
80	6.4862E-04	4.1861E-04	4.1208E-04	1.7982E-04	6.3973E-05	3.9141E-04
160	6.6739E-04	4.5397E-04	4.0471E-04	1.8253E-04	6.1593E-05	3.1935E-04

320	6.3412E-04	4.1778E-04	4.1534E-04	1.9402E-04	5.6318E-05	4.3230E-04
1450	6.9116E-04	4.8254E-04	4.5643E-04	1.8747E-04	5.5706E-05	3.6701E-04

Table 6.3 Attribute weights calculated by MODLR

Mean

K-fold	Weight1	Weight2	Weight3	Weight4	Weight5	Weight6
10	0.4942	0.1862	0.1020	0.0145	0.0232	0.1799
20	0.5068	0.1846	0.0989	0.0075	0.0255	0.1769
40	0.5011	0.1846	0.1005	0.0108	0.0246	0.1783
80	0.4984	0.1848	0.1014	0.0123	0.0241	0.1790
160	0.4927	0.1847	0.1040	0.0148	0.0229	0.1809
320	0.4947	0.1847	0.1030	0.0140	0.0234	0.1802
1450	0.4960	0.1846	0.1025	0.0134	0.0236	0.1799
Average	0.4977	0.1849	0.1018	0.0125	0.0239	0.1793

Variance

K-fold	Weight1	Weight2	Weight3	Weight4	Weight5	Weight6
10	8.3211E-05	2.6849E-05	4.6967E-05	1.4948E-05	4.1610E-06	4.8667E-05
20	5.0139E-05	2.7375E-05	3.2899E-05	9.4225E-06	7.7599E-07	1.5000E-05
40	1.3189E-05	1.2394E-05	2.0046E-05	3.6729E-06	7.7038E-07	1.1531E-05
80	1.4210E-05	7.3865E-06	8.8455E-06	2.8601E-06	4.4637E-07	6.6906E-06
160	4.9882E-06	2.7578E-06	3.9636E-06	1.1300E-06	2.1997E-07	3.1429E-06
320	5.7037E-06	3.2006E-06	4.0616E-06	9.3401E-07	2.1408E-07	3.6225E-06
1450	6.6460E-07	3.7565E-07	5.2755E-07	1.0484E-07	2.4021E-08	3.6023E-07

It can be seen that the total above-floor area is the most significant attribute in all weight determination methods. The average weights of the total above-floor area are 0.6455 by OLS, 0.6356 by GA, 0.4977 by MODAL. The second most important attribute weight is the total below-floor area. The average weights of total below-floor area are 0.2055 by using OLS, 0.2675 by using GA, 0.1849 by using MODAL. The third most significant attribute differs due to the weight determination methods. The third is the no. of storeys below the ground (0.0564) by using OLS, no. of storeys above the ground (0.0351) by using GA, the total height of the building (0.1793) by MODAL. Except for Modal, duration and the total height of the building are the least significant attribute weight for ECCE.

It can be seen that OLS and GA are not much different in the mean value of attribute weight. The total above-floor area is less important in weighting, and the total height of the building is more important when using MODAL.

Variance of weight attribute is used for measuring the weight robustness. K-fold cross-validation uses part of the available data to fit the model, and a different part for testing. The smaller the value of KF in K-fold, the more repeatability among training samples during the calculation of each fold. The changes in the attribute weight of each fold represents the changes in case-base's structure caused by the difference in the training sample. For the long-term use of the system, it is necessary to reduce the fluctuations of the case-base's structure caused by the changes in the training sample while ensuring the model's performance. The weight determination in the ECCE CBR model should be capable of capturing the feature of the mainstream bulk of cases and reduce the unnecessary changes caused by minority or outliers of the data. The variance of weight attributes is used to measure the robustness of attribute weight. The smaller variance of weight attributes represents less sensitivity to the minority of cases and greater robustness of the CBR model.

Table 6.1 to Table 6.3 represents the changes in the attribute weight in each calculation. For example, when the value of KF is ten, and weight determination method is OLS, the average weight of total above floor area is 0.6437; the average weight of total below floor area is 0.2055; the average weight of the no. of storeys above the ground is 0.0353; the average weight of the no. of storeys below the ground is 0.0560; the average weight of duration is 0.0482; the average weight of the total height of the building is 0.0113. Accordingly, the variances of total above-floor area, no. of storeys above the ground, no. of storeys below the ground, duration, total height of the building are 8.49E-05,1.22E-04,1.06E-04, 1.60E-05, 4.54E-06 and 2.36E-05.

Similarly, when using GA and the value of KF is ten, the variances of total above-floor area, no. of storeys above the ground, no. of storeys below the ground, duration, total heights of the building are 4.15E-04, 2.96E-04, 5.23E-04, 1.96E-04, 4.69E-05, 2.67E-04. When using MODAL and the value of KF in K-fold is ten, the variances of weights of total above-floor area, no. of storeys above the ground, no. of storeys below the ground, duration, and total height of the building are 8.32E-05, 2.68E-05, 4.70E-05, 1.49E-05, 4.16E-06 and 4.87E-05.

Figure 6.1 to Figure 6.7 represent the changes of the attribute weight in each calculation based on different values of KF. Despite the differences in the case adaptation value of K-fold, MODAL has the least weight variance compared with other methods. MODAL shows excellent performance in weight stability.

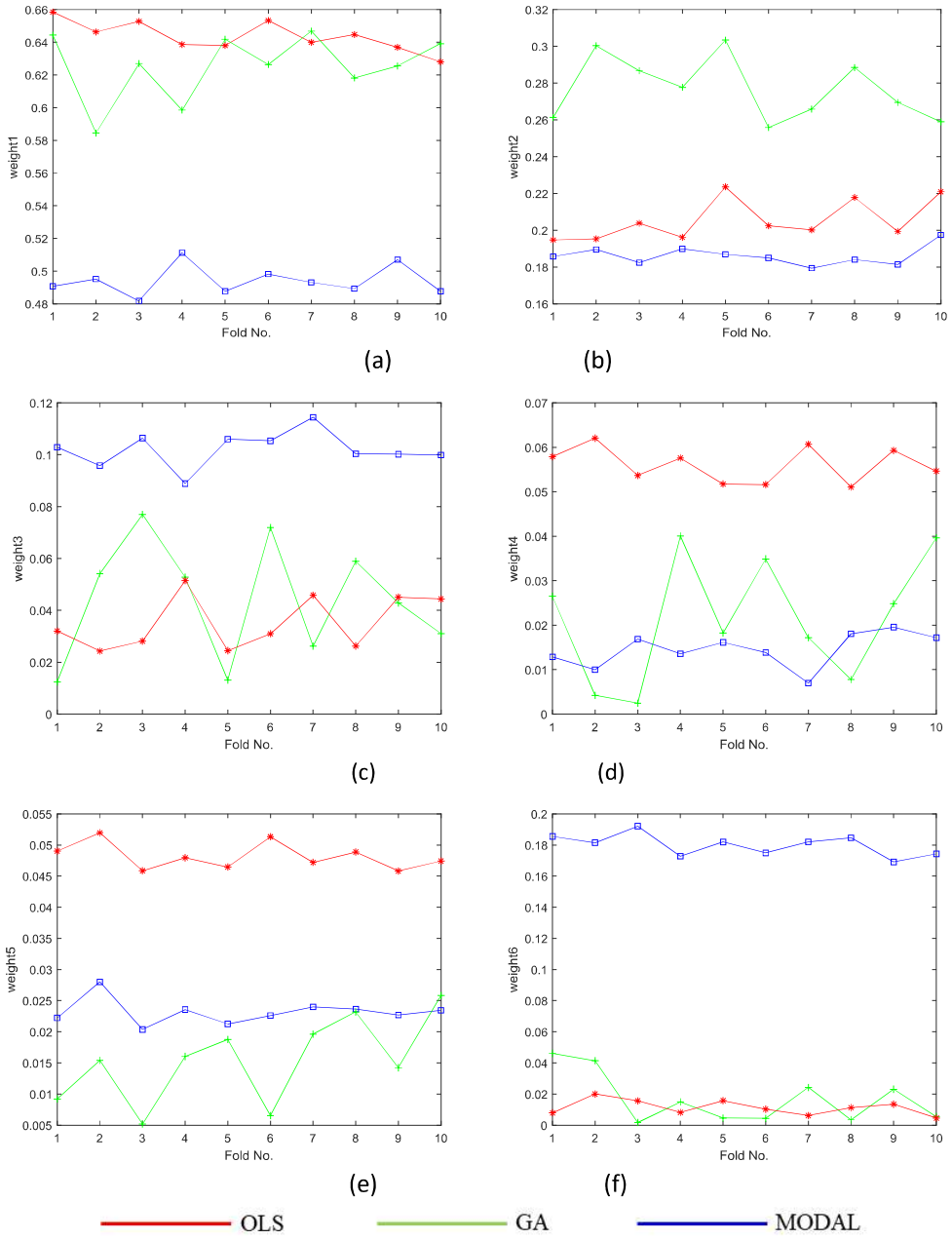


Figure 6.1 Attribute weights of each fold (k=10) : (a) total above floor area; (b) total below floor area; (c) No. of storeys above the ground; (d) No. of storeys below the ground; (e) duration; (f) total height of building

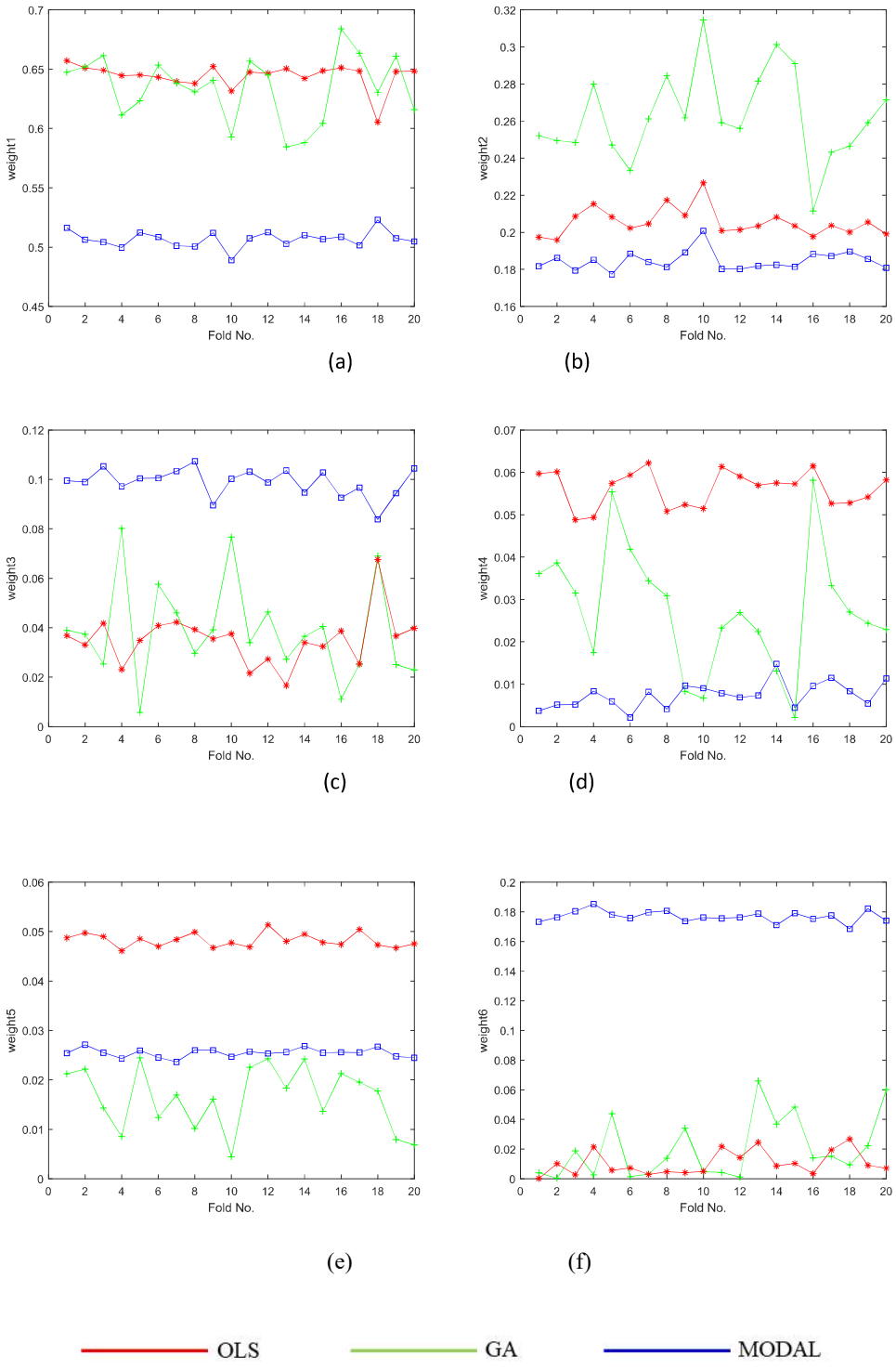


Figure 6.2 Attribute weights of each fold (k=20) : (a) total above floor area; (b) total below floor area; (c) No. of storeys above the ground; (d) No. of storeys below the ground; (e) duration; (f) total height of building

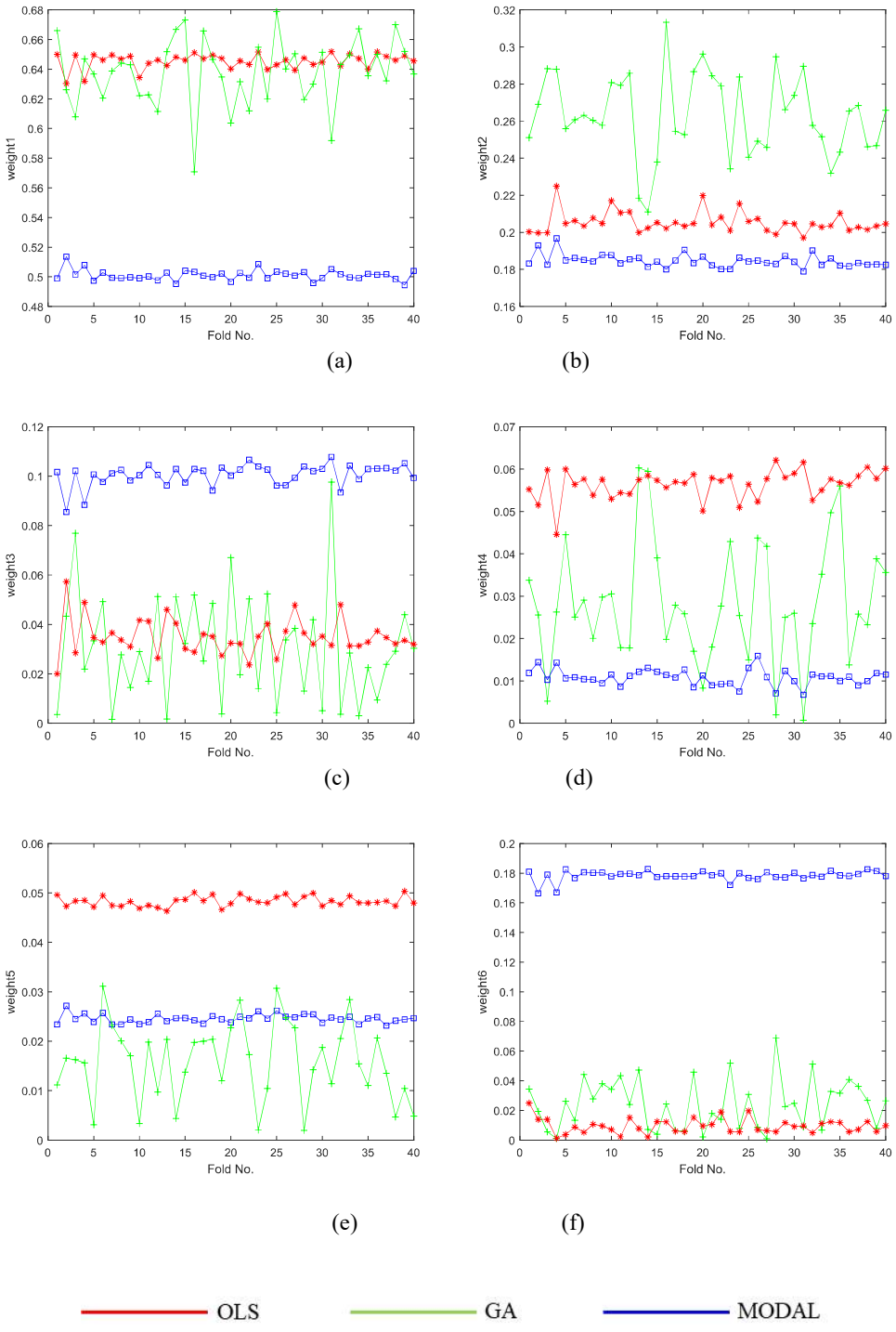


Figure 6.3 Attribute weights of each fold (k=40) : (a) total above floor area; (b) total below floor area; (c) No. of storeys above the ground; (d) No. of storeys below the ground; (e) duration; (f) total height of building

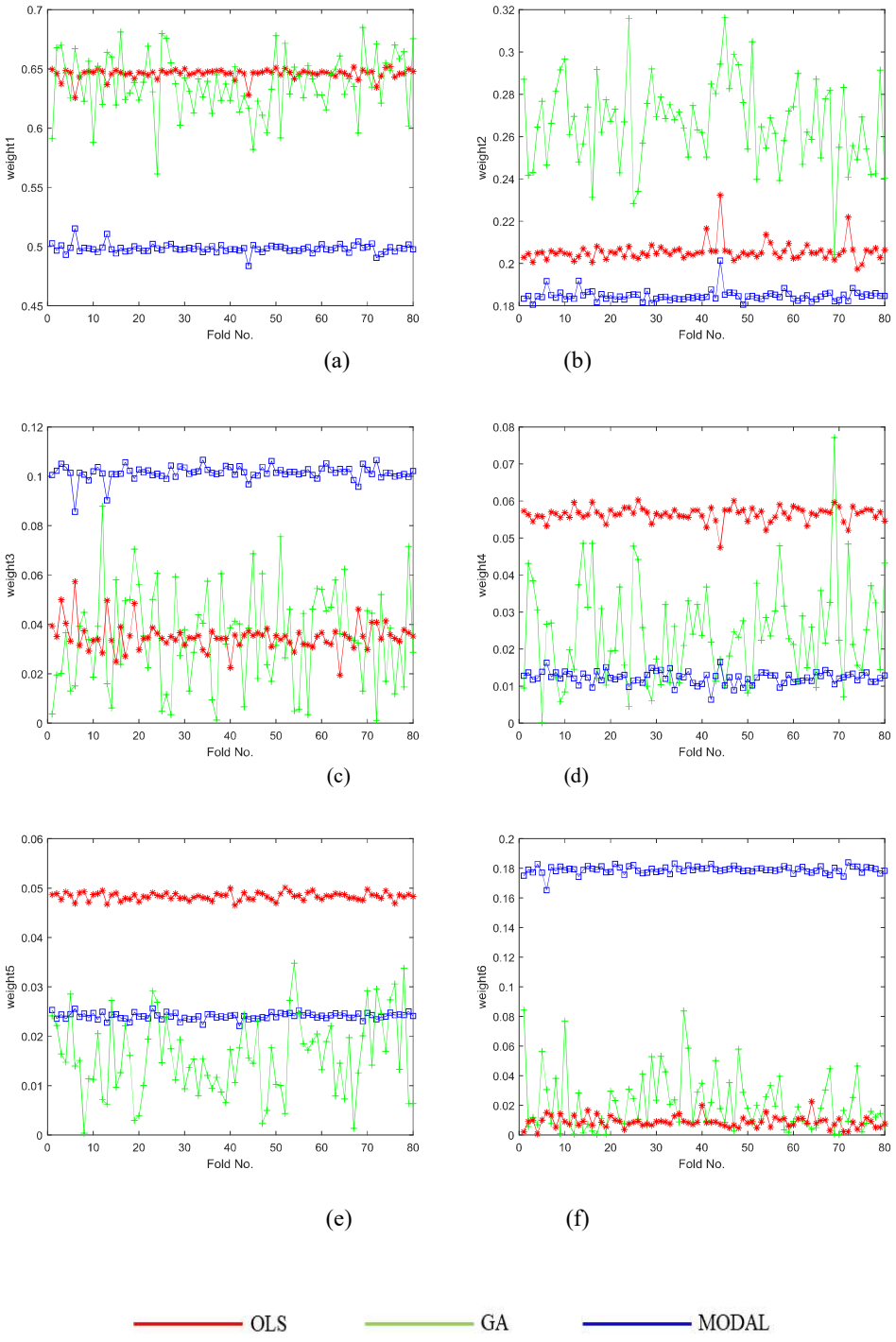


Figure 6.4 Attribute weights of each fold (k=80) : (a) total above floor area; (b) total below floor area; (c) No. of storeys above the ground; (d) No. of storeys below the ground; (e) duration; (f) total height of building

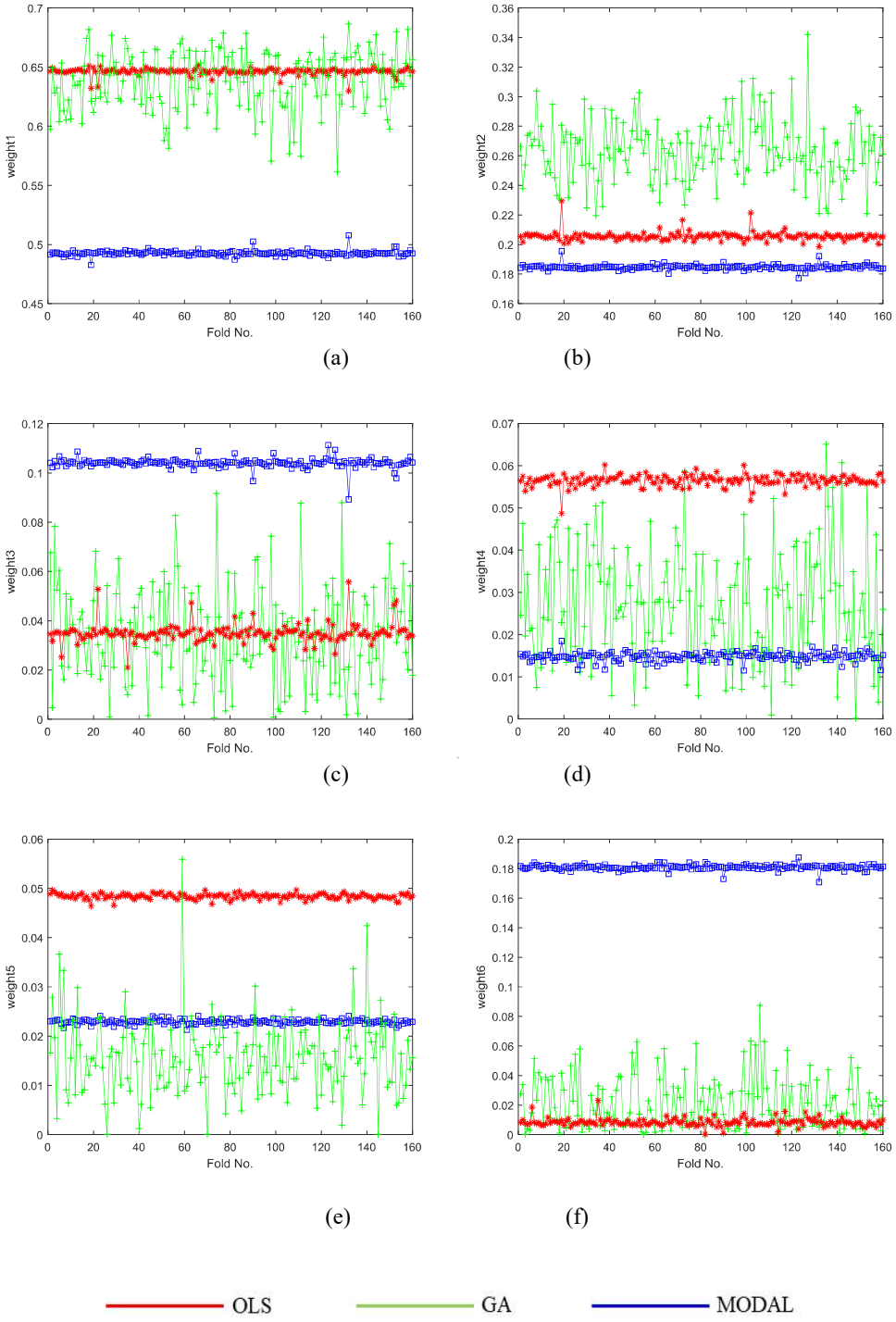


Figure 6.5 Attribute weights of each fold (k=160) : (a) total above floor area; (b) total below floor area; (c) No. of storeys above the ground; (d) No. of storeys below the ground; (e) duration; (f) total height of building

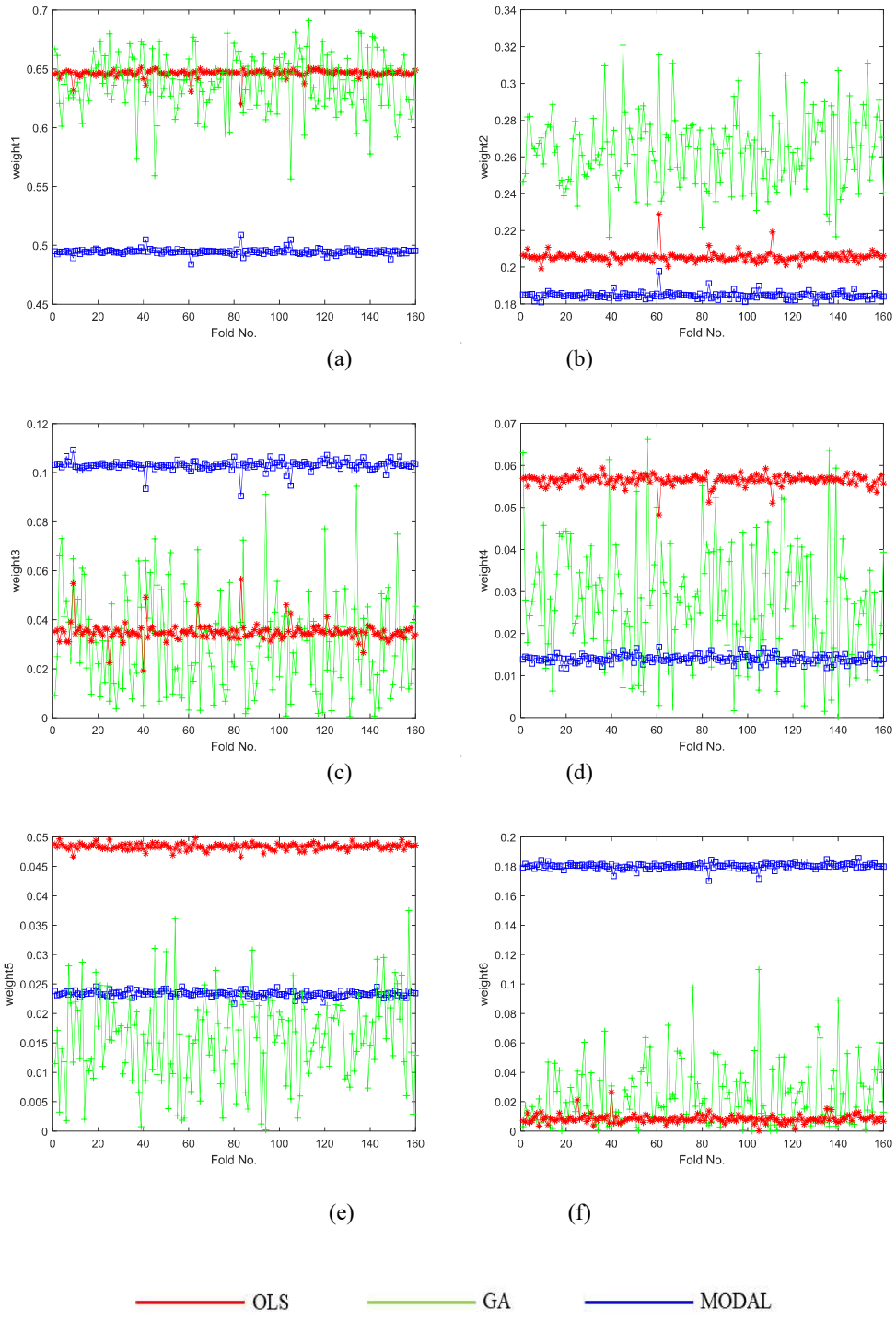


Figure 6.6 Attribute weights of each fold (k=320) : (a) total above floor area; (b) total below floor area; (c) No. of storeys above the ground; (d) No. of storeys below the ground; (e) duration; (f) total height of building

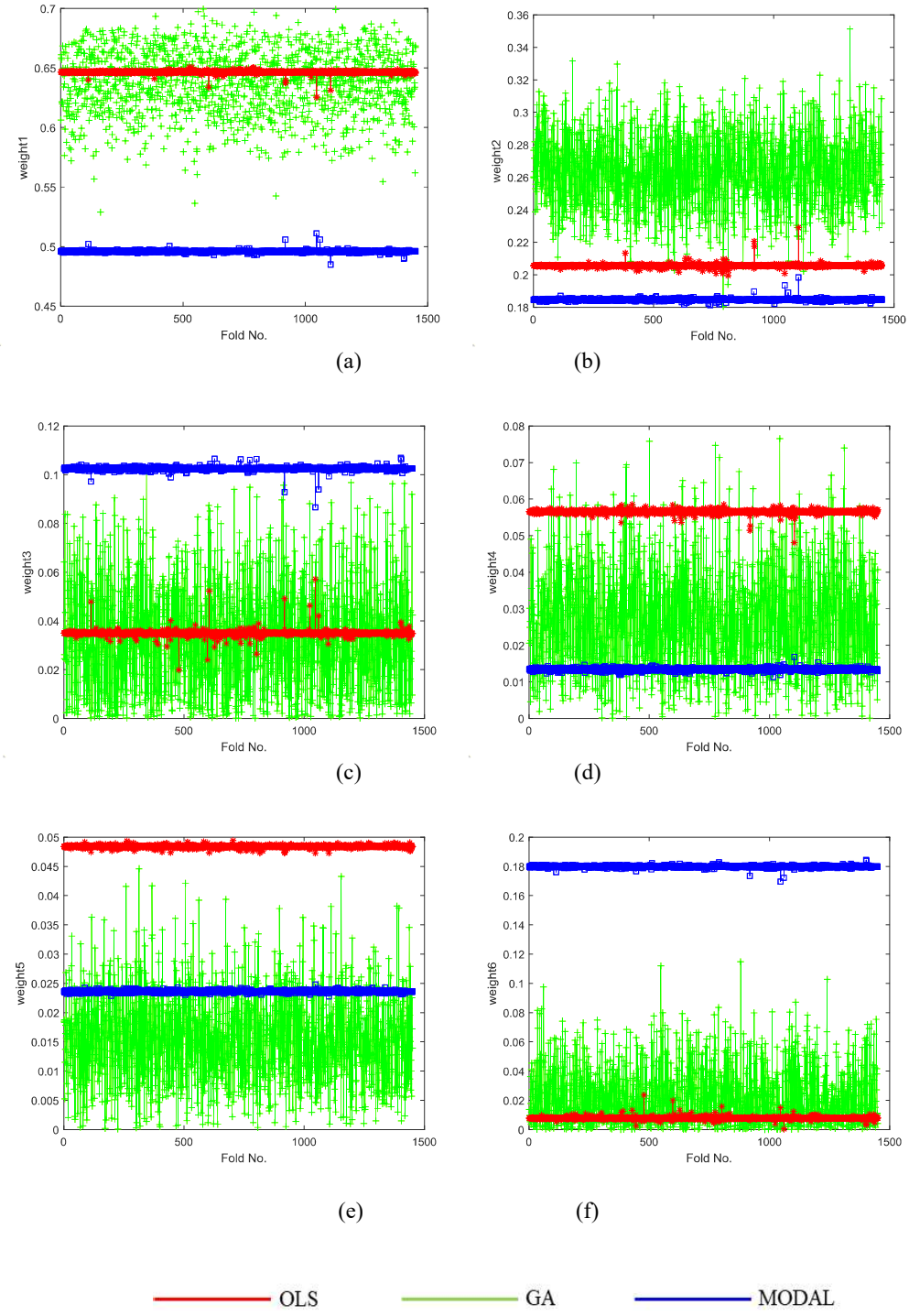


Figure 6.7 Attribute weights of each fold (k=1450) : (a) total above floor area; (b) total below floor area; (c) No. of storeys above the ground; (d) No. of storeys below the ground; (e) duration; (f) total height of building

6.2.2 Error Rate

Tables 6.4 to 6.6 show the results of the CBR model using OLS, GA and MODAL. When the parameters of case adaptation are 5/3/1, the average value of MAPE in the OLS-based CBR model is 0.1770/0.1812/0.2001, the average value of MAPE in GA-based CBR model is 0.1730/ 0.1765/0.1765, the average value of MAPE in MODAL-based CBR model is 0.1528/0.1463/0.1746. The accuracy of the CBR model differs due to the different setting parameters in case adaptation. The OLS and GA show the best results when the parameter of case adaptation is 5, while MODAL has the best performance when the parameter of case adaptation is 3. The parameter of case adaptation represents the number of similar cases needed to produce the final results. Of the three methods, MODAL-based CBR model has the least mean MAPE (0.1463) and mean RMSE (0.1880), followed by GA-based CBR model (MAPE =0.1730, RMSE=0.2128) and OLS-based CBR model (MAPE=0.1770, RMSE=0.2206).

Table 6.4 The result of the OLS-CBR model

K-fold	K=5		K=3		K=1	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
10	0.1791	0.2246	0.1817	0.2304	0.1995	0.2647
20	0.1768	0.2225	0.1794	0.2271	0.1974	0.2621
40	0.1776	0.2230	0.1809	0.2277	0.1998	0.2644
80	0.1766	0.2206	0.1817	0.2268	0.2022	0.2631
160	0.1766	0.2156	0.1822	0.2223	0.2013	0.2571
320	0.1755	0.2141	0.1810	0.2204	0.2012	0.2553
1450	0.1767	0.2236	0.1816	0.2305	0.1990	0.2644
Average	0.1770	0.2206	0.1812	0.2265	0.2001	0.2616

Table 6.5 The result of the GA -CBR model

K-fold	K=5		K=3		K=1	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
10	0.1719	0.2137	0.1762	0.2203	0.1956	0.2556
20	0.1719	0.2147	0.1750	0.2195	0.1966	0.2624
40	0.1739	0.2149	0.1760	0.2211	0.1947	0.2564
80	0.1737	0.2142	0.1752	0.2173	0.1982	0.2574
160	0.1733	0.2111	0.1791	0.2180	0.1998	0.2539
320	0.1722	0.2067	0.1777	0.2138	0.2003	0.2507
1450	0.1741	0.2140	0.1765	0.2219	0.1938	0.2554
Average	0.1730	0.2128	0.1765	0.2188	0.1970	0.2560

Table 6.6 The result of the MODAL-CBR model

K-fold	K=5		K=3		K=1	
	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
10	0.1563	0.1983	0.1492	0.1941	0.1759	0.2334
20	0.1523	0.1939	0.1460	0.1900	0.1750	0.2302
40	0.1530	0.1937	0.1454	0.1893	0.1751	0.2305
80	0.1532	0.1920	0.1472	0.1886	0.1754	0.2284
160	0.1515	0.1858	0.1453	0.1821	0.1732	0.2195
320	0.1519	0.1873	0.1459	0.1816	0.1739	0.2181
1450	0.1513	0.1940	0.1449	0.1906	0.1737	0.2311
Average	0.1528	0.1921	0.1463	0.1880	0.1746	0.2273

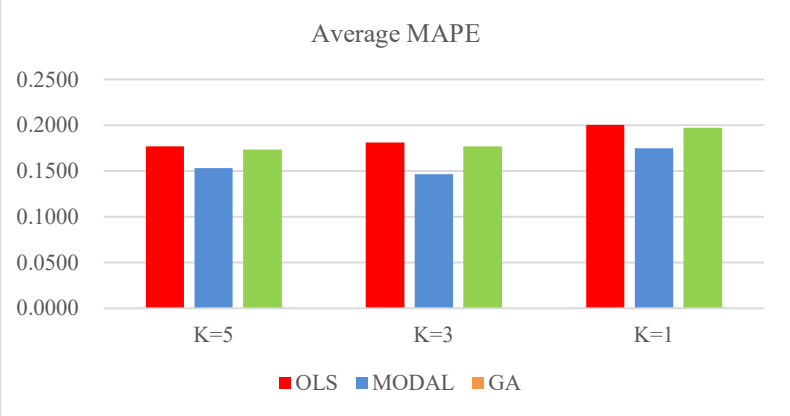


Figure 6.8 Average of MAPE of CBR model based on OLS,MA,GA

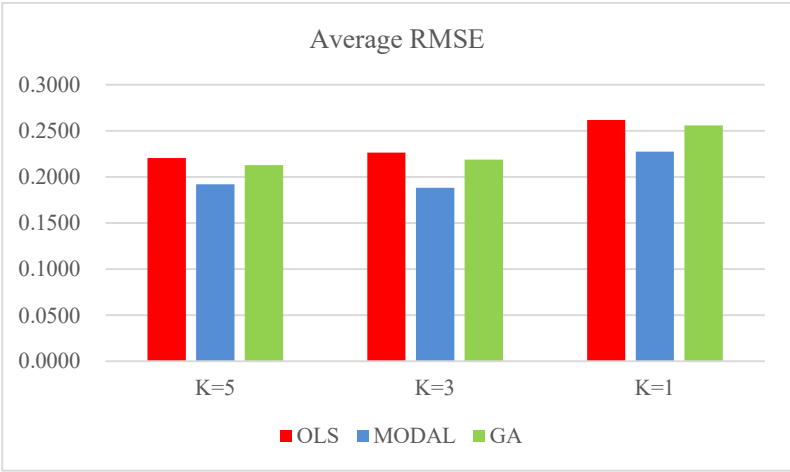


Figure 6.9 Average of RMSE of CBR model based on OLS, MA, GA

6.3 DISCUSSION

6.3.1 Weight Robustness

Different KF value in K-fold is used to measure the differences in the training set during each calculation. The larger the value of KF in K-fold, the more repeatability among training samples during the calculation of each fold. This means a smaller difference in the training set during each calculation. From Figure 6.1 to Figure 6.7, the trend of attribute weight during each calculation follows the fundamental principle that changes in weighting in each calculation decreases with the increase in the KF value of K-fold.

Despite the differences in the KF value of K-fold, Modal regression has the least variance compared with OLS and GA. MODAL shows excellent performance in weight stability. The variance of weight attributes during each calculation is used to measure the robustness of attribute weight. The smaller variance of weight attributes represents greater robustness of the CBR model. In the ECCE CBR model, the changes in the attribute weight of each fold represent the changes in case-base's structure caused by the difference in the training sample. For the long-term use of the system, it is necessary to reduce the fluctuations of the case-base's structure caused by the changes in the training sample while ensuring the model's performance. The weight determination in the ECCE CBR model should be capable of capturing the feature of the mainstream bulk of cases and reducing the unnecessary changes caused by minority or outliers of the data.

Another interesting finding is that GA is not suitable for weight determination in the ECCE CBR model, concerning long-term use. Despite GA's superiority in optimization for a single calculation, the result produced by GA is difficult to repeat, and inconsistent. This is expected to be caused by the randomization evolving process involved in GA. Case-base is deemed as the fundamental component in the CBR system. Since the structure of case-base affects case representation, knowledge storage and model implementation, the stable structure of case-base is significant for long-term use of the CBR model for ECCE.

6.3.2 Comparison of Three Weight Determination Methods

Compared with other methods, MODAL regression shows excellent performance in weight stability. Figure 6.1 to Figure 6.7 illustrates the superiority of

MODAL concerning robustness. Despite the difference in the value of K, the MODAL-CBR model has the most stable performance in weight determination. MODAL-CBR model can reduce the sensitivity of the noisy data and support long-term use of the ECCE CBR model.

The influence of the changes in the training sample on the MODAL-CBR model is not as significant as on the OLS-CBR model and GA-CBR model. The accuracy differs due to the different values of case adaptation. The OLS and GA show the best results when the parameter of case adaptation is 5, while MODAL has the best performance when the setting of case adaptation is 3. The setting of case adaptation represents the number of similar cases needed to produce the final results. The larger the parameter of case adaptation, the more adaptation capability requirement of the model. Accordingly, the MODAL-CBR model requires less case adaptation.

As a robust regression based on the conditional mode of the response value, MODAL takes the value with the highest probability of occurrence and thus provides a more stable performance than other weight determination methods (Yao & Li, 2014). Accordingly, the MODAL-CBR model focuses on the main characteristics of the cases and maintains the stable structure of case-base despite the changes in the cases. The results show that the MODAL-CBR model significantly improves the MAPE from 0.17 to 0.1463 and RMSE from 0.22 to 0.18.

6.4 CHAPTER SUMMARY

This chapter introduces a robust weight determination method for improving the long-term use of the ECCE CBR model. The MODAL regression is used in weight determination to increase the model's robustness. K-folder cross-validation is used for the model validation, and different KF values of 10, 20, 40, 80, 160, 320 and 1450 are used, given their influence on the training sample.

The results show that the total above-floor area is the most significant attribute in all weight determination methods, followed by total underground floor area. The significance of remaining attributes differs due to the methods used. The third-most significant attribute differs due to the weight determination methods. Except for MODAL, duration and the total height of the building are the least weight for ECCE.

The model is evaluated from robustness and accuracy perspectives. Variance of weight attribute of each calculation is used for measuring the weight robustness. The

measure of estimation accuracy includes the MAPE and RMSE. The results are compared with those using the OLS and GA in weight determination, followed by the discussion on the overall robustness of the weight determination. From the robustness perspective, MODAL-CBR model has the least attribute variance compared with those with OLS and GA. The results show the superiority of the MODAL-CBR model in weighting stability. From the accuracy perspective, the results showed that the MODAL- CBR model has the least mean MAPE and mean RMSE, followed by the GA-based CBR model and OLS-based CBR model.

Chapter 7: Improving the ECCE CBR Model by using CBM

7.1 INTRODUCTION

The previous chapter represents the results of OLS-CBR, GA-CBR and MODLR-CBR. The comparative results with respect to robustness and prediction accuracy are provided. The results show that the MODAL-CBR model has the largest robustness compared with those with OLS and GA, which illustrates the superiority of the MODAL-CBR model in weighting stability. From the accuracy perspective, the results showed that the MODAL-CBR model has the least mean MAPE and mean RMSE, followed by the GA-based CBR model and OLS-based CBR model.

Except for robustness, efficiency is also a significant factor to be considered during the long-term running of the ECCE CBR system. A successful ECCE CBR model should not only provide accurate estimation results but should also be easily and conveniently maintained. Despite the advantage of CBR in ECCE, several practical issues need to be considered when using the CBR system to solve the real-world problem. These issues include the limit of the case-base size, the trade-off between long-term and short-term performance, the expected distribution of future problems and the availability of sources of cases (Leake & Wilson, 2000). Among those, the limit of the case-base received most research attention because it is the most common problem when using the CBR system.

During the long-term operation of the ECCE CBR model, the large storage requirement and slow efficiency caused by the continuous increase in the size of the case-base should be carefully addressed. When the size of the case-base becomes large, the requirement of data storage becomes intensive and the retrieval process gets slow. This trend will significantly impair the performance of the ECCE CBR system if not being properly handled. This inevitably raises the question of how to select cases for avoiding excessive storage and time complexity, and possibly to maintain the system's performance by reducing error. The success of the CBR system is not only measured by the range of problems that can be satisfactorily solved, but also by the storage and answer time that is needed to generate solutions for case targets. Thus, this chapter

introduces a CBM approach for case-base reduction during the long-term use of the ECCE CBR model.

An original method based on a CBM for case-base editing is introduced in this chapter, to cleanse the training dataset and improve the efficiency of the ECCE CBR model. CBM is a process of updating and refining the case-base. It is used for revising and adjusting the knowledge of the case-base to facilitate the reasoning system to improve the performance of the CBR system. Case-based editing methods are proposed based on the basic concepts of CBM introduced in Chapter 3. Given a case set X=$\{x_1, x_2, \dots, x_n\}$, let $x \in$ X be a case whose cost should be estimated. The CBM method starts with calculating the coverage contribution weight for any given K, the case adaptation value used in the CBR model. Then, for each training case x_i in training set X_r =$\{x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n\}$, calculating its coverage N (x_i) in the case-base. The next step is calculating the weighted coverage contribution of each training case x_i in the case-base X_r. Later, the weighted coverage contribution is ranked and the threshold of weighted coverage contribution is determined. The next step is to select a subset D whose weighted coverage contribution is lower than a threshold. Finally, the case-base is updated by deleting those in D. Briefly, given one training case, the editing rules consider either to delete or keep the case unchanged according to their weighted coverage contribution in the case-base. The cases with the lowest coverage contribution will be eliminated.

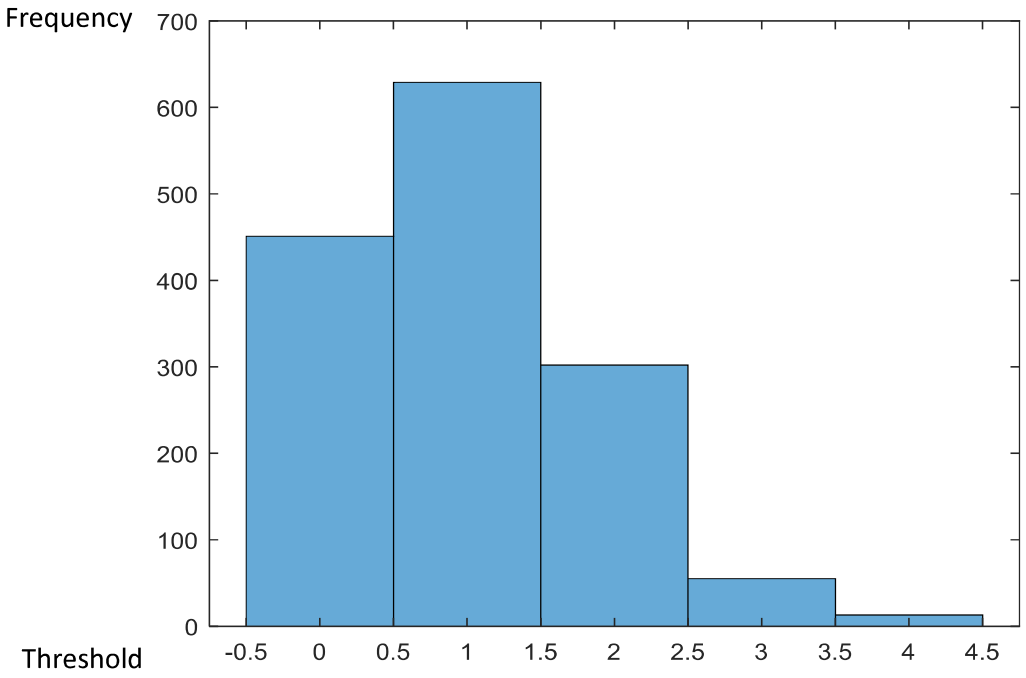
7.2 RESULTS

7.2.1 Edting Threshold

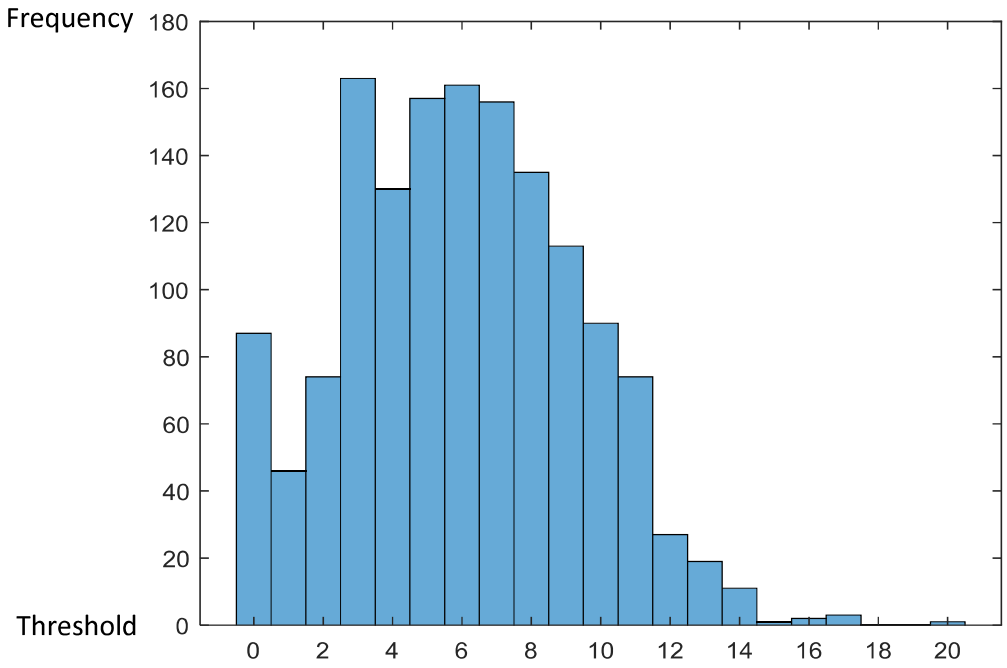
To perform the case-base editing strategy, the first step is to calculate the converage contribution weight for any given K. In this study, the K values of 1,3,5 are used. The coverge weight is determined consequently by the K value based on Equation 3.3. The second step is to calculate the weighted coverage contribution of each of the training cases. Figure 7.1 to Figure 7.3 shows the range of the coverage contribution in different models. In the OLS-CBR model, the range of weighted coverage contribution of the case-base is [0,4] when K is one, the range of weighted coverage contribution of the case-base is [0,20] when K is three, the range of weight coverage contribution of the casebase is [0,45] when K is five. In the GA-CBR model, the range of weighted coverage contribution of the case-base is [0,5] when K is one,

the range of weighted coverage contribution of the case-base is [0,19] when K is three, the range of weight coverage contribution of the casebase is [0,43] when K is five. In the MODAL-CBR model, the range of weighted coverage contribution of the case-base is [0,5] when K is one, the range of weighted coverage contribution of the case-base is [0,17] when K is three, the range of weight coverage contribution of the casebase is [0,40] when K is five.

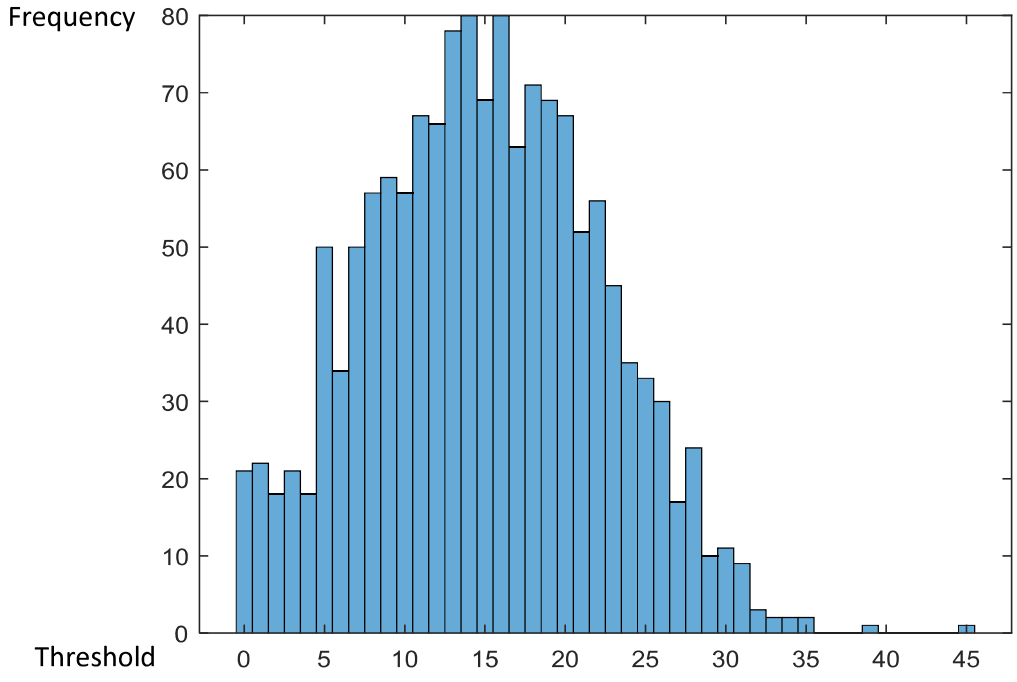
The range of the coverage contribution differs in different models. It can be seen that, with the increase in the K value, the range of the coverage contribution increases. The edting threshold should be in the range of the coverage contribution. For example, if the coverge contribution range is [0,5], then the editing thredholds can be zero, one, two, three, and four. Note that the editing threshhold should be less than the maximum value of the coverage contribution so that the case-base is not empty. Considering the coverage range of the case-base, this study selects 3, 15 and 30 as the maximum value of the editing threshold when K is 1, 3, 5 seperately. Therefore, the editing threshhold ranges from 1 to 30 when K equals 5; ranges from 1 to 15 when K equals 3; and ranges from 1 to 3 when K equals 1. Leave-one cross-validation is used for model validation.



(a)

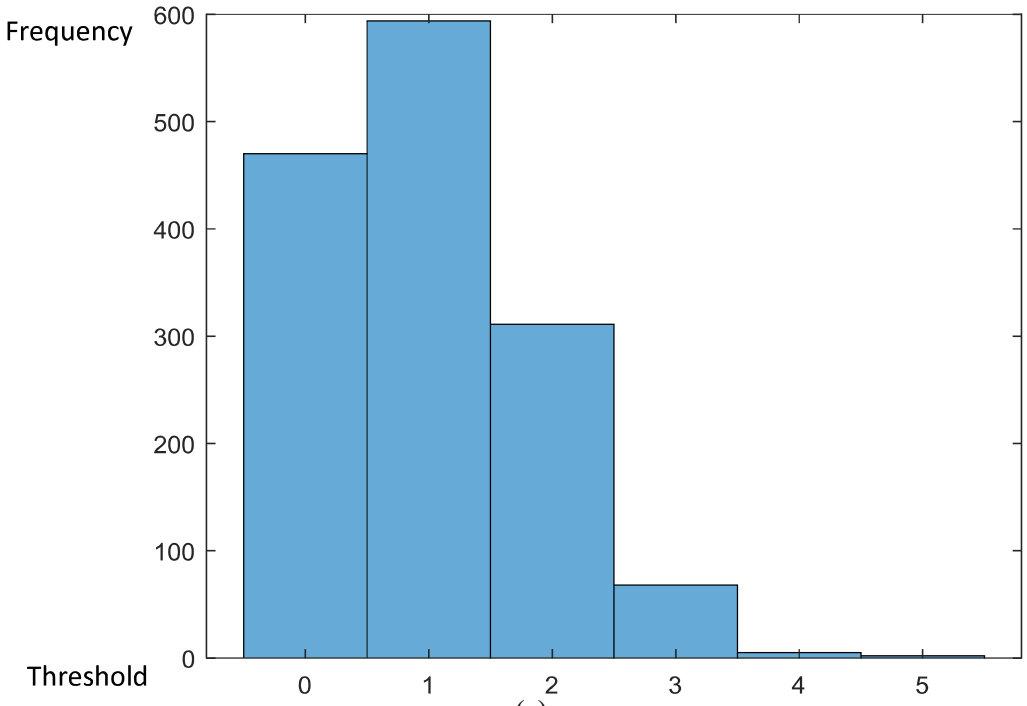


(b)

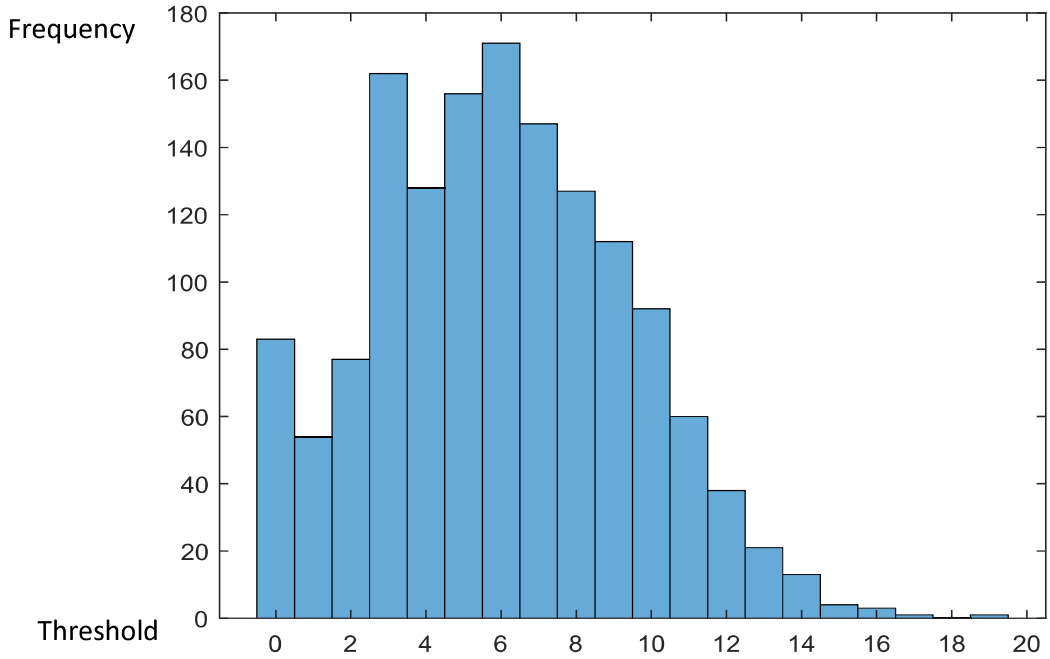


(c)

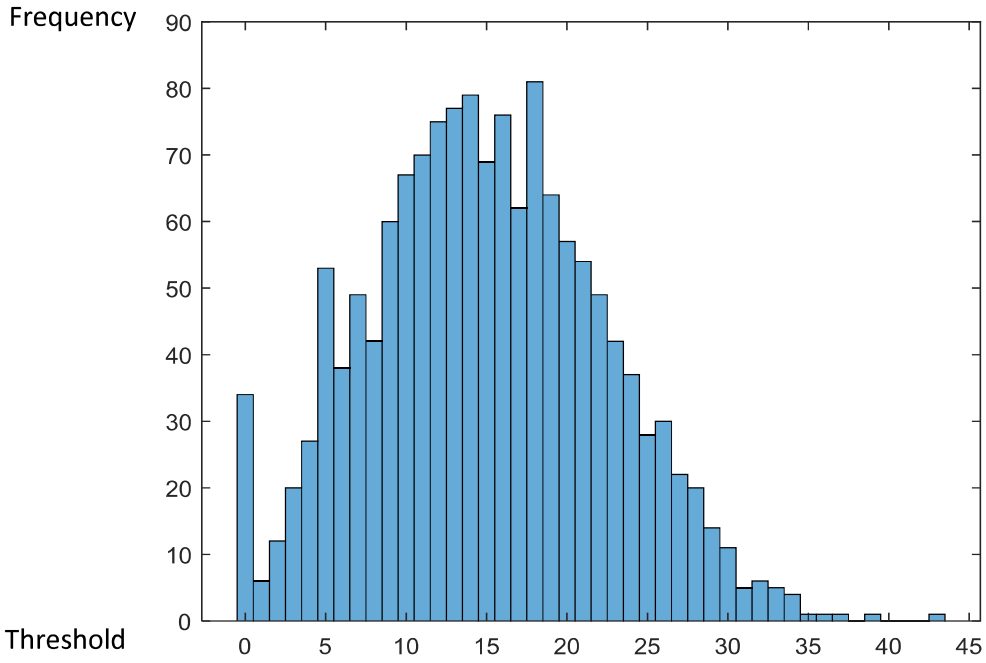
Figure 7.1 The histogram of coverage contribution of case-base in OLS-CBR model: (a) K=1; (b) K=3; (c) K=5;


Frequency
600
500
400
300
200
100
0
Threshold
0
1
2
3
4
5

(a)

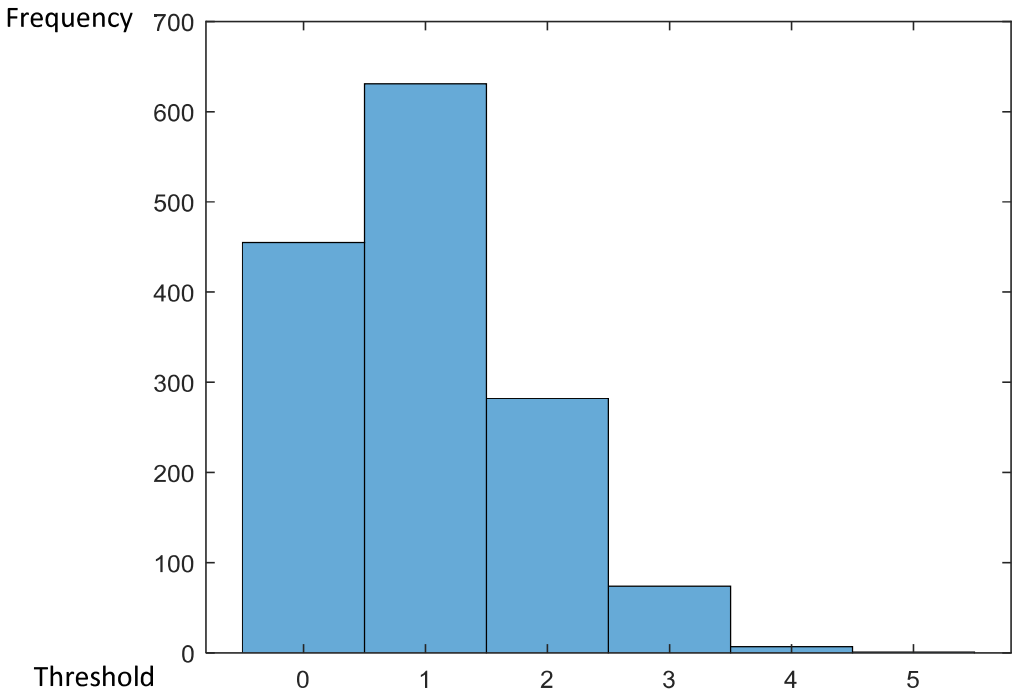

Frequency
180
160
140
120
100
80
60
40
20
0
Threshold
0
2
4
6
8
10
12
14
16
18
20

(b)

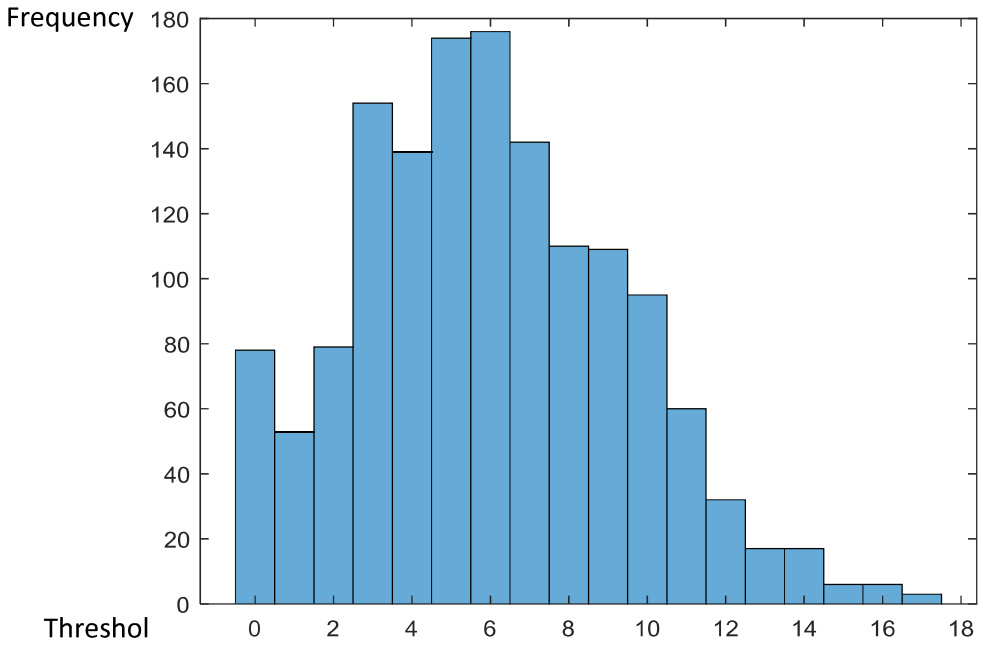


(c)

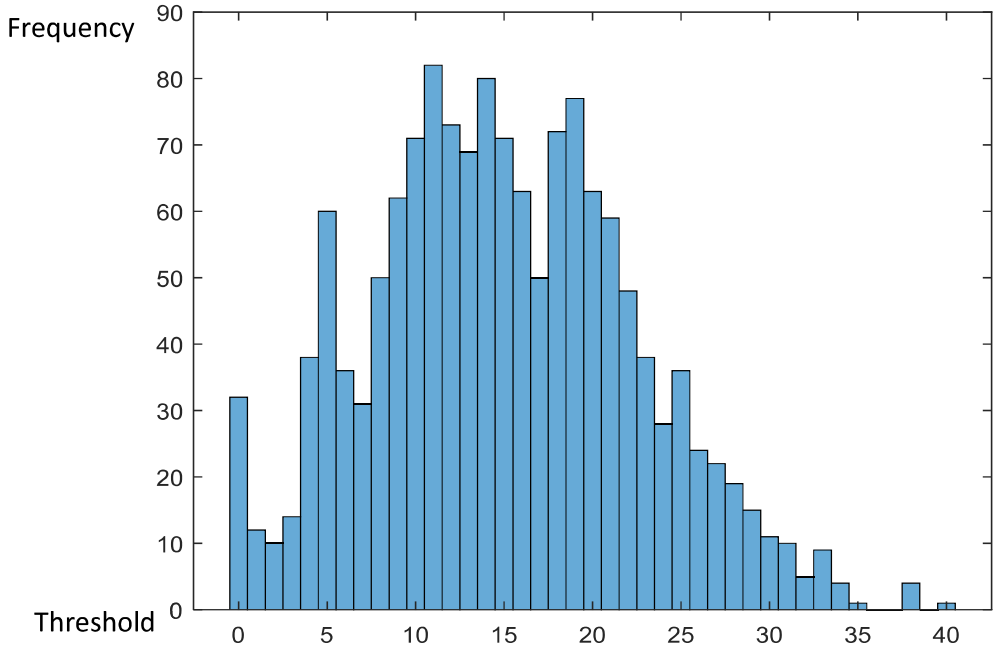
Figure 7.2 The histogram of coverage contribution of case-base in GA-CBR model: (a) K=1; (b) K=3; (c) K=5;



(a)



(b)



(c)

Figure 7.3 The histogram of coverage contribution of case-base in MODAL-CBR model: (a) K=1; (b) K=3; (c) K=5;

7.2.2 Compression Ratio

The performance of the ECCE CBR model was compared before and after using the proposed CBM approach. After editing the case-base, the attribute weight and the case-base both have been updated. Thus, there are three options in parameter setting for the CBR ECCE model: new weight and new training dataset (strategy 1); new weight and original training case-base (strategy 2); original attribute weight and optimized training dataset (strategy 3).

Table 7.1 The number of cases in the case-base after case-base editing

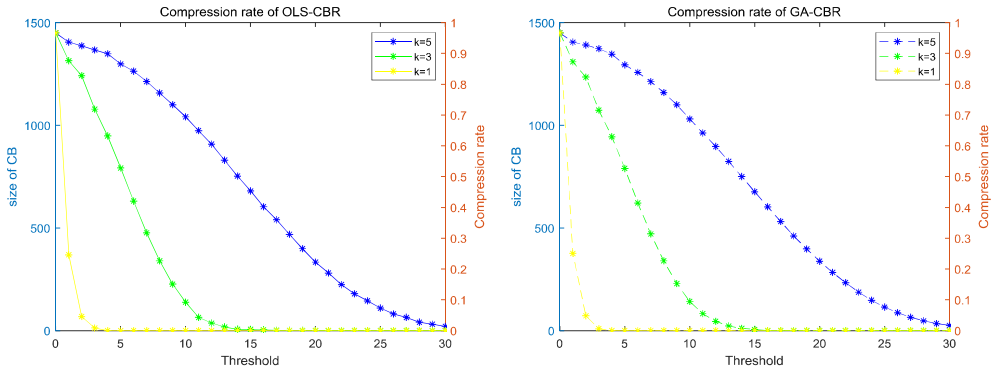
The No. of cases	OLS-CBR			GA-CBR			MODLR-CBR		
Threshold	K=1	K=3	K=5	K=1	K=3	K=5	K=1	K=3	K=5
1	369	1315	1405	376	1309	1405	364	1317	1404
2	68	1242	1387	73	1235	1391	82	1237	1394
3	13	1079	1367	9	1074	1373	7	1085	1379
4	-	948	1348	-	944	1346	-	946	1341
5	-	792	1299	-	790	1295	-	771	1283
6	-	631	1264	-	622	1258	-	598	1247
7	-	476	1213	-	471	1213	-	454	1214
8	-	340	1158	-	340	1159	-	344	1165
9	-	226	1100	-	229	1100	-	236	1103
10	-	138	1041	-	142	1031	-	142	1034
11	-	65	974	-	83	963	-	81	953
12	-	37	908	-	45	897	-	48	878
13	-	18	830	-	23	824	-	30	809
14	-	7	753	-	11	750	-	16	729
15	-	6	681	-	6	676	-	9	659
16	-	-	603	-	-	603	-	-	595
17	-	-	540	-	-	531	-	-	543
18	-	-	469	-	-	461	-	-	475
19	-	-	399	-	-	398	-	-	395
20	-	-	334	-	-	338	-	-	334
21	-	-	280	-	-	283	-	-	272
22	-	-	224	-	-	233	-	-	226
23	-	-	180	-	-	187	-	-	190
24	-	-	145	-	-	148	-	-	161
25	-	-	110	-	-	115	-	-	124
26	-	-	81	-	-	87	-	-	101
27	-	-	65	-	-	65	-	-	80
28	-	-	41	-	-	48	-	-	60
29	-	-	31	-	-	34	-	-	45
30	-	-	20	-	-	25	-	-	33

- represents the not applicable

Table 7.2 The compression rate based on different thresholds

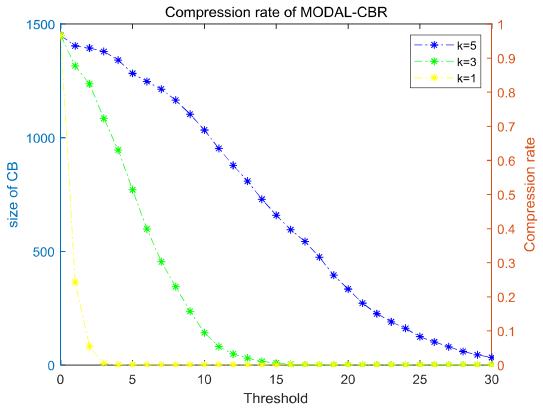
Compression rate	OLS-CBR			GA-CBR			MODLR-CBR		
threshold	K=1	K=3	K=5	K=1	K=3	K=5	K=1	K=3	K=5
1	0.25	0.91	0.97	0.26	0.90	0.97	0.25	0.91	0.97
2	0.05	0.86	0.96	0.05	0.85	0.96	0.06	0.85	0.96
3	0.01	0.74	0.94	0.01	0.74	0.95	0.01	0.75	0.95
4	-	0.65	0.93	-	0.65	0.93	-	0.65	0.93
5	-	0.55	0.90	-	0.55	0.89	-	0.53	0.89
6	-	0.44	0.87	-	0.43	0.87	-	0.41	0.86
7	-	0.33	0.84	-	0.33	0.84	-	0.31	0.84
8	-	0.23	0.80	-	0.23	0.80	-	0.24	0.80
9	-	0.16	0.76	-	0.16	0.76	-	0.16	0.76
10	-	0.10	0.72	-	0.10	0.71	-	0.10	0.71
11	-	0.04	0.67	-	0.06	0.66	-	0.06	0.66
12	-	0.03	0.63	-	0.03	0.62	-	0.03	0.61
13	-	0.01	0.57	-	0.02	0.57	-	0.02	0.56
14	-	0.00	0.52	-	0.01	0.52	-	0.01	0.50
15	-	0.00	0.47	-	0.00	0.47	-	0.01	0.45
16	-	-	0.42	-	-	0.42	-	-	0.41
17	-	-	0.37	-	-	0.37	-	-	0.37
18	-	-	0.32	-	-	0.32	-	-	0.33
19	-	-	0.28	-	-	0.27	-	-	0.27
20	-	-	0.23	-	-	0.23	-	-	0.23
21	-	-	0.19	-	-	0.20	-	-	0.19
22	-	-	0.15	-	-	0.16	-	-	0.16
23	-	-	0.12	-	-	0.13	-	-	0.13
24	-	-	0.10	-	-	0.10	-	-	0.11
25	-	-	0.08	-	-	0.08	-	-	0.09
26	-	-	0.06	-	-	0.06	-	-	0.07
27	-	-	0.04	-	-	0.04	-	-	0.06
28	-	-	0.03	-	-	0.03	-	-	0.04
29	-	-	0.02	-	-	0.02	-	-	0.03
30	-	-	0.01	-	-	0.02	-	-	0.02

- represents the not applicable



(a)

(b)



(c)

Figure 7.4 Compression rate based on Threshold

Table 7.1 and Table 7.2 shows the number of cases and the compression rate in the case-base after editing. For example, when K and threshold and are set as five and one, the number of the cases left in the case-base is 1405 and the compression rate is 0.97; when K and threshold are set to be five and 10 separately, the number of the cases left in the case-base is 1401 and the compression rate is 0.72;

Figure 7.4 shows how the size of the case-base and the compression rate changes based on different thresholds of coverage contribution. The different case adaptation values generate different compression processes. When case adaptation is set to five, the compression rate can be in the range between 1% to 97%. When case adaptation is set to three, the compression rate can be in the range between 1% to 91%. When case adaptation is set to one, the compression rate can be in the range between 1% to 25%.

When conducting case-base editing, the ideal compression process should provide an extensive compression range offering sufficient choices. When the K value is small (K=1), the rapid compression process sacrifices the compression range. This means the editing of the case-base is very inflexible, and there is not much space for choosing the optimal compression ratio. When the K value is larger (K=5), the compression step is smaller, which can provide a broader range of compression interval, which provides more options.

7.2.3 Weight Calculation

After the case-base editing, the attribute weight is updated based on different thresholds. Table 7.3 to Table 7.5 presents the average value of weight according to different editing thresholds when K is set to five. Table 7.6 to Table 7.8 presents the average value of weight according to different editing threshold when K is set to three. Table 7.9 to Table 7.11 presents the average value of weight according to different editing threshold when K is set to one.

Table 7.3 The attribute weight of OLS-CBR after case-base editing (K=5)

OLS-CBR (K=5)						
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.6396	0.2066	0.0349	0.0611	0.0491	0.0086
2	0.6384	0.2089	0.0378	0.0593	0.0486	0.0069
3	0.6374	0.2092	0.0324	0.0604	0.0488	0.0118
4	0.6382	0.2101	0.0258	0.0598	0.0486	0.0175
5	0.6284	0.2076	0.0525	0.0593	0.0478	0.0044
6	0.6124	0.2162	0.0593	0.0564	0.0459	0.0098
7	0.6101	0.2202	0.0608	0.0531	0.0457	0.0100
8	0.6111	0.2284	0.0582	0.0467	0.0484	0.0072
9	0.6129	0.2471	0.0481	0.0399	0.0493	0.0027
10	0.5959	0.2340	0.0708	0.0348	0.0471	0.0174
11	0.6005	0.2235	0.0688	0.0389	0.0497	0.0187
12	0.6046	0.2204	0.0644	0.0443	0.0497	0.0165
13	0.5378	0.2015	0.1165	0.0323	0.0458	0.0662
14	0.5248	0.1820	0.1279	0.0418	0.0434	0.0801
15	0.5266	0.1723	0.1262	0.0466	0.0456	0.0826
16	0.5082	0.2296	0.1185	0.0269	0.0438	0.0729
17	0.5303	0.2311	0.1057	0.0284	0.0477	0.0569
18	0.5962	0.2592	0.0563	0.0288	0.0572	0.0023
19	0.5261	0.2176	0.1085	0.0337	0.0510	0.0631
20	0.5407	0.2544	0.0877	0.0205	0.0540	0.0427
21	0.5605	0.2667	0.0773	0.0025	0.0594	0.0336
22	0.5200	0.3137	0.0830	0.0035	0.0595	0.0202
23	0.4982	0.2628	0.1210	0.0102	0.0457	0.0621

24	0.4302	0.3948	0.0680	0.0551	0.0493	0.0026
25	0.3977	0.4087	0.0537	0.0852	0.0503	0.0044
26	0.3327	0.4084	0.0174	0.1004	0.0534	0.0878
27	0.3780	0.2876	0.0880	0.0723	0.0493	0.1249
28	0.5435	0.1039	0.0862	0.0833	0.0708	0.1122
29	0.5121	0.0986	0.0499	0.1953	0.1056	0.0385
30	0.3813	0.3095	0.0942	0.1234	0.0630	0.0285

Table 7.4 The attribute weight of GA-CBR after case-base editing (K=5)

GA-CBR (K=5)						
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.6278	0.2772	0.0307	0.0250	0.0165	0.0228
2	0.6270	0.2771	0.0308	0.0251	0.0166	0.0233
3	0.6261	0.2782	0.0309	0.0253	0.0162	0.0232
4	0.6267	0.2793	0.0313	0.0241	0.0160	0.0226
5	0.6251	0.2823	0.0299	0.0243	0.0168	0.0216
6	0.6236	0.2836	0.0305	0.0234	0.0173	0.0216
7	0.6242	0.2832	0.0309	0.0232	0.0177	0.0207
8	0.6214	0.2851	0.0303	0.0233	0.0185	0.0214
9	0.6077	0.3051	0.0290	0.0192	0.0184	0.0205
10	0.6014	0.3129	0.0295	0.0190	0.0173	0.0198
11	0.6022	0.3114	0.0292	0.0195	0.0176	0.0202
12	0.6031	0.3117	0.0292	0.0190	0.0173	0.0196
13	0.6060	0.3083	0.0295	0.0192	0.0180	0.0189
14	0.6126	0.3021	0.0300	0.0180	0.0181	0.0191
15	0.6141	0.2998	0.0303	0.0172	0.0177	0.0208
16	0.6149	0.3013	0.0306	0.0145	0.0185	0.0201
17	0.6131	0.3057	0.0297	0.0125	0.0194	0.0196
18	0.6097	0.3106	0.0282	0.0109	0.0201	0.0205
19	0.6081	0.3120	0.0291	0.0110	0.0209	0.0188
20	0.6058	0.3143	0.0284	0.0102	0.0225	0.0187
21	0.6069	0.3143	0.0274	0.0093	0.0242	0.0179
22	0.6000	0.3223	0.0266	0.0090	0.0238	0.0183
23	0.5985	0.3234	0.0278	0.0076	0.0246	0.0180
24	0.6035	0.3132	0.0299	0.0068	0.0269	0.0196
25	0.6046	0.3093	0.0309	0.0071	0.0262	0.0220
26	0.6040	0.3088	0.0314	0.0066	0.0256	0.0236
27	0.5937	0.3152	0.0326	0.0075	0.0264	0.0246
28	0.5801	0.3285	0.0335	0.0068	0.0268	0.0243
29	0.5650	0.3360	0.0359	0.0085	0.0292	0.0254
30	0.5536	0.3429	0.0366	0.0087	0.0317	0.0266

Table 7.5 The attribute weight of MODLR-CBR after case-base editing (K=5)

MODLR-CBR (K=5)						
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.4944	0.1863	0.1000	0.0155	0.0240	0.1797
2	0.4943	0.1854	0.1000	0.0165	0.0239	0.1798
3	0.4928	0.1850	0.1001	0.0176	0.0244	0.1800
4	0.4891	0.1840	0.0978	0.0194	0.0232	0.1865
5	0.4886	0.1792	0.1033	0.0230	0.0236	0.1824
6	0.4957	0.1803	0.1014	0.0237	0.0233	0.1756
7	0.4869	0.1784	0.1063	0.0215	0.0234	0.1836
8	0.4904	0.1773	0.1068	0.0230	0.0252	0.1773
9	0.6135	0.2234	0.0844	0.0382	0.0305	0.0101
10	0.6242	0.2316	0.0516	0.0364	0.0317	0.0246
11	0.5050	0.1899	0.0903	0.0281	0.0266	0.1602
12	0.4752	0.1972	0.1022	0.0384	0.0249	0.1621
13	0.4687	0.2025	0.1034	0.0362	0.0258	0.1634
14	0.4532	0.2044	0.1066	0.0361	0.0252	0.1745
15	0.4468	0.2089	0.1097	0.0287	0.0279	0.1780
16	0.5154	0.0722	0.1401	0.0352	0.0327	0.2043
17	0.5153	0.0750	0.1355	0.0410	0.0362	0.1970
18	0.5769	0.0230	0.1367	0.0031	0.0326	0.2277
19	0.5563	0.0132	0.1488	0.0170	0.0236	0.2410
20	0.5313	0.0551	0.1359	0.0313	0.0252	0.2212
21	0.5032	0.1895	0.1144	0.0257	0.0303	0.1369
22	0.5504	0.0039	0.1973	0.0297	0.0307	0.1880
23	0.5696	0.0061	0.1900	0.0200	0.0226	0.1917
24	0.4879	0.1754	0.1455	0.0174	0.0220	0.1517
25	0.3473	0.1579	0.2462	0.0652	0.0133	0.1701
26	0.1777	0.1789	0.3135	0.0628	0.0090	0.2580
27	0.1860	0.4404	0.1653	0.1047	0.0113	0.0922
28	0.3674	0.0127	0.3271	0.0472	0.0096	0.2361
29	0.5849	0.0503	0.2064	0.0853	0.0270	0.0460
30	0.3339	0.0062	0.3021	0.0853	0.0392	0.2333

Table 7.6 The attribute weight of OLS-CBR after case-base editing (K=3)

	OLS-CBR (K=3)					
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.6313	0.2189	0.0281	0.0600	0.0474	0.0142
2	0.6233	0.2321	0.0303	0.0507	0.0481	0.0155
3	0.6359	0.2164	0.0255	0.0513	0.0507	0.0202
4	0.6383	0.2140	0.0250	0.0517	0.0506	0.0203
5	0.5920	0.2742	0.0234	0.0323	0.0518	0.0262
6	0.4647	0.2381	0.1350	0.0297	0.0379	0.0946
7	0.4739	0.3111	0.1106	0.0112	0.0406	0.0526
8	0.4708	0.2857	0.1226	0.0203	0.0425	0.0581
9	0.3868	0.3611	0.1342	0.0112	0.0339	0.0730
10	0.3803	0.4407	0.0436	0.0645	0.0360	0.0349
11	0.4424	0.3478	0.0066	0.0584	0.0504	0.0944
12	0.5045	0.0397	0.2149	0.0074	0.0693	0.1642
13	0.4868	0.0202	0.1459	0.0805	0.0736	0.1929
14	0.1686	0.3271	0.0999	0.3229	0.0557	0.0258
15	0.2169	0.3047	0.0000	0.3080	0.0539	0.1165

Table 7.7 The attribute weight of GA-CBR after case-base editing (K=3)

	GA-CBR (K=3)					
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.6253	0.2792	0.0315	0.0247	0.0173	0.0221
2	0.6248	0.2819	0.0304	0.0240	0.0169	0.0220
3	0.6258	0.2808	0.0293	0.0250	0.0186	0.0205
4	0.6269	0.2768	0.0304	0.0269	0.0180	0.0210
5	0.6152	0.2973	0.0296	0.0191	0.0177	0.0209
6	0.5993	0.3205	0.0293	0.0158	0.0150	0.0201
7	0.5989	0.3240	0.0294	0.0133	0.0155	0.0189
8	0.6027	0.3216	0.0286	0.0102	0.0183	0.0186
9	0.5912	0.3299	0.0300	0.0095	0.0195	0.0199
10	0.5762	0.3465	0.0308	0.0083	0.0188	0.0194
11	0.5924	0.3177	0.0321	0.0126	0.0246	0.0207
12	0.6088	0.2705	0.0382	0.0243	0.0312	0.0270
13	0.6004	0.2510	0.0444	0.0271	0.0453	0.0318
14	0.5551	0.2535	0.0549	0.0274	0.0662	0.0428
15	0.5112	0.2062	0.0721	0.0479	0.1005	0.0620
16	0.6253	0.2792	0.0315	0.0247	0.0173	0.0221

Table 7.8 The attribute weight of MODAL-CBR after case-base editing (K=3)

MODAL-CBR (K=3)						
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.4921	0.1816	0.1044	0.0191	0.0241	0.1786
2	0.4918	0.1799	0.1070	0.0202	0.0241	0.1770
3	0.4898	0.1742	0.1121	0.0223	0.0251	0.1764
4	0.4803	0.1748	0.1072	0.0298	0.0237	0.1842
5	0.5767	0.0500	0.1181	0.0124	0.0290	0.2138
6	0.5930	0.2508	0.0487	0.0543	0.0333	0.0199
7	0.3213	0.1302	0.2305	0.0088	0.0223	0.2869
8	0.4013	0.1757	0.1420	0.0295	0.0233	0.2283
9	0.5709	0.2512	0.0122	0.0488	0.0215	0.0953
10	0.5402	0.1217	0.1274	0.0372	0.0109	0.1627
11	0.2692	0.4087	0.1306	0.0793	0.0025	0.1098
12	0.3210	0.3161	0.1149	0.0899	0.0194	0.1388
13	0.2325	0.4075	0.1339	0.0656	0.0212	0.1393

Table 7.9 The attribute weight of OLS -CBR after case-base editing (K=1)

OLS-CBR (K=1)						
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.5236	0.3484	0.0006	0.0128	0.0487	0.0658
2	0.3254	0.0086	0.3283	0.0285	0.0305	0.2787
3	0.1640	0.0449	0.3626	0.0569	0.0131	0.3585

Table 7.10 The attribute weight of GA -CBR after case-base editing (K=1)

GA-CBR (K=1)						
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.6158	0.2877	0.0262	0.0251	0.0179	0.0272
2	0.6053	0.2985	0.0360	0.0205	0.0152	0.0245
3	0.5827	0.2798	0.0391	0.0228	0.0441	0.0315

Table 7.11 The attribute weight of MODAL-CBR after case-base editing (K=1)

MODAL-CBR (K=1)						
Threshold	weight1	weight2	weight3	weight4	weight5	weight6
1	0.3595	0.1690	0.1776	0.0251	0.0243	0.2445
2	0.5212	0.0062	0.2502	0.0046	0.0124	0.2054
3	0.0436	0.3728	0.2501	0.1441	0.0125	0.1769

7.2.4 Error Rate

Tables 7.12 to 7.14 show results of the MAPE and RMSE of OLS-, GA-, MODAL-CBR (K=5) based on the editing threshold ranging from 1 to 30. For example, Table 7.12 gives the results of the MAPE and RMSE in the OLS-CBR (K=5) based on the editing threshold ranging from 1 to 30. When the case adaptation and threshold are

set 5 and 10, the number of cases in the case-base after case-base editing is 1041, the compression rate is 0.7187, the MAPE and RMSE of the Strategy 1 are 0.1747 and 0.2213; the MAPE and RMSE of the Strategy 2 are 0.1809 and 0.2305; the MAPE and RMSE of the Strategy 3 are 0.1671 and 0.2305. The remaining results of other parameter settings (K=3,K=1) can be seen in the Appendix C-1 to C-6.

Table 7.12 The error rate of OLS -CBR after case-base editing (K=5)

OLS-CBR (K=5)								
Threshold	No. of cases	CR	Strategy 1		Strategy 2		Strategy 3	
			MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	1405	0.9698	0.1750	0.2219	0.1764	0.2238	0.1750	0.2238
2	1387	0.9573	0.1751	0.2217	0.1768	0.2250	0.1744	0.2250
3	1367	0.9434	0.1730	0.2188	0.1772	0.2252	0.1720	0.2252
4	1348	0.9305	0.1731	0.2188	0.1772	0.2252	0.1719	0.2252
5	1299	0.8965	0.1746	0.2204	0.1769	0.2249	0.1724	0.2249
6	1264	0.8720	0.1743	0.2202	0.1769	0.2253	0.1720	0.2253
7	1213	0.8373	0.1756	0.2215	0.1781	0.2272	0.1721	0.2272
8	1158	0.7989	0.1755	0.2219	0.1786	0.2285	0.1719	0.2285
9	1100	0.7593	0.1757	0.2226	0.1797	0.2288	0.1702	0.2288
10	1041	0.7187	0.1747	0.2213	0.1809	0.2305	0.1671	0.2305
11	974	0.6722	0.1774	0.2237	0.1823	0.2312	0.1689	0.2312
12	908	0.6268	0.1792	0.2248	0.1838	0.2324	0.1680	0.2324
13	830	0.5730	0.1802	0.2263	0.1882	0.2357	0.1640	0.2357
14	753	0.5195	0.1810	0.2284	0.1887	0.2366	0.1648	0.2366
15	681	0.4697	0.1833	0.2329	0.1899	0.2370	0.1652	0.2370
16	603	0.4164	0.1869	0.2351	0.1912	0.2389	0.1679	0.2389
17	540	0.3727	0.1891	0.2380	0.1910	0.2389	0.1704	0.2389
18	469	0.3236	0.1921	0.2390	0.1944	0.2412	0.1763	0.2412
19	399	0.2755	0.1918	0.2398	0.1946	0.2422	0.1700	0.2422
20	334	0.2302	0.1950	0.2452	0.1971	0.2451	0.1697	0.2451
21	280	0.1935	0.1983	0.2493	0.1967	0.2444	0.1669	0.2444
22	224	0.1546	0.2028	0.2544	0.2014	0.2520	0.1621	0.2520
23	180	0.1239	0.2089	0.2622	0.2037	0.2567	0.1629	0.2567
24	145	0.0999	0.2066	0.2651	0.2044	0.2577	0.1676	0.2577
25	110	0.0758	0.2144	0.2738	0.2074	0.2606	0.1724	0.2606
26	81	0.0562	0.2230	0.2817	0.2159	0.2677	0.1711	0.2677
27	65	0.0446	0.2553	0.3139	0.2425	0.2944	0.1734	0.2944
28	41	0.0284	0.2953	0.3570	0.2824	0.3366	0.1731	0.3366
29	31	0.0214	0.3072	0.3567	0.2939	0.3454	0.1803	0.3454
30	20	0.0140	0.4739	0.4724	0.4782	0.4752	0.1775	0.4752

CR is the compression rate.

Table 7.13 The error rate of GA -CBR after case-base editing (K=5)

GA-CBR (K=5)								
Threshold	No. of cases	CR	Strategy 1		Strategy 2		Strategy 3	
			MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	1405	0.9698	0.1715	0.2145	0.1728	0.2146	0.1704	0.2136
2	1387	0.9573	0.1738	0.2178	0.1729	0.2147	0.1728	0.2171
3	1367	0.9434	0.1731	0.2170	0.1724	0.2142	0.1727	0.2166
4	1348	0.9305	0.1711	0.2139	0.1730	0.2147	0.1698	0.2133
5	1299	0.8965	0.1743	0.2184	0.1743	0.2160	0.1717	0.2166
6	1264	0.8720	0.1747	0.2180	0.1745	0.2164	0.1725	0.2160
7	1213	0.8373	0.1743	0.2194	0.1744	0.2163	0.1723	0.2174
8	1158	0.7989	0.1759	0.2184	0.1745	0.2163	0.1723	0.2158
9	1100	0.7593	0.1751	0.2191	0.1751	0.2181	0.1722	0.2157
10	1041	0.7187	0.1803	0.2243	0.1769	0.2202	0.1739	0.2167
11	974	0.6722	0.1802	0.2228	0.1775	0.2210	0.1729	0.2164
12	908	0.6268	0.1793	0.2236	0.1793	0.2230	0.1714	0.2150
13	830	0.5730	0.1800	0.2256	0.1802	0.2241	0.1703	0.2150
14	753	0.5195	0.1879	0.2318	0.1840	0.2285	0.1726	0.2163
15	681	0.4697	0.1827	0.2287	0.1853	0.2305	0.1701	0.2136
16	603	0.4164	0.1849	0.2312	0.1862	0.2328	0.1725	0.2153
17	540	0.3727	0.1865	0.2335	0.1886	0.2358	0.1694	0.2138
18	469	0.3236	0.1879	0.2365	0.1889	0.2372	0.1711	0.2146
19	399	0.2755	0.1894	0.2378	0.1897	0.2373	0.1705	0.2152
20	334	0.2302	0.1931	0.2422	0.1919	0.2396	0.1720	0.2157
21	280	0.1935	0.1925	0.2418	0.1913	0.2399	0.1709	0.2146
22	224	0.1546	0.1915	0.2433	0.1919	0.2413	0.1722	0.2176
23	180	0.1239	0.1993	0.2490	0.1982	0.2458	0.1716	0.2174
24	145	0.0999	0.2065	0.2567	0.2025	0.2517	0.1714	0.2163
25	110	0.0758	0.2123	0.2624	0.2094	0.2586	0.1714	0.2162
26	81	0.0562	0.2215	0.2690	0.2154	0.2636	0.1691	0.2140
27	65	0.0446	0.2329	0.2830	0.2273	0.2780	0.1699	0.2131
28	41	0.0284	0.2583	0.3105	0.2453	0.2981	0.1690	0.2129
29	31	0.0214	0.2795	0.3369	0.2671	0.3228	0.1731	0.2185
30	20	0.0140	0.3078	0.3659	0.3000	0.3590	0.1697	0.2142

CR is the compression rate.

Table 7.14 The error rate of MODAL-CBR after case-base editing (K=5)

MODAL-CBR (K=5)								
Threshold	No. of cases	CR	Strategy 1		Strategy 2		Strategy 3	
			MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	1404	0.9690	0.1535	0.1964	0.1518	0.1946	0.1533	0.1961
2	1394	0.9623	0.1541	0.1975	0.1522	0.1952	0.1534	0.1965
3	1379	0.9519	0.1548	0.1982	0.1528	0.1957	0.1545	0.1975
4	1341	0.9258	0.1559	0.1989	0.1535	0.1967	0.1538	0.1964
5	1283	0.8855	0.1573	0.2010	0.1560	0.2000	0.1537	0.1968
6	1247	0.8608	0.1575	0.2007	0.1575	0.2006	0.1526	0.1957
7	1214	0.8377	0.1598	0.2032	0.1592	0.2023	0.1522	0.1951
8	1165	0.8039	0.1671	0.2103	0.1610	0.2043	0.1585	0.2017
9	1103	0.7616	0.1714	0.2160	0.1625	0.2061	0.1609	0.2061
10	1034	0.7134	0.1736	0.2186	0.1650	0.2098	0.1631	0.2060
11	953	0.6575	0.1751	0.2205	0.1674	0.2130	0.1625	0.2075
12	878	0.6060	0.1739	0.2200	0.1709	0.2164	0.1560	0.2014
13	809	0.5582	0.1753	0.2222	0.1730	0.2183	0.1547	0.1984
14	729	0.5033	0.1789	0.2268	0.1736	0.2196	0.1572	0.2022
15	659	0.4547	0.1814	0.2272	0.1751	0.2206	0.1599	0.2047
16	595	0.4107	0.1852	0.2346	0.1812	0.2295	0.1581	0.2040
17	543	0.3750	0.1877	0.2356	0.1834	0.2325	0.1600	0.2060
18	475	0.3276	0.1932	0.2432	0.1874	0.2364	0.1651	0.2122
19	395	0.2724	0.1925	0.2438	0.1883	0.2395	0.1652	0.2105
20	334	0.2306	0.1928	0.2460	0.1892	0.2412	0.1574	0.2047
21	272	0.1878	0.1982	0.2542	0.1915	0.2463	0.1583	0.2038
22	226	0.1558	0.2008	0.2574	0.1951	0.2479	0.1604	0.2075
23	190	0.1311	0.1995	0.2561	0.2011	0.2553	0.1599	0.2057
24	161	0.1113	0.2090	0.2712	0.2078	0.2656	0.1634	0.2109
25	124	0.0859	0.2164	0.2820	0.2063	0.2630	0.1646	0.2099
26	101	0.0700	0.2321	0.3002	0.2188	0.2808	0.1681	0.2142
27	80	0.0551	0.2567	0.3299	0.2332	0.2954	0.1679	0.2144
28	60	0.0411	0.2653	0.3561	0.2450	0.3232	0.1658	0.2105
29	45	0.0310	0.2978	0.3844	0.2707	0.3427	0.1705	0.2168
30	33	0.0228	0.3617	0.4435	0.3228	0.3870	0.1763	0.2234

CR is the compression rate.

Figure 7.5 to Figure 7.23 provide the comparative results of MAPE and RMSE before and after using the proposed CBM approach. There are three options in parameter setting for the CBR ECCE model: new weight and new training dataset (Strategy 1); original attribute weight and optimized training dataset (Strategy 2); new weight and original training case-base (Strategy 3). Among the three strategies, both Strategy 1 and Strategy 2 involve the new case training dataset. Strategy 3 shows the

performance of the new attribute weight after the case-base editing. The results can be seen in Appendix C-1 to Appendix C-9.

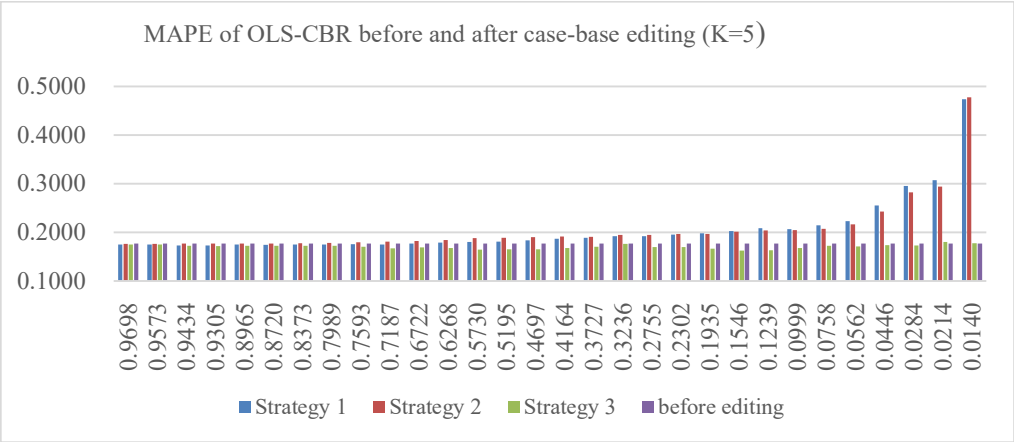


Figure 7.5 The changes in MAPE of OLS-CBR (K=5)

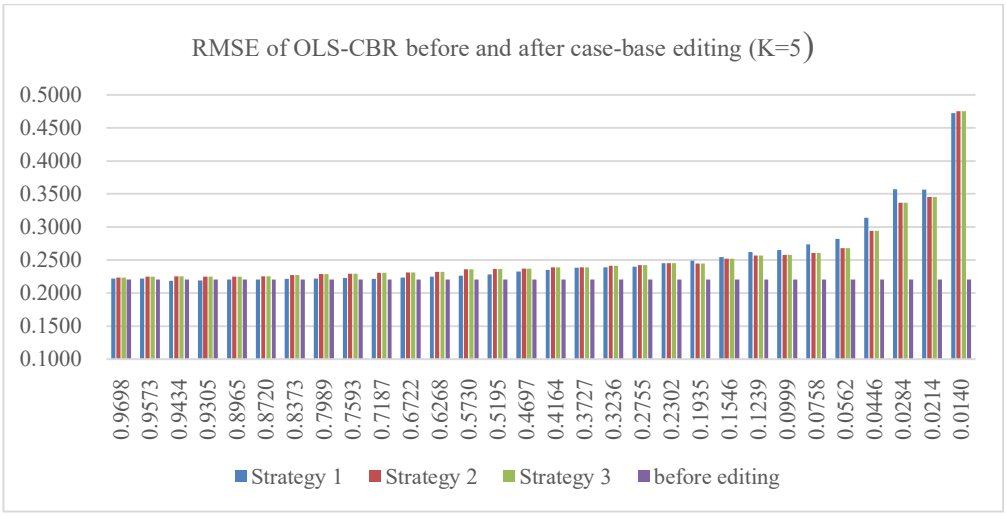


Figure 7.6 The changes in RMSE of OLS-CBR (K=5)

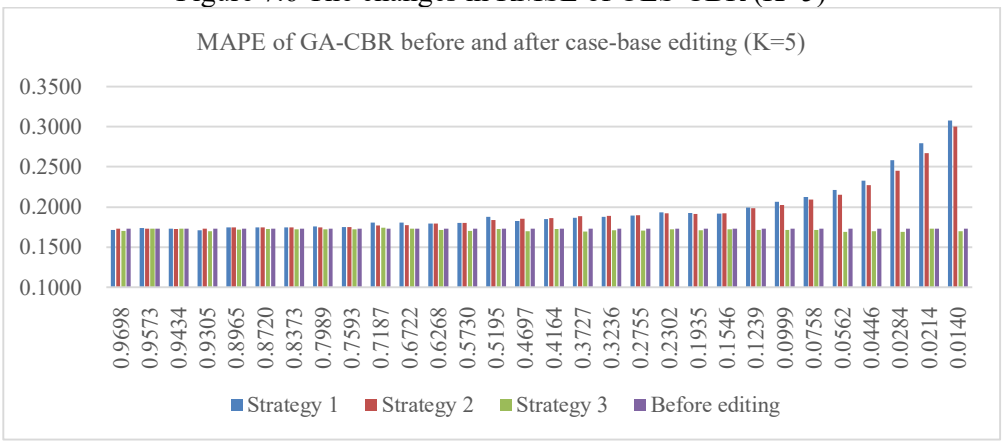


Figure 7.7 The changes in MAPE of GA-CBR (K=5)

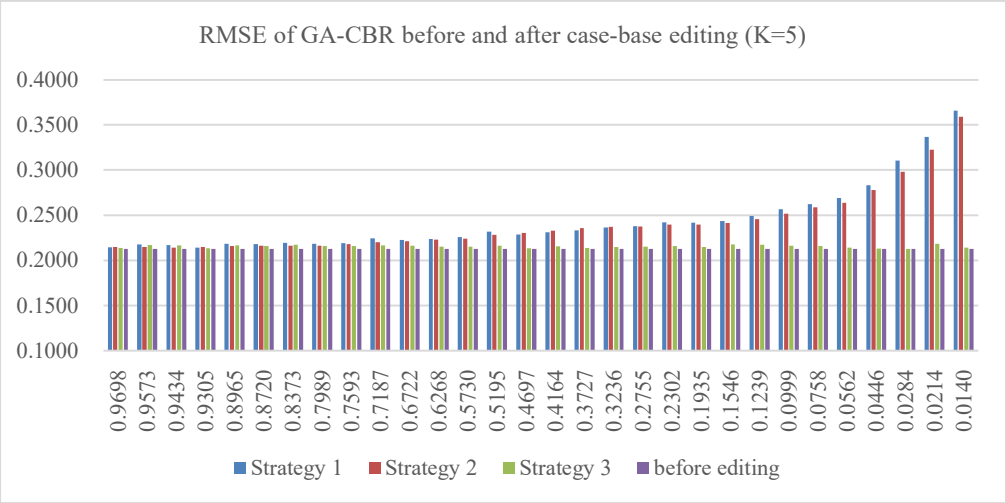


Figure 7.8 The changes in RMSE of GA-CBR (K=5)

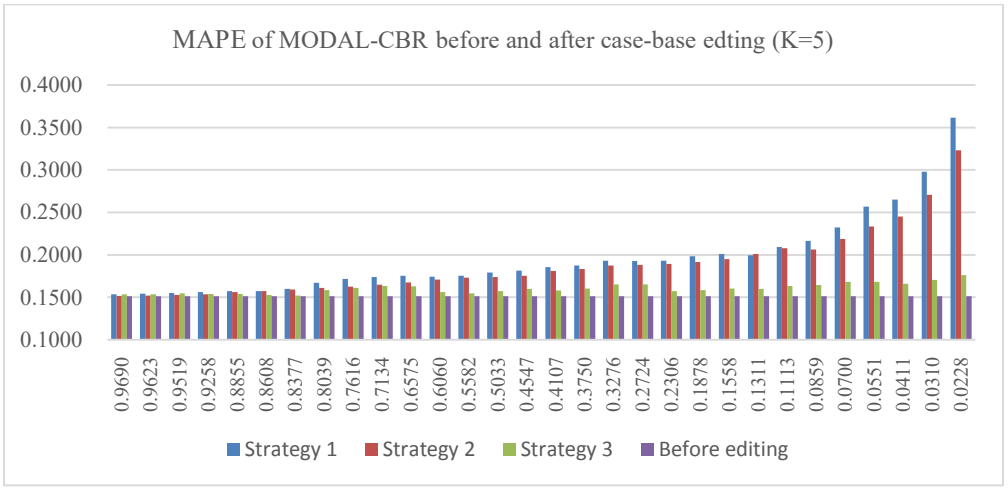


Figure 7.9 The changes in MAPE of MODAL-CBR (K=5)

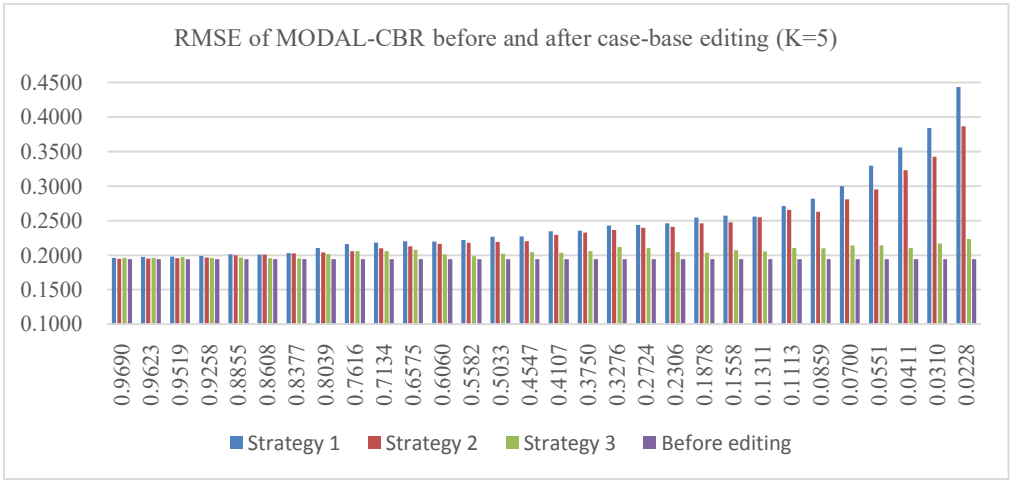


Figure 7.10 The changes in RMSE of MODAL-CBR (K=5)

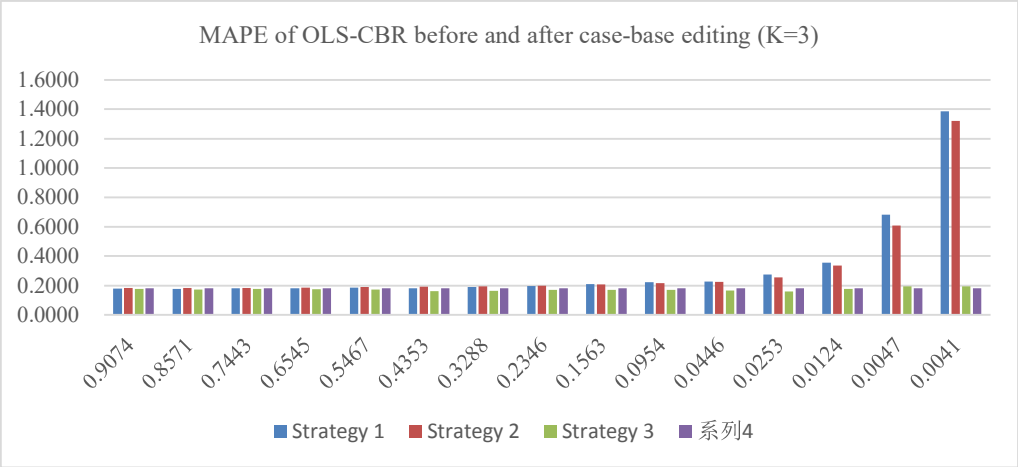


Figure 7.11 The changes in MAPE of OLS-CBR (K=3)

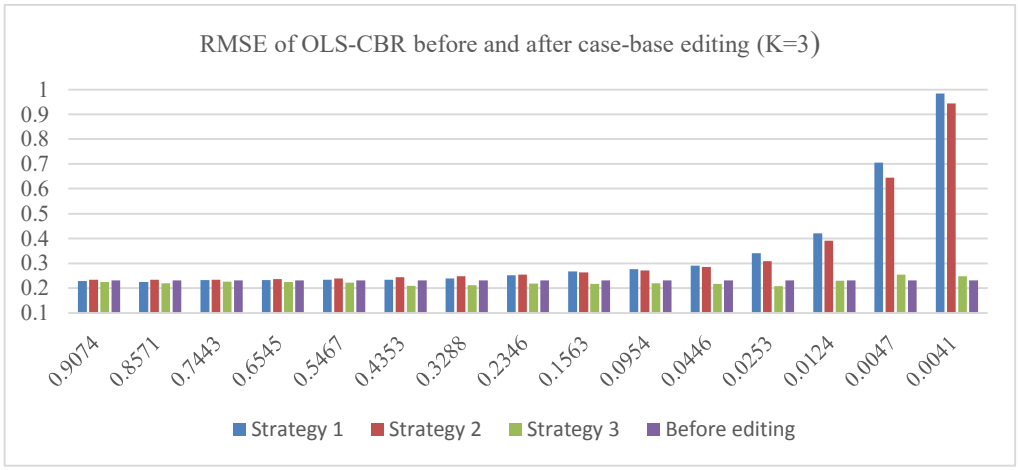


Figure 7.12 The changes in RMSE of OLS-CBR (K=3)

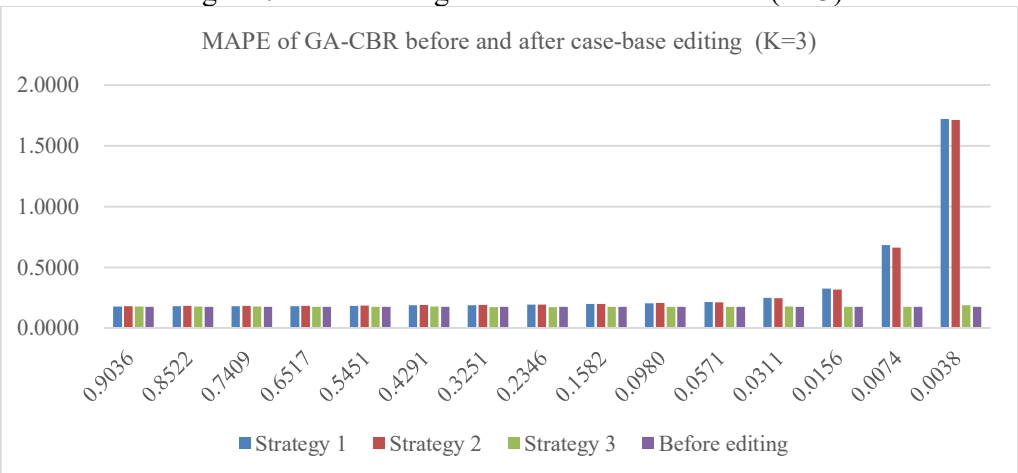


Figure 7.13 The changes in MAPE of GA-CBR (K=3)

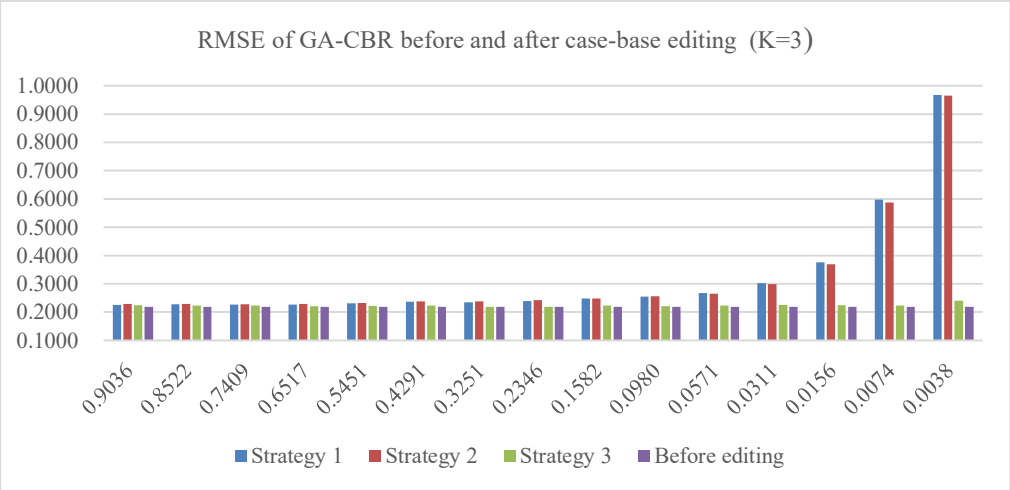


Figure 7.14 The changes in RMSE of GA-CBR (K=3)

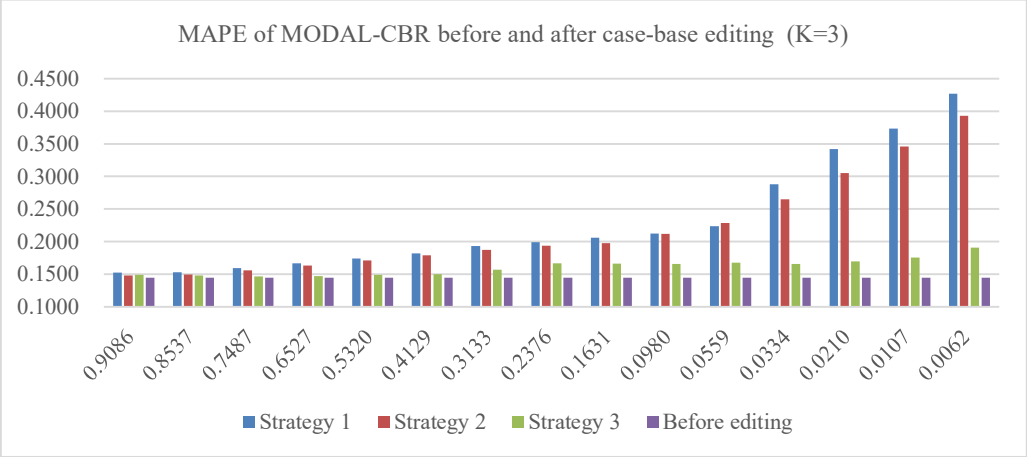


Figure 7.15 The changes in MAPE of MODAL-CBR (K=3)

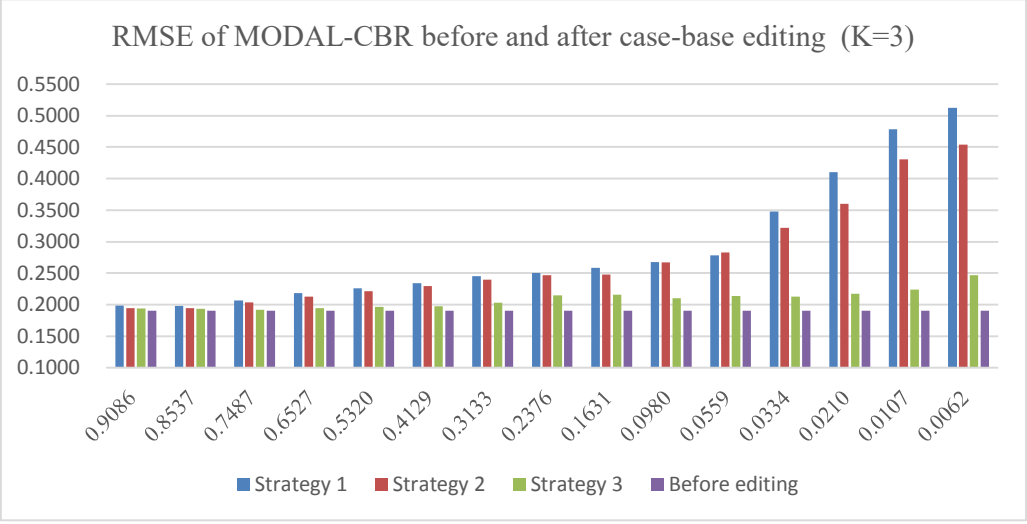


Figure 7.16 The changes in RMSE of MODAL-CBR (K=3)

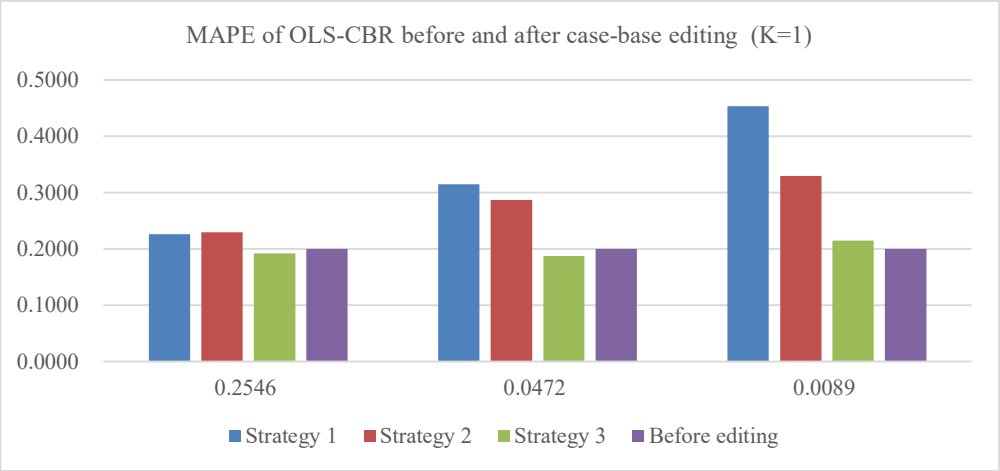


Figure 7.17 The changes in MAPE of OLS-CBR (K=1)

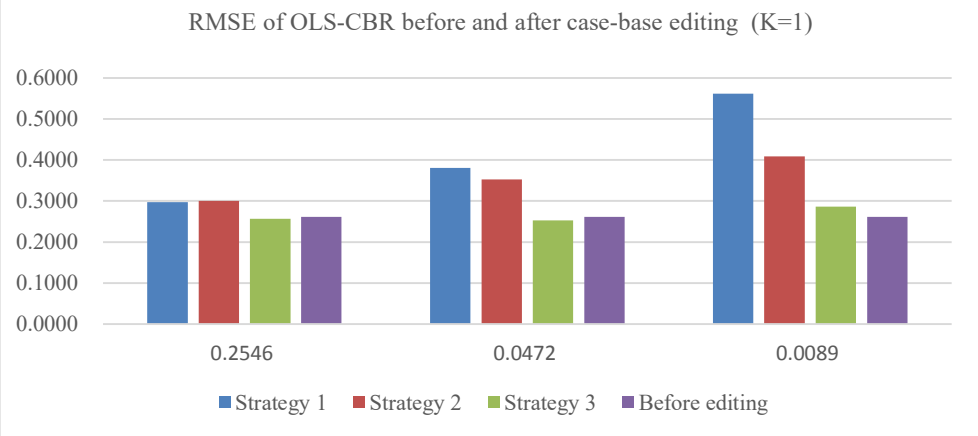


Figure 7.18 The changes in RMSE of LOS-CBR (K=1)

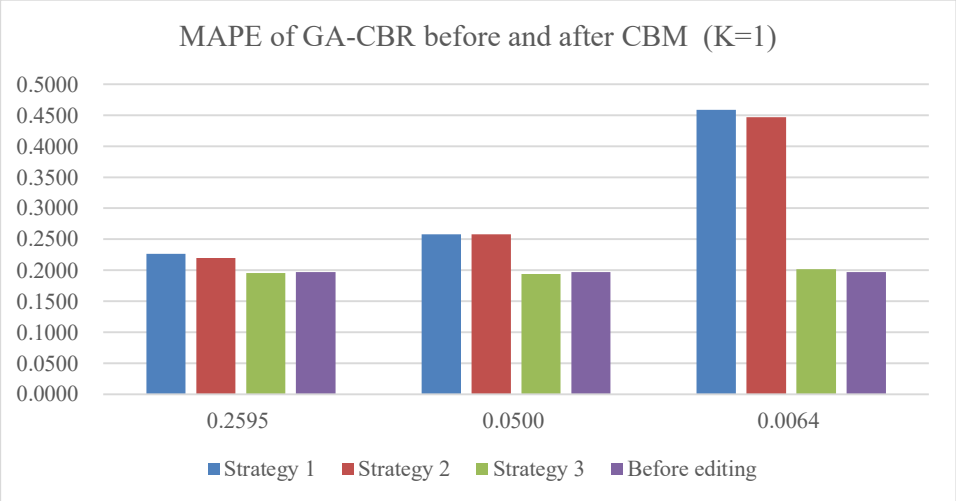


Figure 7.19 The changes in MAPE of GA-CBR (K=1)

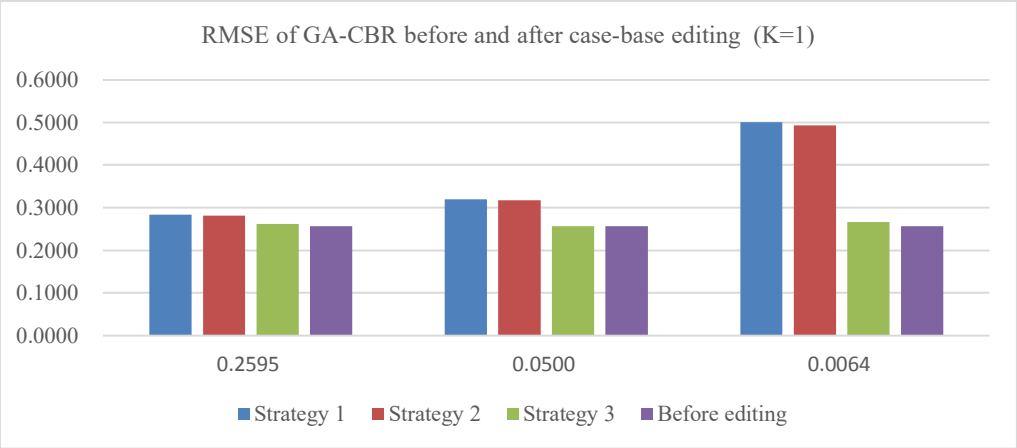


Figure 7.20 The changes in RMSE of GA-CBR (K=1)

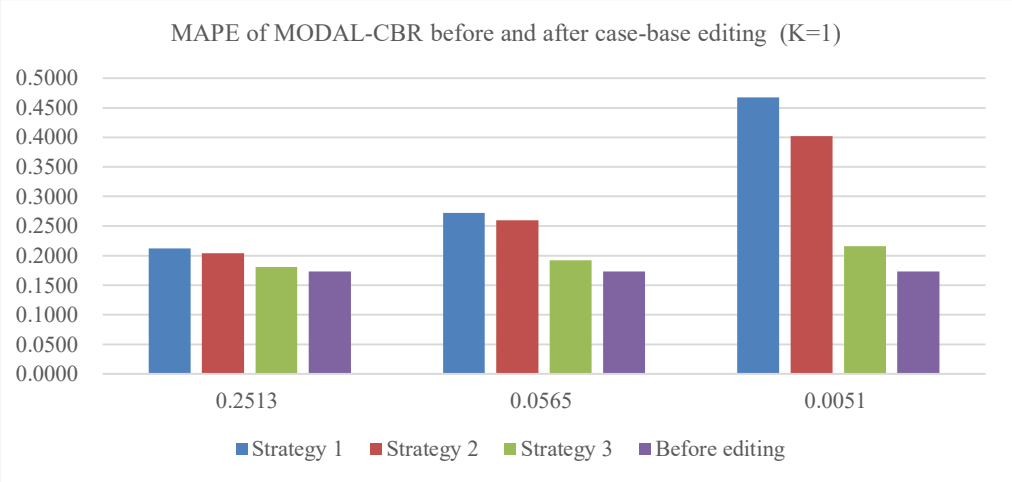


Figure 7.21 The changes in MAPE of MODAL-CBR (K=1)

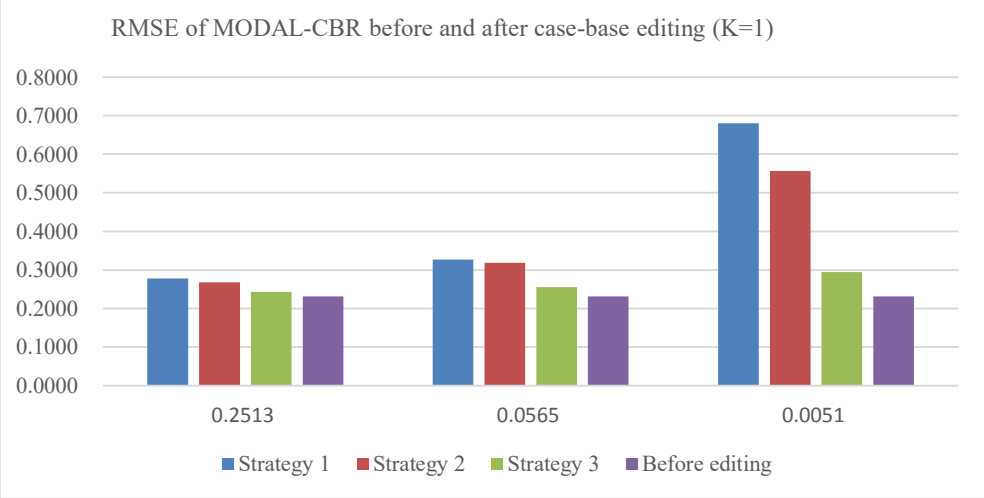


Figure 7.22 The changes in RMSE of MODAL-CBR (K=1)

Figure 7.5 and Figure 7.6 illustrate how the MAPE and RMSE of the OLS-CBR (K=5) model change according to the editing threshold. Before case-base editing, the MAPE and RMSE of OLS-CBR (K=5) are 17.67% and 22.06%. In Strategy 1, the MAPE slightly decreases when the threshold is low; when the compression rate reaches 0.7188, the MAPE of OLS-CBR (K=5) decreases to the lowest point. The decrease in the compression rate after 0.7188 produces the slight increase in the MAPLE of OLS-CBR (K=5). In Strategy 2, both MAPE and RMSE slightly increase with the increasing of the threshold; the greater compression rate, the larger MAPE and RMSE. In Strategy 3, the MAPE ranges from 16.21% to 17.75% and the best MAPE is 16.21% when the compression rate reaches 0.1546, illustrating the possible optimization effects of Strategy 3 on attribute weighting in OLS-CBR model. The detailed results can be seen in Appendix C-1.

Figure 7.7 and Figure 7.8 illustrate how the MAPE and RMSE of the GA-CBR (K=5) model change according to the editing threshold. Before case-base editing, the MAPE and RMSE of GA-CBR (K=5) are 17.41% and 21.40%. In Strategy 1, the MAPE slightly decreases when compression starts; when the compression rate reaches 0.9305, the MAPE of GA-CBR (K=5) decreases to the lowest point. After reaching this lowest point, MAPE of GA-CBR (K=5) begins to grow slightly with the continuous increase in threshold. The situation remains the same in Strategy 2. In Strategy 2, the MAPE slightly decreases when compression starts; when the compression rate reaches 0.9305, the MAPE of GA-CBR (K=5) decreases to the lowest point and begins to grow slightly afterwards. In Strategy 3, the MAPE ranges from 16.90% to 17.39%, illustrating the possible optimization effects of Strategy 3 on the GA-CBR model. The detailed results can be seen in Appendix C-2.

Figure 7.9 and Figure 7.10 illustrate how the MAPE and RMSE of the MODAL-CBR (K=5) model change according to the editing threshold. Before case-base editing, the MAPE and RMSE of MODAL-CBR (K=5) are 15.13% and 19.40%. In Strategy 1 and 2, both MAPE and RMSE keep growing with the increasing of the compression rate; the greater compression rate, the large MAPE and RMSE of MODAL-CBR (K=5). For example, when the compression rate reaches 0.8855, the MAPE and RMSE of MODAL-CBR (K=5) are 15.73% and 20.10%; when the compression rate reaches to 0.5582, the MAPE and RMSE of MODAL-CBR (K=5) increase to 17.09% and

21.64%. In Strategy 3, the MAPE ranges from 15.33% to 17.63%. The detailed results can be seen in Appendix C-3.

Figure 7.11 and Figure 7.12 illustrate how the MAPE and RMSE of the OLS-CBR (K=3) model change according to the editing threshold. Before case-base editing, the MAPE and RMSE of OLS-CBR (K=3) are 18.16% and 23.06%. In Strategy 1, the MAPE slightly decreases when compression starts; when the compression rate reaches 0.6517, the MAPE of OLS-CBR (K=3) decreases to the lowest point (MAPE = 17.95%) and begins to grow slightly. In Strategy 2, both MAPE and RMSE keep increasing with the growth of compression starts; the greater compression rate, the large MAPE and RMSE. In Strategy 3, the MAPE increases from 17.22% and reaches 73.94% when the compression rate becomes too small. The detailed results can be seen in Appendix C-4.

Figure 7.13 and Figure 7.14 illustrate how the MAPE and RMSE of the GA-CBR (K=3) model change according the editing threshold. Before case-base editing, the MAPE and RMSE of GA-CBR (K=3) are 17.65% and 22.19%. In Strategy 1 and 2, both MAPE and RMSE slightly increase when compression starts; the greater the degree of compression, the large MAPE and RMSE. In Strategy 3, the MAPE ranges from 17.22% to 18.68% and the MAPE remains the same level before case-base editing, except when the compression rate becomes too small. The detailed results can be seen in Appendix C-5.

Figure 7.15 and Figure 7.16 illustrate how the MAPE and RMSE of MODAL-CBR (K=3) model change according to the editing threshold. Before case-base editing, the MAPE and RMSE of MODAL-CBR (K=3) are 14.49% and 19.06%. In Strategy 1 and 2, both MAPE and RMSE of MODAL-CBR (K=3) slightly increase when compression starts; when the compression rate reaches 0.5320, the increase in MAPE and RMSE start growing fast. In Strategy 3, the MAPE increases from 14.89% to 19.05% when the editing threshold decrease from 1 to 15. The detailed results can be seen in Appendix C-6.

Figure 7.17 and Figure 7.18 illustrate how the MAPE and RMSE of OLS-CBR (K=1) model change according to the editing threshold. Before case-base editing, the MAPE and RMSE of OLS-CBR (K=1) are 19.90% and 26.44%. In Strategy 1 and 2, both MAPE and RMSE of OLS-CBR (K=1) keep increasing with the increase in

compression rate; Strategy 3 fails to reduce the MAPE and RMSE of OLS-CBR (K=1) effectively. The detailed results can be seen in Appendix C-7.

Figure 7.19 and Figure 7.20 illustrate how the MAPE and RMSE of the GA-CBR (K=1) model change according to the editing threshold. Before case-base editing, the MAPE and RMSE of GA-CBR (K=1) are 19.38% and 25.54%. In Strategy 1 and 2, both MAPE and RMSE of GA-CBR (K=1) increase with the increase in compression rate; Strategy 3 fails to reduce the MAPE and RMSE of GA-CBR (K=1). The detailed results can be seen in Appendix C-8.

Figure 7.21 and Figure 7.23 illustrate how the MAPE and RMSE of the MODAL-CBR (K=1) model change according to the editing threshold. Before case-base editing, the MAPE and RMSE of MODAL-CBR (K=1) are 17.37% and 23.11%. In all strategies, both MAPE and RMSE of MODAL-CBR (K=1) increase with the increase in the compression rate. The detailed results can be seen in Appendix C-8.

7.3 DISCUSSION

The results after case-base editing differ due to the different setting. The weight determination method has a significant effect on the performance of the case-base editing. For example, Strategy 1 works better in OLS-CBR (K=5), while Strategy 2 produces better results in MODAL-CBR (K=5). Generally, the significance of weight1 tends to decrease with the increase of the compression rate. The parameter setting in the CBR model has a significant influence on its performance. Except for weight determination, the value of case adaptation becomes important when editing case-base. From Table 7.1 and Table 7.2, it can be seen that the compression speed is directly linked with the value of case adaptation. When case adaptation value is set as 5, the compression process is slower and the compression range is bigger compared when case adaptation is set to 1. Generally, the bigger value of case adaptation, the slower speed of the compression process and the bigger the compression range.

The three strategies in this study illustrate the effect of case-base editing from different perspectives. Strategy 1 uses the new weight and new case-base after case-base editing; Strategy 2 uses the original weight before case-base editing and new case-base after case-base editing; Strategy 3 uses the new weight after case-base editing and the original case-base before case-base editing. Only Strategies 1 and 2 have a compression effect on the CBR model since they use the case-base after the editing.

The major difference between Strategy 1 and 2 is the attribute weighting value. Strategy 1 and 2 are designed to compress the size of the case-base while Strategy 3 is designed to test the potential optimization on attribute weight by case-base editing.

The performances of Strategy 1, 2 and 3 in ECCE CBR are compared before and after using the case-base editing. Generally, compressing the size of the case-base inevitably sacrifices the accuracy. Thus Strategy 1 and 2 are designed to minimize this trend when compressing the size of case-base. It's interesting to find that there is a slight decrease in the MAPE of OLS-CBR (K=5) when using the strategy 1 and threshold is set low. When the threshold is set to 10, the compression rate of OLS-CBR (K=5) is 0.7187; the MAPE and RMSE of OLS-CBR (K=5) are 17.47% and 22.13% separately. Compared with the MAPE (17.41%) and RMSE (21.40%) generated from OLS-CBR (K=5) model before case-base editing, the results show that the Strategy 1 can maintain the CBR's performance while reducing the size of case-base by 28.13% when using Strategy 1 in OLS-CBR (K=5). Compared with the Strategy 1, OLS-CBR (K=5) has slightly higher MAPE and RMSE when using Strategy 2. The potential reason behind this, is that strategy 1 helps to eliminate the noisy data and produce more accurate attribute weight when the editing threshold is set low. The result that MAPE and RMSE slightly decrease when the threshold is set low using Strategy 3, also confirms the potential effect on the attribute weight optimization. For the GA-CBR model, there is no significance between the results in Strategy 1 and Strategy 2; With the increase of the compression rate, both MAPE and RMSE keep growing. The MAPE and RMSE of GA-CBR (K=5) can be slightly decreased to 16.90% and 21.29% when using the Strategy 3. For the MODAL-CBR model, Strategy 2 produces better results than Strategy 1, and Strategy 3 has no optimization effect.

The results help us understand how MAPE and RMSE change with decreases in the number of cases stored in the case-base. By slightly compressing the OLS-CBR model, the results show that MAPE and RMSE can be even slightly improved. The underlying reason behind that may be the potential effect of the proposed method on purifying the training dataset and improving the performance of CBR for ECCE. Except for slightly increasing prediction performance in the OLS-CBR (K=5) model, the proposed method has the advantage of needing less storage requirements.

7.4 CHAPTER SUMMARY

This chapter introduces an original method based on case-base editing to improve the efficiency of the ECCE CBR model. The case-based editing methods are proposed based on the basic concepts of CBM and the weight coverage contribution introduced in Chapter 3. Given one training case, the editing rules consider either to delete or keep the case unchanged according to the evaluation criterion. This rule is used in all cases in the case-base, and the weighted coverage contribution is calculated. The case with the lowest coverage contribution will be eliminated. The performance of the ECCE CBR model was compared before and after using the proposed CBM approach. The editing threshold differs according to the value of the K. To better illustrate how this method works, the editing threshold ranges from 1 to 30 when K equals 5; ranges from 1 to 15 when K equals 3; and ranges from 1 to 3 when K equals 1. After editing the case-base, the attribute weight and the case-base both have been updated. Thus, there are three options in parameter setting for the CBR ECCE model: new weight and new training dataset (strategy 1); new weight and original training case-base (strategy 2); original attribute weight and optimized training dataset (strategy 3). Leave-one cross-validation is used for model validation. Two error measures, namely Mean Average Percent Error (MAPE) and the Root Mean Squared Error (RMSE) of the log values were used to evaluate the performance of the model.

The results after case-base editing differ due to the different settings. The weight determination method has a significant result on the performance of the case-base editing. For example, Strategy 1 works better in OLS-CBR (K=5), while Strategy 2 produces better results in MODAL-CBR (K=5). The potential reason behind this is the changes caused by case-base editing.

Generally, compressing the size of case-base inevitably results in the loss of accuracy. Strategies 1 and 2 are designed to minimize this trend when compressing the size of case-base. It's interesting to find that there is a slight decrease in the MAPE of OLS-CBR (K=5) when the threshold is low using the strategy 1. When the threshold is set to 10, the compression rate of OLS-CBR (K=5) is 0.7187; the MAPE and RMSE of OLS-CBR (K=5) are 0.1823 and 0.2312 separately. The results show that the Strategy 1 can maintain the CBR's performance while reducing the size of case-base by 28.13% when using Strategy 1 in OLS-CBR (K=5). Compared with the Strategy 1, OLS-CBR (K=5) has slightly higher MAPE and RMSE when using Strategy 2. The

potential reason behind this is that strategy 1 helps to eliminate the noisy data and produce more accurate attribute weights when the editing threshold is set to be low. The result that MAPE and RMSE slightly decrease when using Strategy 3, also confirms the potential effect on the attribute optimization. For the GA-CBR model, there is no significance between the results in Strategy 1 and Strategy 2; With the increase of the compression rate, both MAPE and RMSE keep growing. The MAPE and RMSE of GA-CBR (k=5) can be slightly decreased to 16.90% and 21.29% when using Strategy 3. For the MODAL-CBR model, Strategy 2 produces better results than Strategy 1, and Strategy 3 has no optimization effect.

Chapter 8: Conclusion

8.1 OVERVIEW OF RESEARCH OBJECTIVES

The main aim of this research was to improve the CBR model in ECCE for long-term use. The research objectives were developed to achieve the proposed research aim. Four research objectives were built upon four research questions, as follows:

Research Question 1: *What limitations exist in the current ECCE CBR studies with respect to long-term use?*

Research Objective 1: To provide a comprehensive literature on the previous studies on ECCE CBR mode.

Research Question 2: *What are the main differences caused by different parameter settings of ECCE CBR model?*

Research Objective 2: To conduct an empirical study to compare the methods for calculating weight and similarity, as well as exploring the influence of sample size on ECCE CBR.

Research Question 3: *How to maintain a stable knowledge structure of the CBR model during long-term use?*

Research Objective 3: To improve the robustness of the ECCE CBR model by combining the CBR system and robust method.

Research Question 4: *How to improve the efficiency of the ECCE CBR model for long-term use?*

Research Objective 4: To develop a method to enhance the efficiency of the ECCE CBR model and maintain its performance during long-term use.

8.2 CONCLUSIONS OF THE RESEARCH OBJECTIVES

8.2.1 Research Objective 1

RO1: To provide a comprehensive literature on the previous studies on ECCE CBR model.

The first research objective answered the question of, *" what limitations exist in the current ECCE CBR studies with respect to long-term use?"* The research question was answered by providing a comprehensive literature review in related areas. The background of the construction cost estimation, inaccuracy in construction cost estimation, and influence factors in construction cost performance were reviewed together with the significance and challenges in ECCE. The application of CBR in ECCE carefully examined each step of the existing CBR model in ECCE. After briefly introducing the CBR and its advantage in ECCE, problem formulation, case retrieval, case reuse, case revision, and CBM were reviewed to provide an in-depth understanding of the existing research. Section CBM included the definition of CBM, the criteria for evaluating case-base, influencing factors in CBM, and classification of CBM strategy, case-base reduction strategy, case-base partitioning strategy, and case-base optimization strategy.

After literature review, numerous limitations were found in the existing research. Firstly, despite some research suggesting the potential superiority of the CBR model for long-term use, there was no empirical evidence supporting this assumption. Secondly, the existing research lacked a systematic understanding of the parameter setting in the CBR model. Various weight determination methods and similarity functions were used in the previous CBR model, yet the question of which methods are better for calculating weight and similarity remained. A comparative study could help to identify the optimal parameters in the CBR model and provide a valuable understanding of ECCE CBR.

Thirdly, the research attention on improving the CBR's performance for long-term use was found currently to be far from enough. The majority of ECCE CBR modes focus on case retrieval or case reuse, largely ignoring the CBM. Based on the research findings in this chapter, the following research objectives were proposed.

8.2.2 Research Objective 2

RO2: To conduct an empirical study to compare the methods for calculating weight and similarity, as well as exploring the influence of sample size on ECCE CBR.

The second objective was set up to respond to the question of, "*What are the main differences caused by different parameter settings of ECCE CBR model?"* Based on this research question, this study conducted a comparative study of the different

model settings of CBR based on a collected ECCE database of 1450 completed Chinese apartment building projects. Given the sample size range in previous studies and the size of the collected database, seven sample sizes of 50, 100, 200, 400, 600, 800, and 1000 contracts were used in this study, covering the range of sample sizes used in previous studies. Three weight determination methods (MRA, GA, FC) and two similarity functions were compared by random selections from the database, with a 20% project hold-out sample. Two errors, including MAPE and RMSE, were used to measure the performance of the ECCE CBR model.

The results provided a better understanding of the settings in the CBR models. GA had more advantages when dealing with smaller sample sizes. After the size of the case-base becomes large (>400), there was no significant difference between models using MRA and GA. FC had the weakest performance in terms of accuracy, especially when combining S1 function. Besides, on average, the S2 function had a better performance than S1. Also, adjusting case adaption value could improve the performance of the CBR model. For different sample sizes, the best case adaptation value might be different. Therefore, multiple case adaptation values should be tried in the CBR study to have the best performance.

Some discussions were made based on the justification of sample size and explanation of data-oriented results. Despite the differences in the parameter settings, the MAPE of the model based on 1000 cases largely decreased compared with that of 50 cases. For example, in the CBR-W1S1 model, the MAPE was 35.81% when the sample size was 50 and decreased to 25.31% when the sample size became 1000. The results in the study confirmed the previous hypothesis that the performance of CBR improved with the increase of the sample size generally (Ji et al., 2010b). This research also found that the marginal contribution of the increasing size of case-base decreases. For example, the accuracy largely increased when the sample size was raised from 50 to 400. However, the increase in the accuracy became unapparent when the sample size increased from 400 to 1000.

What is also becoming apparent is that expanding the database to contain more cases containing a small number of highly influential predictor variables (six in this case) may be a better use of resources in improving the accuracy of the ECCE when the size of case-base is extremely small. Therefore, even though ECCE lacks detailed target project information at the early stage, increasing the sample size may be

sufficient to produce reliable results by objective methods. The long-term use of the ECCE CBR system results in continuous growth in the size of the case-base. Although this improves the performance of the accuracy of the ECCE CBR system, the improvement in accuracy is quite limited when the size of the case-base becomes large. Therefore, case-base maintenance during the long-term use of the ECCE CBR system was proposed to improve the ECCE CBR model.

8.2.3 Research Objective 3

RO3: To improve the robustness of the ECCE CBR model by combining the ECCE CBR system and robust method.

The third objective was initiated to justify the question of, "*How to maintain a stable knowledge structure of the CBR model during long-term use ?*" Based on this research question, a robust method was introduced in the ECCE CBR system for improving the long-term use of the ECCE CBR model. The MODLR was used in weight determination to increase the model's robustness. K-folder cross-validation was used for the model validation. Different K values of 10, 20, 40, 80, 160, 320 and 1450 were used, given their influence on the training sample.

The results showed that the total above-floor area was the most significant attribute in all weight determination methods, followed by the total underground-floor area. The significance of remaining attributes differed due to the methods used. The third-most significant attribute differed due to the weight determination methods. Except for MODAL, duration and the total height of the building were the least significant attribute weights for ECCE.

The model was evaluated from the perspectives of robustness and accuracy. The variance of weight attribute of each calculation was used for measuring the weight robustness. The measure of the estimation accuracy included the MAPE and RMSE. The results were compared with those using the ordinary least squares (OLS) regression and GA in weight determination, followed by the discussion on the overall robustness of the weight determination. From the robustness perspective, the MODAL-CBR model had the least attribute variance compared with those with OLS and GA. The results showed the superiority of the MODAL-CBR model in weight stability. From the accuracy perspective, the results showed that the MODAL-CBR model has the least mean MAPE and mean RMSE, followed by the GA-CBR model

and OLS-CBR model. In summary, MODAL-CBR produced stable attribute weights during the multiple calculations of the CBR model in long-term use. It better prepares construction cost agencies and organisations to tackle the massive growth in the scale of the data and help practitioners have a consistent understanding of the knowledge stored in the case-base.

The results showed that the knowledge structure inevitably changes during long-term use of the CBR model. Therefore, the data quality must be considered in practice. Data quality implies that one needs to look beyond traditional concerns with the accuracy of the data (Tayi & Ballou, 1998). There are four dimensions in data quality: accuracy, completeness, consistency, and timeliness (Ballou & Pazer, 1985). 'Accuracy' is defined as the correctness of the fact recoded 'completeness' as the relevence of the information recorded, 'consistency' as the uniformity in the information recored, and 'timeliness' as the recording of information on time. Poor data quality would result in negative impact on operation of the system (Coetzer & Vlok, 2019). Furthermore, poor data quality cannot fullfil the requirement of the system, consequently resulting in fail in achieving the expected results (Karkouch, et al., 2016; Laranjeiro, et al., 2015; Taleb, et al., 2016).

Construction cost estimation usually needs the cost data of the historical projects. The historical cost data will be useful for cost estimation only if they are collected and organized in a way that is compatible with future applications. The consistency of the construction cost data is critical since it provides a reliable baseline for the new project. Therefore, the information must be updated with respect to changes that will inevitably occur (Hendrickson, et al., 2008). Without sufficient refining and updating, historical cost data shouldn't be used carelessly. Changes in relative prices may have substantial impacts on construction costs that have increased relatively in price.

Unfortunately, systematic changes over a long period of time for such factors are inevitable. In particular, the changed resource costs, construction methods, design styles, and economic conditions create the outdated and inconsistent data. Also, the size of the case-base can grow quickly with the long-term use of a CBR model (Smiti & Elouedi, 2018a). The efficiency of solving a new problem thus becomes increasingly slow, resulting in compromised overall performance of the CBR model (Khan, et al., 2019b; Lupiani, et al., 2014b). Without proper handling, these problems caused by long-term use will impair the performance of the CBR model, the typical issues being

the unstable knowledge structure and low efficiency because of the continuously increasing size of the case-base during long-term use. This inevitably raises the problem between the benefits of having more data and the deficiencies of having inappropriate data.

8.2.4 Research Objective 4

RO4: To develop a CBM strategy for ECCE CBR models to maintain its efficiency during long-term use.

The fourth objective was proposed to answer the question of, "*How to improve the efficiency of the ECCE CBR model for long-term use?* " This study introduces an original method based on case-base editing to improve the efficiency of the ECCE CBR model. The case-based editing method is proposed based on the basic concepts of CBM and the weight coverage contribution introduced in Chapter 3.

The performance of the ECCE CBR model is compared before and after using the proposed CBM approach. The editing threshold differs according to the value of the K. The editing threshold ranges from 1 to 30 when K equals 5, ranges from 1 to 15 when K equals 3, and the editing threshold ranges from 1 to 3 when K equals 1. After editing the case-base, the attribute weight and the case-base both have been updated.

The results show that the weight determination method has a significant result on the performance of the case-base editing. For instance, Strategy 1 works better in OLS-CBR (K=5), while Strategy 2 produces better results in MODAL-CBR (K=5). Generally, compressing the size of case-base inevitably results in the loss of accuracy. The Strategies 1 and 2 are designed to minimize this trend when compressing the size of case-base. It's interesting to find that there is a slight decrease in the MAPE of OLS-CBR (K=5) when using Strategy 1 to compress the case-base lightly. The results show that the Strategy 1 can maintain the CBR's performance while reducing the size of case-base by 28.13% when using Strategy 1 in OLS-CBR (K=5). Compared with the Strategy 1, OLS-CBR (K=5) has slightly higher MAPE and RMSE when using Strategy 2. The potential reason behind this is that Strategy 1 helps to eliminate the noisy data and produce more accurate attribute weight when the editing threshold is set low. The result that MAPE and RMSE slightly decrease when using Strategy 3 also confirms the potential effect of the proposed CBM on the attribute optimization. For the GA-CBR model, there is no significance between the results in Strategy 1 and

Strategy 2; with the increase of the compression rate, both MAPE and RMSE keep growing. The MAPE and RMSE of GA-CBR (k=5) can be slightly decreased to 16.90% and 21.29% when using Strategy 3. For the MODAL-CBR model, Strategy 2 produces better results than Strategy 1; Strategy 3 has no optimization effect. Altogether, this study could serve as a reliable reference for optimizing the overall performance of the CBR ECCE model.

8.3 RESEARCH CONTRIBUTION

This study aims to improve the practice of long-term use of case-based reasoning in early construction cost estimation through parameter setting, by improving the robustness of the CBR system and enhancing the efficiency of the CBR system. This study makes theoretical contributions to the body of knowledge domain in the early construction cost estimation. It provides new insights in optimizing the ECCE CBR model during its long-term operation. This study fills the gap in the multi-dimensional optimization of ECCE CBR models by addressing its robustness and efficiency performance. It provides theoretical insights into the role of the sample size played in the ECCE CBR model, and how to maintain the CBR model after the ECCE model has been established. The value-added contributions toward theory development are discussed in detail as follows.

Firstly, this study conducted a comprehensive literature review on the existing ECCE CBR model. The literature review in this thesis helps reseachers and practitioners achieve a better understanding of the history, development, status quo, and further trend of ECCE CBR. By providing a careful examination of each step of CBR models in ECCE, the limitations in the previous studies are identified and are summarized. Furthermore, the knowledge gaps and the future research directions may serve as a motivation for reseachers and practitioners to work on the next generation of research to assist the development of ECCE CBR around the world.

Secondly, this study conducted an empirical study on sample size affecting the accuracy of the ECCE CBR model. By analysing the influence of sample size on the accuracy of the CBR model, this thesis helps understand the changes in accuracy with increase in sample size. This part of the research could provide a justified reason for the contradictory results of the previous studies. Also, comparison among different weight determination methods and similarity functions helps researchers and

practitioners understand which approach works better in the ECCE CBR model. The research findings in this study could reduce the data collection effort for construction cost estimation at the early stage. In the construction industry, data collection for small companies or agencies remains difficult. The results in this study will assist small companies to understand how much effort they need to make in data collection for estimating a project.

Thirdly, this study address the robustness of case-base structure during its long-term use. It considers the knowledge structure of the case-base in the ECCE CBR model from the long-term perspective. This study addresses the problem caused by noisy data or the situation where the actual cost data distribution does not satisfy the hypothesis. The weight stability of the case-retrieving process has been improved by using MODAL, which focuses on the mainstream bulk of cases. Combining the ECCE CBR model with robust weighting effectively reduces the changes in the attribute weight raised, by continuously updating knowledge in the case-base. It helps maintain a stable knowledge structure of the case-base, which can better handle the changes in the case-base data. This research enriches the existing studies in the literature by not only considering the single-time performance of the model, but also improving the stability of long-term use. It provides some new insights for of ECCE CBR optimization.

Finally, this study improves the application of the CBR system in ECCE by introducing a case-base maintenance strategy. This study proposed a case-base editing method based on a prototype selection to compress the size of the case-base. This study solves the efficiency issues raised by the constant increase in the size of the case-base. It assists ECCE CBR models in satisfying the efficiency and storage requirements during its long-term use. The size of the case-base can be maintained at a stable level despite continually adding and updating of the case-base. It can reduce the inconsistent data raised by the changed design styles and economic conditions in the case-base. The deficiencies of having inappropriate data can be minimized by retaining the most useful cases. It prepares the ECCE CBR model with an updating system for handling the redundancy in the case-base during its long-term use.

This study contributes to improving the overall performance of the ECCE CBR models by addressing the real-world problems raised during its long-term use. By considering the stability in weight determination and efficiency in case retrieval, this

study provides valuable insights in optimizing ECCE CBR from multidimensional perspectives. Except for the accuracy, the robustness of knowledge structure and the case-base maintenance are considered. The proposed ECCE CBR system produces more reliable feature weighting and estimation results during the long-term use. With the data growth in the construction industry and the popularity of the public and free database in times to come, the value of this research will become more apparent in the future.

8.4 RESEARCH LIMITATIONS

Despite the research significance and contribution, this study has several limitations as follows:

The scope of the study involves only the early stage of construction cost estimation of apartment buildings. The model is built exclusively for handing the predictors at the early stage. These may result in the potential limitation of this study by using only a few predictors. Given the amount of project information available in the early stage, this study considers the data availability and the effectiveness of the predictors in data collection. Therefore, the same study should be conducted with more predictors and different types of buildings. Considering the research scope and research aims, this study primarily relied on data-based analysis. The predictors in this study lacked further validation by practitioners. The data selection and categorisation is limited as it lacks careful examination. Nevertheless, the results in this study make contributions to improving the robustness and efficiency of the model. It can also be considered as a reliable source for any cost estimation CBR model at other project stages.

The literature review in this thesis is limited to journal papers because peer-reviewed journal articles are the most valuable sources of information. Other sources, such as the International Construction Measurement Standards (ICMS 1 & 2), which defines the cost categories and explores the different ways that costs can be collected and used, as well as work done by professional societies in Canada, USA, UK, Malaysia, and Sweden on how costs can be captured and manipulated using systems, have been excluded as outside the scope of this review. These construction cost estimation practices may be different from what is in the literature review of this research.

This study aims to improve the ECCE CBR model for long-term use. Despite that the research hypothesis proposed in this study provides the research direction, this thesis is limited because it can not be rigorously tested. Although the number of cases in this study has exceeded the number of samples in previous studies, the sample size may not be large enough to show the significance of this research. The data collection is limited due to the confidentiality of construction costs in the business industry. Despite collecting 1800 building projects in China, this is still far from enough for taking full advantage of the methods in data mining. Since the study attempts to explore the influence of the case-base on estimation accuracy, the insufficient sample size of case-base will make the results less convincing. This study is limited since it takes no account of the huge databases that organisations such as highways agencies, housing providers, multilateral development banks, and large multi-disciplined professional practices keep. With the data growth in the construction industry and the popularity of the public and free data set, the value of this research will become more apparent in the future.

Moreover, the majority of the variables used in this study are numerical. Therefore, this research may have limitations in ignoring nominal and categorical data. The generalisation of this study may also be limited because of the lack of sufficient diversity of attributes. This research only uses apartment buildings in China, and further research should validate the use of the proposed method with other building types for higher generalization. The reliability of the proposed methods needs to be tested using various sources of data.

Also, this research is limited in the reasoning process by only mimicking what is conventionally done by experienced estimators. This research lacks the comparison between the algorithm and the experienced estimator and the results in this study lack the industry context. This study exploits the classic ECCE CBR model as the foundation and compares the results based on three weight determinations, two similarity functions, and three case adaptation values. The performance of the CBR model based on the combination of other parameter settings remains a question. Future research will require comparison studies with more complicated parameter settings, as well as different methodologies.

8.5 RECOMMENDATION FOR FUTURE RESEARCH

This research contributes to providing various perspectives to study the ECCE, especially on improving the ECCE CBR model for long-term use in practice. Several recommendations are made for further research, as below.

More literature reviews can be conducted based on the extensive work that has been done in the construction sector by companies, organisations, and professional institutions on the techniques of early cost estimating. The International Construction Measurement Standards (ICMS 1 & 2) that define the cost categories can be further explored. The work that has been done by professional societies in Canada, USA, UK, Malaysia, and Sweden on how costs can be captured and manipulated can be fully examined and studied. A detailed understanding of how costs are incurred, captured, manipulated and used in the construction sector can be provided in the future.

This study explores which methods are better for calculating weight and similarity based on different sample sizes. Except for sample size, other factors, including the case indexing, the settings of case retrieval, and case reuse, also influence the performance of the CBR model, making it difficult to understand the ECCE CBR model comprehensively. Therefore, to deepen the understanding of the ECCE CBR model, greater research attention should be paid to the influence of different CBR parameters and the inclusion of more predictor variables of residential building. As well, the validation of the prediction selection by experienced practitioners is recommended. More effort can be put into how the data are categorised before plugging into the model. Even though various ECCE CBR applications have been developed, there is still a big gap between research and cost estimation practice. The huge data bases that organisations such as highway agencies, housing providers, multilateral development banks, and large multi-disciplined professional practices hold, can be considered and explored. The popular trend of focusing on studying specific steps such as case retrieval, reuse, and revision has resulted in the other steps being ignored, which is hindering the overall development of ECCE CBR models. Further work is required to improve various aspects of the CBR model, including robustness, stability, accuracy and efficiency. More robust weight determination methods are encouraged to be combined with the CBR model, and a comparative study among different robust methods should be conducted in the future. Besides, as a database prediction system, a more diversified data processing method can be used in

the future (i.e., transforming the single factors into normal distributions for performing optimization computations).

Since CBR is a problem-solving process, variously described as addressing a current issue by recalling and reusing previous knowledge and experience, the ECCE CBR model is a dynamic system that requires continuously renewing and updating. The ECCE CBR applications should include not only the initial model development but also the maintenance of the case-base, and optimization of the system during long-term use. The existing research in combining CBM and CBR optimization in ECCE is insufficient. Further research is encouraged into the maintenance and operation of the ECCE CBR model. Moreover, various CBM strategies, including case-base reduction, partitioning, and optimization, should be explored in the ECCE CBR model to enhance the overall performance of the system.

As one of the artificial intelligence methods, the application of CBR can be extended to more research areas. Since the construction industry is facing large volumes of data throughout the life cycle stages of a project, CBR is encouraged to combine with more data mining techniques. Diversified information collection and processing should be used in the CBR model. Aside from using the numerical and nominal variables, the text should also be encoded as variables in the CBR system. Thus, CBR can become an experience mining system for solving problems, when combined with data and text mining techniques. In this way, experiences in previous construction projects can be turned into valuable resources. The CBR system can be further extended as a recommendation system that can accurately provide all the information for solving certain problems. Therefore, this study encourages future research to develop a more automated and intelligent CBR system with more diversified and complex structures for addressing the problems in the construction industry. Furthermore, more research work need to be conducted to answer the ultimate question, "*Is the CBR model better than an estimator's experience or not?"* The ECCE CBR model should be optimized from multiple perspectives to be competent in practice.The comparison of the performance between the ECCE CBR model and the expert is strongly encouraged to be explored. More research effort should also be made in the context of the industry.

Bibliography

(U.S.), P. M. I. (2017). *A guide to the project management body of knowledge (PMBOK guide)* (Six ed.): Newtown Square, Pennsylvania : Project Management Institute, Inc.

Aamodt, A., & Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications, 7*(1), 39-59. Retrieved from

Abdel-Aziz, A., & Hüllermeier, E. (2015). Case Base Maintenance in Preference-Based CBR. In (pp. 1-14): Springer International Publishing.

Adam, A., Josephson, P.-E. B., Lindahl, G. J. E., construction, & management, a. (2017). Aggregation of factors causing cost overruns and time delays in large public construction projects: trends and implications. *24*(3), 393-406. Retrieved from

Adeli, H., & Wu, M. (1998a). Regularization Neural Network for Construction Cost Estimation. *Journal of Construction Engineering and Management, 124*(1), 18-24. doi: doi:10.1061/(ASCE)0733-9364(1998)124:1(18)

Adeli, H., & Wu, M. (1998b). Regularization Neural Network for Construction Cost Estimation. *124*(1), 18-24. doi: doi:10.1061/(ASCE)0733-9364(1998)124:1(18)

Aha, D. W. (1992). Tolerating noisy, irrelevant and novel attributes in instance-based learning algorithms. *International Journal of Man-Machine Studies*
36(2), 267-287. Retrieved from

Aha, D. W., Kibler, D., & Albert, M. K. (1991a). Instance-based learning algorithms. *Machine learning*
6(1), 37-66. Retrieved from

Aha, D. W., Kibler, D., & Albert, M. K. (1991b). Instance-based learning algorithms. *Machine learning, 6*(1), 37-66. Retrieved from

Aha, D. W., Marling, C., & Watson, I. (2005). Case-based reasoning commentaries: introduction. *The Knowledge Engineering Review, 20*(03), 201-202. Retrieved from

Ahn, J., Ji, S.-H., Park, M., Lee, H.-S., Kim, S., & Suh, S.-W. (2014). The attribute impact concept: Applications in case-based reasoning and parametric cost estimation. *Automation in Construction, 43*, 195-203. doi: https://doi.org/10.1016/j.autcon.2014.03.011

Ahn, J., Park, M., Lee, H.-S., Ahn, S. J., Ji, S.-H., Song, K., & Son, B.-S. (2017). Covariance effect analysis of similarity measurement methods for early

construction cost estimation using case-based reasoning. *Automation in Construction, 81*, 254-266. doi: https://doi.org/10.1016/j.autcon.2017.04.009

Akinsiku, E. O., Babatunde, S. O., & Opawole, A. (2011). Comparative accuracy of floor area, storey enclosure and cubic methods in preparing preliminary estimate in Nigeria. *Journal of Building Appraisal, 6*(3-4), 315-322. Retrieved from

Akintoye, A., & Fitzgerald, E. (2000). A survey of current cost estimating practices in the UK. *Construction Management*
Economics, 18(2), 161-172. Retrieved from

Allison, P. D. (2001). *Missing data* (Vol. 136): Sage publications.

An, S.-H., Kim, G.-H., & Kang, K.-I. (2007). A case-based reasoning cost estimating model using experience by analytic hierarchy process. *Building and Environment, 42*(7), 2573-2579. doi: https://doi.org/10.1016/j.buildenv.2006.06.007

Anastasopoulos, P. C., Haddock, J. E., & Peeta, S. (2014). Cost overrun in public-private partnerships: Toward sustainable highway maintenance and rehabilitation. *Journal of Construction Engineering*
Management
140(6), 04014018. Retrieved from

Arafa, M., & Alqedra, M. (2011a). Early stage cost estimation of buildings construction projects using artificial neural networks. *4*(1). Retrieved from

Arafa, M., & Alqedra, M. (2011b). Early stage cost estimation of buildings construction projects using artificial neural networks. *Journal of Artificial Intelligence, 4*(1), 63-75. Retrieved from

ARCADIS. (2020). *Construction Cost Handbook: Malaysia*. Retrieved from https://images.arcadis.com/media/F/B/D/%7BFBD7F4D6-E86A-48AC-A33B-3069422CC9EB%7DConstruction%20Cost%20Handbook_Malaysia%202020.pdf

Arditi;, D., & Tokdemir, B. (1999). Comparison of Case-Based Reasoning and Artificial Neural Networks. *Journal of Computing in Civil Engineering, 13*(3), 162-169. doi: doi:10.1061/(ASCE)0887-3801(1999)13:3(162)

Ashley, K. D. (1991). *Modeling legal arguments: Reasoning with cases and hypotheticals*: MIT press.

Ashworth, A., & Perera, S. (2015). *Cost studies of buildings*: Routledge.

Azzeh, M., & Elsheikh, Y. (2017). Learning best K analogies from data distribution for case-based software effort estimation. *arXiv preprint arXiv:.04567*. Retrieved from

Balali, V., Noghabaei, M., Heydarian, A., & Han, K. (2018). Improved Stakeholder Communication and Visualizations: Real-Time Interaction and Cost Estimation within Immersive Virtual Environments. In *Construction Research Congress 2018* (pp. 522-530).

Ballou, D. P., & Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management science, 31*(2), 150-162. Retrieved from

Bannon Jr, W. (2015). Missing data within a quantitative research study: How to assess it, treat it, and why you should care. *Journal of the American Association of Nurse Practitioners, 27*(4), 230-232. Retrieved from

Barletta, R. (1991). An introduction to case-based reasoning. *AI expert, 6*(8), 42-49. Retrieved from

Barraza, G. A., Back, W. E., & Mata, F. (2000). Probabilistic monitoring of project performance using SS-curves. *Journal of Construction Engineering Management, 126*(2), 142-148. Retrieved from

Bartlett, F. C., & Burt, C. (1933). Remembering: A study in experimental and social psychology. *British Journal of Educational Psychology, 3*(2), 187-192. Retrieved from

Barua, S., Begum, S., & Ahmed, M. U. (2018). Towards distributed k-nn similarity for scalable case retrieval. In *XCBR: First Workshop on Case-Based Reasoning for the Explanation of Intelligent Systems. Workshop at the 26th International Conference on Case-Based Reasoning (ICCBR 2018)* (pp. 151-160).

Bergmann, R. (2001). Highlights of the European INRECA projects. In *Case-Based Reasoning Research and Development* (pp. 1-15): Springer.

Bertisen, J., & Davis, G. A. (2008a). Bias and error in mine project capital cost estimation. *The Engineering Economist, 53*(2), 118-139. Retrieved from

Bertisen, J., & Davis, G. A. J. T. E. E. (2008b). Bias and error in mine project capital cost estimation. *53*(2), 118-139. Retrieved from

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., . . . Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. *Advanced Engineering Informatics, 30*(3), 500-521. doi: https://doi.org/10.1016/j.aei.2016.07.001

Bode, J. (1998). Neural networks for cost estimation. *Cost Engineering 40*(1), 25. Retrieved from

Bode, J. (2000). Neural networks for cost estimation: Simulations and pilot application. *International Journal of Production Research, 38*(6), 1231-1254. doi: 10.1080/002075400188825

Bordat, C., McCullouch, B. G., Labi, S., & Sinha, K. C. (2004). An analysis of cost overruns and time delays of INDOT projects. Retrieved from

Botev, Z. I., Grotowski, J. F., & Kroese, D. P. J. T. a. o. S. (2010). Kernel density estimation via diffusion. *38*(5), 2916-2957. Retrieved from

Bowen, P., & Edwards, P. (1985). Cost modelling and price forecasting: practice and theory in perspective. *Construction Management Economics*
3(3), 199-215. Retrieved from

Briggs, A. H., Mooney, C. Z., & Wonderling, D. E. J. S. i. m. (1999). Constructing confidence intervals for cost‐effectiveness ratios: an evaluation of parametric and non‐parametric techniques using Monte Carlo simulation. *18*(23), 3245-3262. Retrieved from

Brighton, H., & Mellish, C. (1999). On the consistency of information filters for lazy learning algorithms. In *European conference on principles of data mining and knowledge discovery* (pp. 283-288): Springer.

Bromilow, F. (1969). Contract time performance expectations and the reality. In *Building Forum* (Vol. 1, pp. 70-80).

Burke, R. (2013). Project management: planning and control techniques. *26*. Retrieved from

Byung Soo, K. (2011). The approximate cost estimating model for railway bridge project in the planning phase using CBR method. *KSCE Journal of Civil Engineering, 15*(7), 1149-1159. doi: 10.1007/s12205-011-1342-2

Canesi, R., & Marella, G. (2017). Residential construction cost: An Italian survey. *Data in Brief, 11*, 231-235. doi: https://doi.org/10.1016/j.dib.2017.02.005

Cao, G., Shiu, S., & Wang, X. (2001). A fuzzy-rough approach for case base maintenance. In *International Conference on Case-Based Reasoning* (pp. 118-130): Springer.

Carr, R. I. J. J. o. C. E., & Management. (1989). Cost-estimating principles. *115*(4), 545-551. Retrieved from

CBIS. (2014). ISO 21500: Guidance on Project Management, . Retrieved from http://www.cbisco.com.au/iso21500-guidance-on-project-management/

Chan, A. P. (1999). Modelling building durations in Hong Kong. *Construction Management & Economics, 17*(2), 189-196. Retrieved from

Chan, V. K., & Wong, W. E. (2007). Outlier elimination in construction of software metric models. In *Proceedings of the 2007 ACM symposium on Applied computing* (pp. 1484-1488).

Chang, N. B., Chen, Y., Chen, H. J. J. o. E. S., & A, H. P. (1997). A fuzzy regression analysis for the construction cost estimation of wastewater treatment plants (I) theoretical development. *32*(4), 885-899. Retrieved from

Changchien, S. W., & Lin, M.-C. (2005). Design and implementation of a case-based reasoning system for marketing plans. *Expert systems with applications, 28*(1), 43-53. Retrieved from

Chau, A. D. (2018). *Conceptual Cost Estimation Decision Support System in University Construction Projects* (Ph.D.). The University of Alabama, Ann Arbor. Retrieved from https://gateway.library.qut.edu.au/login?url=https://search.proquest.com/docview/2186898161?accountid=13380
https://qut.primo.exlibrisgroup.com/openurl/61QUT_INST/61QUT_INST:61QUT??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:dissertation&genre=dissertations+%26+theses&sid=ProQ:ProQuest+Dissertations+%26+Theses+Global&atitle=&title=Conceptual+Cost+Estimation+Decision+Support+System+in+University+Construction+Projects&issn=&date=2018-01-01&volume=&issue=&spage=&au=Chau%2C+Anh+Duc&isbn=978-0-438-88833-3&jtitle=&btitle=&rft_id=info:eric/&rft_id=info:doi/Retrieved from ProQuest Dissertations & Theses Global database. (10976275).

Chen, D., & Burrell, P. (2001). Case-Based Reasoning System and Artificial Neural Networks: A Review. *Neural Computing & Applications, 10*(3), 264-276. doi: https://doi.org/10.1007/PL00009897

Cheng, M.-Y., Tsai, H.-C., & Hsieh, W.-S. (2009). Web-based conceptual cost estimates for construction projects using Evolutionary Fuzzy Neural Inference Model. *Automation in Construction, 18*(2), 164-172. Retrieved from

Cheung, F. K., & Skitmore, M. (2006a). Application of cross validation techniques for modelling construction costs during the very early design stage. *Building Environment, 41*(12), 1973-1990. Retrieved from

Cheung, F. K., & Skitmore, M. (2006b). A modified storey enclosure model. *Construction Management Economics*
24(4), 391-405. Retrieved from

Cheung, F. K. T., Rihan, J., Tah, J., Duce, D., & Kurul, E. (2012). Early stage multi-level cost estimation for schematic BIM models. *Automation in Construction, 27*(0), 67-77. doi: http://dx.doi.org/10.1016/j.autcon.2012.05.008

Cheung, F. K. T., & Skitmore, M. (2006c). Application of cross validation techniques for modelling construction costs during the very early design stage. *Building and Environment, 41*(12), 1973-1990. doi: https://doi.org/10.1016/j.buildenv.2005.09.011

Chiu, C. (2002). A case-based customer classification approach for direct marketing. *Expert Systems with Applications, 22*(2), 163-168. Retrieved from

Choi, S., Kim, D. Y., Han, S. H., & Kwak, Y. H. (2013). Conceptual cost-prediction model for public road planning via rough set theory and case-based reasoning. *Journal of Construction Engineering and Management, 140*(1). Retrieved from

Choi, S., Kim, D. Y., Han, S. H., & Kwak, Y. H. (2014). Conceptual Cost-Prediction Model for Public Road Planning via Rough Set Theory and Case-Based Reasoning. *140*(1), 04013026. doi: doi:10.1061/(ASCE)CO.1943-7862.0000743

Choong-Wan, K., TaeHoon, H., Chang-Taek, H., Park, S. H., & Joon-oh, S. (2010). 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), 121-137. doi: 10.3846/ijspm.2010.10

ChoongWan, K., TaeHoon, H., & ChangTaek, H. (2011). The development of a construction cost prediction model with improved prediction capacity using the advanced CBR approach. *Expert Systems with Applications, 38*(7), 8597-8606. doi: 10.1016/j.eswa.2011.01.063

ChoongWan, K., TaeHoon, H., ChangTaek, H., & KyoJin, K. (2010). 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*(5), 739-752. doi: 10.1139/L10-007

Chou, J.-S. (2008). Applying AHP-Based CBR to Estimate Pavement Maintenance Cost. *Tsinghua Science & Technology, 13, Supplement 1*, 114-120. doi: http://dx.doi.org/10.1016/S1007-0214(08)70136-6

Chou, J.-S. (2009). Web-based CBR system applied to early cost budgeting for pavement maintenance project. *Expert Systems with Applications, 36*(2, Part 2), 2947-2960. doi: https://doi.org/10.1016/j.eswa.2008.01.025

Chou, J.-S., Lin, C.-W., Pham, A.-D., & Shao, J.-Y. (2015). Optimized artificial intelligence models for predicting project award price. *Automation in Construction, 54*, 106-115. doi: https://doi.org/10.1016/j.autcon.2015.02.006

Coakley, C. W., & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association*
88(423), 872-880. Retrieved from

Coetzer, E. O., & Vlok, P. (2019). A standardised model to quantify the financial impact of poor engineering information quality in the oil and gas industry. *South African Journal of Industrial Engineering*
30(4), 131-142. Retrieved from

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2013). *Applied multiple regression/correlation analysis for the behavioral sciences*: Routledge.

Craw, S., Wiratunga, N., & Rowe, R. C. (2006). Learning adaptation knowledge to improve case-based reasoning. *Artificial Intelligence, 170*(16), 1175-1192. doi: https://doi.org/10.1016/j.artint.2006.09.001

Creese, R. C., & Li, L. (1995). Cost estimation of timber bridges using neural networks. *Cost Engineering, 37*(5), 17. Retrieved from

Croux, C., Rousseeuw, P. J., & Hössjer, O. (1994). Generalized S-estimators. *Journal of the American Statistical Association, 89*(428), 1271-1281. Retrieved from

Cummins, L., & Bridge, D. (2011). On dataset complexity for case base maintenance. In *International Conference on Case-Based Reasoning* (pp. 47-61): Springer.

Damir, A., Elkhatib, A., & Nassef, G. J. I. J. o. F. (2007). Prediction of fatigue life using modal analysis for grey and ductile cast iron. *29*(3), 499-507. Retrieved from

Dang, C. N., & Le-Hoai, L. J. E., Construction. (2018). Revisiting storey enclosure method for early estimation of structural building construction cost. *Engineering, Construction Architectural Management*
Retrieved from

Dang, C. N., Le-Hoai, L. J. E., Construction, & Management, A. (2018). Revisiting storey enclosure method for early estimation of structural building construction cost. *25*(7), 877-895. Retrieved from

Darko, A., & Chan, A. P. C. (2017). Review of Barriers to Green Building Adoption. [https://doi.org/10.1002/sd.1651]. *Sustainable Development, 25*(3), 167-179. doi: https://doi.org/10.1002/sd.1651

Daschbach, J. M., & Apgar, H. (1988). Design analysis through techniques of parametric cost estimation. *Engineering Costs and Production Economics, 14*(2), 87-93. doi: https://doi.org/10.1016/0167-188X(90)90111-T

Deckro, R. F., Hebert, J. E., Verdini, W. A., Grimsrud, P. H., & Venkateshwar, S. (1995). Nonlinear time/cost tradeoff models in project management. *Computers & Industrial Engineering, 28*(2), 219-229. Retrieved from

Delany, S. J., & Cunningham, P. (2004). An analysis of case-base editing in a spam filtering system. In *European Conference on Case-Based Reasoning* (pp. 128-141): Springer.

Dietterich, T. G. J. M. L. (1986). Learning at the knowledge level. *1*(3), 287-315. Retrieved from

Dikmen, I., Birgonul, M. T., & Gur, A. K. (2007). A case-based decision support tool for bid mark-up estimation of international construction projects. *Automation in Construction, 17*(1), 30-44. Retrieved from

Dogan, S. Z., Arditi, D., & Guenaydin, H. M. (2006). Determining attribute weights in a CBR model for early cost prediction of structural systems. *Journal of Construction Engineering and Management-Asce, 132*(10), 1092-1098. doi: 10.1061/(asce)0733-9364(2006)132:10(1092)

Dogan, S. Z., Arditi, D., & Guenaydin, H. M. (2008). Using decision trees for determining attribute weights in a case-based model of early cost prediction. *Journal of Construction Engineering and Management-Asce, 134*(2), 146-152. doi: 10.1061/(asce)0733-9364(2008)134:2(146)

Doğan, S. Z., Arditi, D., & Günaydin, H. M. (2008). Using Decision Trees for Determining Attribute Weights in a Case-Based Model of Early Cost Prediction. *134*(2), 146-152. doi: doi:10.1061/(ASCE)0733-9364(2008)134:2(146)

Doğan, S. Z., Arditi, D., & Günaydın, H. M. (2006a). Determining Attribute Weights in a CBR Model for Early Cost Prediction of Structural Systems. *132*(10), 1092-1098. doi: doi:10.1061/(ASCE)0733-9364(2006)132:10(1092)

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

Dogan, S. Z., Arditi, D., & Murat Gunaydin, H. (2008). Using decision trees for determining attribute weights in a case-based model of early cost prediction. *Journal of Construction Engineering and Management, 134*(2), 146-152. doi: 10.1061/(ASCE)0733-9364(2008)134:2(146)

Doğan, S. Z., Arditi, D., & Murat Günaydin, H. (2008). Using decision trees for determining attribute weights in a case-based model of early cost prediction. [Article]. *Journal of Construction Engineering and Management, 134*(2), 146-152. doi: 10.1061/(ASCE)0733-9364(2008)134:2(146)

Du, J., & Bormann, J. (2014a). Improved Similarity Measure in Case-Based Reasoning with Global Sensitivity Analysis: An Example of Construction Quantity Estimating. *28*(6), 04014020. doi: doi:10.1061/(ASCE)CP.1943-5487.0000267

Du, J., & Bormann, J. (2014b). Improved similarity measure in case-based reasoning with global sensitivity analysis: An example of construction quantity estimating. *Journal of Computing in Civil Engineering*. Retrieved from

Dysert, L., & Elliott, B. G. E. (2002). The estimate review and validation process. *Cost Engineering, 44*(1), 17. Retrieved from

Famiyeh, S., Amoatey, C. T., Adaku, E., & Agbenohevi, C. S. (2017). Major causes of construction time and cost overruns: A case of selected educational sector projects in Ghana. *Journal of Engineering, Design and Technology, 15*(2), 181-198. Retrieved from

Flyvbjerg, B. (2007). Policy and planning for large-infrastructure projects: problems, causes, cures. *Environment Planning B: planning design*
34(4), 578-597. Retrieved from

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? *Transport reviews, 23*(1), 71-88. Retrieved from

Fortune, C., & Lees, M. J. U. R. R. S. C. o. T. (1989). An investigation into Methods of Early Cost Advice for Clients. Retrieved from

Fulkerson, D. R. (1961). A network flow computation for project cost curves. *Management science, 7*(2), 167-178. Retrieved from

García-Gil, D., Luengo, J., García, S., & Herrera, F. J. I. S. (2019). Enabling smart data: noise filtering in big data classification. *479*, 135-152. Retrieved from

Gardner, B. J., Gransberg, D. D., & Jeong, H. D. (2016). Reducing data-collection efforts for conceptual cost estimating at a highway agency. *Journal of Construction Engineering and Management, 142*(11), 04016057. Retrieved from

Gardner, B. J., Gransberg, D. D., Jeong, H. D. J. J. o. C. E., & Management. (2016). Reducing data-collection efforts for conceptual cost estimating at a highway agency. *142*(11), 04016057. Retrieved from

Gardner Brendon, J., Gransberg Douglas, D., & Jeong, H. D. (2016). Reducing Data-Collection Efforts for Conceptual Cost Estimating at a Highway Agency. *Journal of Construction Engineering and Management, 142*(11), 04016057. doi: 10.1061/(ASCE)CO.1943-7862.0001174

Gates, G. (1972). The reduced nearest neighbor rule *IEEE transactions on information theory*
18(3), 431-433. Retrieved from

Gervini, D., & Yohai, V. J. (2002). A class of robust and fully efficient regression estimators. *The Annals of Statistics, 30*(2), 583-616. Retrieved from

Graves, D., & Pedrycz, W. (2009). Fuzzy prediction architecture using recurrent neural networks. *Neurocomputing, 72*(7-9), 1668-1678. doi: 10.1016/j.neucom.2008.07.009

Greene, D., Freyne, J., Smyth, B., & Cunningham, P. (2008). An analysis of research themes in the CBR conference literature. In *Advances in case-based reasoning* (pp. 18-43): Springer.

Group, A. (2019). *Canadian Cost Guide 2019*. Retrieved from https://www.altusgroup.com/services/reports/canadian-cost-guide-2019/

Guan, D., Yuan, W., Lee, Y.-K., & Lee, S. (2009). Nearest neighbor editing aided by unlabeled data. *Information Sciences*
179(13), 2273-2282. Retrieved from

Günaydın, H. M., & Doğan, S. Z. (2004). A neural network approach for early cost estimation of structural systems of buildings. *International Journal of Project Management, 22*(7), 595-602. Retrieved from

Haavisto, A. (2015). Early stage cost estimation of medium-sized construction company's new residential building production. Retrieved from

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2010). *Multivariate data analysis* (Vol. 5): Prentice hall Upper Saddle River, NJ.

Hammond, K. J. (1986). CHEF: A Model of Case-based Planning. In *AAAI* (pp. 267-271).

Hampton, G., Baldwin, A. N., & Holt, G. (2012). Project delays and cost: stakeholder perceptions of traditional v. PPP procurement. *Journal of Financial Management of Property*
Construction
17(1), 73-91. Retrieved from

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Handschin, E., Schweppe, F. C., Kohlas, J., & Fiechter, A. (1975). Bad data analysis for power system state estimation. *IEEE Transactions on Power Apparatus*
Systems, 94(2), 329-337. Retrieved from

Haouchine, M.-K., Chebel-Morello, B., & Zerhouni, N. (2007, 2007-06). Case Base Maintenance Approach. In *International Conference on Industrial Engineering and Systems Management, IESM'2007.* (pp. sur CD ROM - 10 pages): Yang Shanlin, Chen Guoqing, Thomas Andre, Artiba Abdelhakim, Xu Zongwei.

Haouchine, M.-K., Chebel-Morello, B., & Zerhouni, N. (2008). Competence-Preserving Case-Deletion Strategy for Case-Base Maintenance. In.

Hart, P. (1968). The condensed nearest neighbor rule (Corresp.). *IEEE transactions on information theory*
14(3), 515-516. Retrieved from

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*: Springer Science & Business Media.

Hegazy, T., & Ayed, A. (1998). Neural network model for parametric cost estimation of highway projects. *Journal of Construction Engineering and Management, 124*(3), 210-218. Retrieved from

Hegazy, T., Ayed, A. J. J. o. C. E., & Management. (1998). Neural network model for parametric cost estimation of highway projects. *124*(3), 210-218. Retrieved from

Hegazy, T., & MOSELHI, O. (1994). Estimator: a prototype of an integrated bid preparation system. *Engineering, Construction Architectural Management* *1*(1), 51-67. Retrieved from

Hendrickson, C., Hendrickson, C. T., & Au, T. (2008). *Project management for construction: Fundamental concepts for owners, engineers, architects, and builders*: Chris Hendrickson.

Hong, T., Hyun, C., & Moon, H. (2011). CBR-based cost prediction model-II of the design phase for multi-family housing projects. *Expert Systems with Applications, 38*(3), 2797-2808. doi: https://doi.org/10.1016/j.eswa.2010.08.071

Howell, D., Rogier, M., Yzerbyt, V., & Bestgen, Y. (1998). *Statistical methods in human sciences*: New York: Wadsworth.

Hu, X., Xia, B., Skitmore, M., & Chen, Q. The application of case-based reasoning in construction management research: An overview. *Automation in Construction*. doi: http://dx.doi.org/10.1016/j.autcon.2016.08.023

Huber, P. J. (2011). *Robust statistics*: Springer.

Huber, P. J. J. T. A. o. S. (1984). Finite sample breakdown of M-and P-estimators. 119-126. Retrieved from

Hwang, S. J. J. o. C. E., & Management. (2009). Dynamic regression models for prediction of construction costs. *135*(5), 360-367. Retrieved from

Hyari, K. H., Al-Daraiseh, A., & El-Mashaleh, M. (2015). Conceptual cost estimation model for engineering services in public construction projects. *Journal of Management in Engineering, 32*(1), 04015021. Retrieved from

Ilyas, I. F., Chu, X. J. F., & Databases, T. i. (2015). Trends in cleaning relational data: Consistency and deduplication. *5*(4), 281-393. Retrieved from

Irfan, M., Khurshid, M. B., Anastasopoulos, P., Labi, S., & Moavenzadeh, F. (2011). Planning-stage estimation of highway project duration on the basis of anticipated project cost, project type, and contract type. *International Journal of Project Management, 29*(1), 78-92. doi: https://doi.org/10.1016/j.ijproman.2010.01.001

Jaeckel, L. A. J. T. A. o. M. S. (1972). Estimating regression coefficients by minimizing the dispersion of the residuals. 1449-1458. Retrieved from

Jafarzadeh, R., Ingham, J., Walsh, K., Hassani, N., Ghodrati Amiri, G. J. J. o. C. E., & Management. (2014). Using statistical regression analysis to establish

construction cost models for seismic retrofit of confined masonry buildings. *141*(5), 04014098. Retrieved from

Jafarzadeh, R., Ingham, J. M., Wilkinson, S., González, V., & Aghakouchak, A. A. (2014). Application of Artificial Neural Network Methodology for Predicting Seismic Retrofit Construction Costs. *140*(2), 04013044. doi: doi:10.1061/(ASCE)CO.1943-7862.0000725

Jahren, C. T., Ashe, A. M. J. J. o. C. E., & management. (1990). Predictors of cost-overrun rates. *116*(3), 548-552. Retrieved from

Jalali, V., & Leake, D. (2014). Adaptation-guided case base maintenance. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Ji, C., Hong, T., & Hyun, C. (2010a). CBR revision model for improving cost prediction accuracy in multifamily housing projects. *Journal of Management in Engineering*. Retrieved from

Ji, C., Hong, T., & Hyun, C. (2010b). CBR revision model for improving cost prediction accuracy in multifamily housing projects. *Journal of Management in Engineering, 26*(4), 229-236. Retrieved from

Ji, C., Hong, T., & Hyun, C. (2010c). CBR Revision Model for Improving Cost Prediction Accuracy in Multifamily Housing Projects. *26*(4), 229-236. doi: doi:10.1061/(ASCE)ME.1943-5479.0000018

Ji, S.-H., Ahn, J., Lee, E.-B., & Kim, Y. (2018). Learning method for knowledge retention in CBR cost models. *Automation in Construction, 96*, 65-74. doi: https://doi.org/10.1016/j.autcon.2018.08.019

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

Ji, S.-H., Park, M., & Lee, H.-S. (2012a). Case Adaptation Method of Case-Based Reasoning for Construction Cost Estimation in Korea. *Journal of Construction Engineering and Management, 138*(1), 43-52. doi: doi:10.1061/(ASCE)CO.1943-7862.0000409

Ji, S.-H., Park, M., & Lee, H.-S. (2012b). Case Adaptation Method of Case-Based Reasoning for Construction Cost Estimation in Korea. *138*(1), 43-52. doi: doi:10.1061/(ASCE)CO.1943-7862.0000409

Ji, S.-H., Park, M., Lee, H.-S., Ahn, J., Kim, N., & Son, B. (2011a). Military Facility Cost Estimation System Using Case-Based Reasoning in Korea. *Journal of Computing in Civil Engineering, 25*(3), 218-231. doi: 10.1061/(asce)cp.1943-5487.0000082

Ji, S.-H., Park, M., Lee, H.-S., Ahn, J., Kim, N., & Son, B. (2011b). Military Facility Cost Estimation System Using Case-Based Reasoning in Korea. *25*(3), 218-231. doi: doi:10.1061/(ASCE)CP.1943-5487.0000082

Jin, R., Cho, K., Hyun, C., & Son, M. (2012). MRA-based revised CBR model for cost prediction in the early stage of construction projects. *Expert Systems with Applications, 39*(5), 5214-5222. doi: https://doi.org/10.1016/j.eswa.2011.11.018

Jin, R., Han, S., Hyun, C., & Kim, J. (2014). Improving Accuracy of Early Stage Cost Estimation by Revising Categorical Variables in a Case-Based Reasoning Model. *Journal of Construction Engineering and Management, 140*(7). Retrieved from

Jin, R., Han, S., Hyun, C., Kim, J. J. J. o. C. E., & Management. (2014). Improving accuracy of early stage cost estimation by revising categorical variables in a case-based reasoning model. *140*(7), 04014025. Retrieved from

Juszczyk, M. (2017). The challenges of nonparametric cost estimation of construction works with the use of artificial intelligence tools. *Procedia engineering, 196*, 415-422. Retrieved from

Juszczyk, M., Leśniak, A., & Zima, K. (2018). ANN based approach for estimation of construction costs of sports fields. *Complexity, 2018*. Retrieved from

Kaka, A., & Price, A. D. (1991). Relationship between value and duration of construction projects. *Construction Management and Economics, 9*(4), 383-400. Retrieved from

Kanj, S., Abdallah, F., Denœux, T., Tout, K. J. P. A., & Applications. (2016). Editing training data for multi-label classification with the k-nearest neighbor rule. [journal article]. *19*(1), 145-161. doi: 10.1007/s10044-015-0452-8

KARANCI, H. (2010). *A comparative study of regression analysis, neural networks and case–based reasoning for early range cost estimation of mass housing projects*. Middle East Technical University.

Karkouch, A., Mousannif, H., Al Moatassime, H., & Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. *Journal of Network and Computer Applications, 73*, 57-81. doi: https://doi.org/10.1016/j.jnca.2016.08.002

Keller, G. (2015). *Statistics for Management and Economics*: Cengage Learning.

Khan, M. J., Hayat, H., & Awan, I. (2019a). Hybrid case-base maintenance approach for modeling large scale case-based reasoning systems. *Human-centric Computing Information Sciences*
9(1), 9. Retrieved from

Khan, M. J., Hayat, H., & Awan, I. (2019b). Hybrid case-base maintenance approach for modeling large scale case-based reasoning systems. [journal article].

Human-centric Computing and Information Sciences, 9(1), 9. doi: 10.1186/s13673-019-0171-z

Khosrowshahi, F., Kaka, A. P. J. B., & Environment. (1996). Estimation of project total cost and duration for housing projects in the UK. *31*(4), 375-383. Retrieved from

Kim, B.-s., & Hong, T. (2012a). Revised Case-Based Reasoning Model Development Based on Multiple Regression Analysis for Railroad Bridge Construction. *Journal of Construction Engineering and Management-Asce, 138*(1), 154-162. doi: 10.1061/(asce)co.1943-7862.0000393

Kim, B.-s., & Hong, T. (2012b). Revised Case-Based Reasoning Model Development Based on Multiple Regression Analysis for Railroad Bridge Construction. *138*(1), 154-162. doi: doi:10.1061/(ASCE)CO.1943-7862.0000393

Kim, B. S. J. K. J. o. C. E. (2011). The approximate cost estimating model for railway bridge project in the planning phase using CBR method. [journal article]. *15*(7), 1149. doi: 10.1007/s12205-011-1342-2

Kim, G.-H., An, S.-H., & Kang, K.-I. (2004a). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building Environmental Impact Assessment Review, 39*(10), 1235-1242. Retrieved from

Kim, G.-H., An, S.-H., & Kang, K.-I. (2004b). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. *Building and Environment, 39*(10), 1235-1242. doi: https://doi.org/10.1016/j.buildenv.2004.02.013

Kim, G., Seo, D., & Kang, K. I. J. J. o. C. i. C. E. (2005). Hybrid models of neural networks and genetic algorithms for predicting preliminary cost estimates. *19*(2), 208-211. Retrieved from

Kim, H.-J., Seo, Y.-C., & Hyun, C.-T. (2012). A hybrid conceptual cost estimating model for large building projects. *Automation in Construction, 25*, 72-81. doi: https://doi.org/10.1016/j.autcon.2012.04.006

Kim, K. J., & Kim, K. (2010a). Preliminary Cost Estimation Model Using Case-Based Reasoning and Genetic Algorithms. *24*(6), 499-505. doi: doi:10.1061/(ASCE)CP.1943-5487.0000054

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

Kim, M., Lee, S., Woo, S., & Shin, D. (2012). Approximate cost estimating model for river facility construction based on case-based reasoning with genetic algorithms. *KSCE Journal of Civil Engineering, 16*(3), 283-292. doi: 10.1007/s12205-012-1482-z

Kim, S.-Y., Choi, J.-W., Kim, G.-H., & Kang, K.-I. (2005). Comparing cost prediction methods for apartment housing projects: CBR versus ANN. *Journal of Asian Architecture and Building Engineering, 4*(1), 113-120. Retrieved from

Kim, S. (2012). Interval estimation of construction cost using case-based reasoning and genetic algorithms. *Journal of Asian Architecture and Building Engineering, 11*(2), 327-334. Retrieved from

Kim, S. (2013). Hybrid forecasting system based on case-based reasoning and analytic hierarchy process for cost estimation. *Journal of Civil Engineering and Management, 19*(1), 86-96. doi: 10.3846/13923730.2012.737829

Kim, S., & Shim, J. H. (2013a). Combining case-based reasoning with genetic algorithm optimization for preliminary cost estimation in construction industry. *Canadian Journal of Civil Engineering, 41*(1), 65-73. doi: 10.1139/cjce-2013-0223

Kim, S., & Shim, J. H. J. C. J. o. C. E. (2013b). Combining case-based reasoning with genetic algorithm optimization for preliminary cost estimation in construction industry. *41*(1), 65-73. Retrieved from

Kim, S. J. J. o. C. E., & Management. (2013). Hybrid forecasting system based on case-based reasoning and analytic hierarchy process for cost estimation. *19*(1), 86-96. Retrieved from

Kocaguneli, E., Menzies, T., Bener, A., & Keung, J. W. (2011). Exploiting the essential assumptions of analogy-based effort estimation. *IEEE transactions on software engineering, 38*(2), 425-438. Retrieved from

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (Vol. 14, pp. 1137-1145): Montreal, Canada.

Kolodneer, J. L. (1991). Improving human decision making through case-based decision aiding. *AI magazine*
12(2), 52-52. Retrieved from

Kolodner, J. (2014). *Case-based reasoning*: Morgan Kaufmann.

Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review, 6*(1), 3-34. Retrieved from

Koo, C., Hong, T., & Hyun, C. (2011). The development of a construction cost prediction model with improved prediction capacity using the advanced CBR approach. *Expert Systems with Applications, 38*(7), 8597-8606. doi: https://doi.org/10.1016/j.eswa.2011.01.063

Koo, C., Hong, T., Hyun, C., & Koo, K. (2010). 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*(5), 739-752. doi: https://doi.org/10.1139/L10-007

Koo, C. W., Hong, T., Hyun, C. T., Park, S. H., & Seo, J. o. (2010). 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), 121-137. doi: https://doi.org/10.3846/ijspm.2010.10

Kouskoulas, V., & Koehn, E. (1974). Predesign cost-estimation function for buildings. *Journal of the Construction Division*
100(4), 589-604. Retrieved from

Lai, C.-c., & Lee, W.-l. (2006). A WICE approach to real-time construction cost estimation. *Automation in construction, 15*(1), 12-19. Retrieved from

Laranjeiro, N., Soydemir, S. N., & Bernardino, J. (2015, 18-20 Nov. 2015). A Survey on Data Quality: Classifying Poor Data. In *2015 IEEE 21st Pacific Rim International Symposium on Dependable Computing (PRDC)* (pp. 179-188).

Larsen, J. K., Shen, G. Q., Lindhard, S. M., & Brunoe, T. D. (2015a). Factors affecting schedule delay, cost overrun, and quality level in public construction projects. *Journal of Management in Engineering, 32*(1), 04015032. Retrieved from

Larsen, J. K., Shen, G. Q., Lindhard, S. M., & Brunoe, T. D. (2015b). Factors affecting schedule delay, cost overrun, and quality level in public construction projects. *Journal of Management in Engineering*
32(1), 04015032. Retrieved from

Lawanna, A., & Daengdej, J. (2010). Hybrid technique and competence-preserving case deletion methods for case maintenance in case-based reasoning. *International Journal of Engineering Science Technology*
64. Retrieved from

Leake, D., & Schack, B. (2015). Flexible Feature Deletion: Compacting Case Bases by Selectively Compressing Case Contents. In (pp. 212-227): Springer International Publishing.

Leake, D., & Schack, B. (2016). Adaptation-guided feature deletion: testing recoverability to guide case compression. In *International Conference on Case-Based Reasoning* (pp. 234-248): Springer.

Leake, D., & Schack, B. (2018). Exploration vs. Exploitation in Case-Base Maintenance: Leveraging Competence-Based Deletion with Ghost Cases. In (pp. 202-218): Springer International Publishing.

Leake, D. B., & Wilson, D. C. (1998a, 1998//). Categorizing case-base maintenance: Dimensions and directions. In B. Smyth & P. Cunningham (Eds.), *Advances in Case-Based Reasoning* (pp. 196-207): Springer Berlin Heidelberg.

Leake, D. B., & Wilson, D. C. (1998b). Categorizing case-base maintenance: Dimensions and directions. In *European Workshop on Advances in Case-Based Reasoning* (pp. 196-207): Springer.

Leake, D. B., & Wilson, D. C. (2000). Guiding case-base maintenance: Competence and performance. In *Proceedings of the 14th European Conference on Artificial Intelligence Workshop on Flexible Strategies for Maintaining Knowledge Containers*: Citeseer.

Lee, S., Jin, Y., Woo, S., & Shin, D. H. (2013a). Approximate cost estimating model of eco-type trade for river facility construction using case-based reasoning and genetic algorithms. *Ksce Journal of Civil Engineering, 17*(2), 292-300. doi: 10.1007/s12205-013-1638-5

Lee, S., Jin, Y., Woo, S., & Shin, D. H. J. K. J. o. C. E. (2013b). Approximate cost estimating model of eco-type trade for river facility construction using case-based reasoning and genetic algorithms. [journal article]. *17*(2), 292-300. doi: 10.1007/s12205-013-1638-5

Lee, Y., MacEachern, S. N., & Jung, Y. (2012). Regularization of case-specific parameters for robustness and efficiency. *Statistical Science*
27(3), 350-372. Retrieved from

Leśniak, A., & Zima, K. (2018). Cost Calculation of Construction Projects Including Sustainability Factors Using the Case Based Reasoning (CBR) Method. *10*(5), 1608. Retrieved from https://www.mdpi.com/2071-1050/10/5/1608

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764-766. Retrieved from

Li, H. (1995). Neural networks for construction cost estimation. *Building Research & Information, 23*(5), 279-284. doi: 10.1080/09613219508727476

Li, H., Love, P. E. J. C. M., & Economics. (1999). Combining rule-based expert systems and artificial neural networks for mark-up estimation. *17*(2), 169-176. Retrieved from

Li, H., Shen, Q., & Love, P. E. (2005). Cost modelling of office buildings in Hong Kong: an exploratory study. *Facilities, 23*(9/10), 438-452. Retrieved from

Liao, Y., Vemuri, V. R. J. C., & security. (2002). Use of k-nearest neighbor classifier for intrusion detection. *21*(5), 439-448. Retrieved from

Lieber, J. (1994). A criterion of comparison between two case bases. In *European Workshop on Advances in Case-Based Reasoning* (pp. 87-100): Springer.

Liu, Z., & Xie, K. (2013, 29-30 June 2013). Research on the Early-Warning for Construction Project Cost Risk. In *2013 Fourth International Conference on Digital Manufacturing & Automation* (pp. 1127-1129).

Lopez De Mantaras, R., McSherry, D., Bridge, D., Leake, D., Smyth, B., Craw, S., . . . Watson, I. A. N. (2005). Retrieval, reuse, revision and retention in case-based reasoning. *The Knowledge Engineering Review, 20*(3), 215-240. Retrieved from https://gateway.library.qut.edu.au/login?url=https://search.proquest.com/docview/217515215?accountid=13380
https://qut.primo.exlibrisgroup.com/openurl/61QUT_INST/61QUT_INST:61QUT??url_ver=Z39.88-2004&rft_val_fmt=info:ofi/fmt:kev:mtx:journal&genre=article&sid=ProQ:ProQ%3Aabiglobal&atitle=Retrieval%2C+reuse%2C+revision+and+retention+in+case-based+reasoning&title=The+Knowledge+Engineering+Review&issn=02698889&date=2005-09-01&volume=20&issue=3&spage=215&au=LOPEZ+DE+MANTARAS%2C+RAMON%3BMCSHERRY%2C+DAVID%3BBRIDGE%2C+DEREK%3BLEAKE%2C+DAVID%3BSMYTH%2C+BARRY%3BCRAW%2C+SUSAN%3BFALTINGS%2C+BOI%3BMAHER%2C+MARY+LOU%3BCOX%2C+MICHAEL+T%3BFORBUS%2C+KENNETH%3BKEANE%2C+MARK%3BAAMODT%2C+AGNAR%3BWATSON%2C+IAN&isbn=&jtitle=The+Knowledge+Engineering+Review&btitle=&rft_id=info:eric/&rft_id=info:doi/

Lowe, D. J., Emsley, M. W., & Harding, A. (2006). Predicting construction cost using multiple regression techniques. *Journal of construction engineering management, 132*(7), 750-758. Retrieved from

Lowe, D. J., Emsley, M. W., Harding, A. J. J. o. c. e., & management. (2006). Predicting construction cost using multiple regression techniques. *132*(7), 750-758. Retrieved from

Lupiani, E., Juarez, J. M., & Palma, J. (2014a). Evaluating Case-Base Maintenance algorithms. *Knowledge-Based Systems, 67*, 180-194. doi: https://doi.org/10.1016/j.knosys.2014.05.014

Lupiani, E., Juarez, J. M., & Palma, J. J. K.-B. S. (2014b). Evaluating case-base maintenance algorithms. *67*, 180-194. Retrieved from

Lupiani, E., Massie, S., Craw, S., Juarez, J. M., & Palma, J. J. J. o. I. I. S. (2016). Case-base maintenance with multi-objective evolutionary algorithms. [journal article]. *46*(2), 259-284. doi: 10.1007/s10844-015-0378-z

Madhusudan, T., Zhao, J. L., & Marshall, B. (2004). A case-based reasoning framework for workflow model management. *Data & Knowledge Engineering, 50*(1), 87-115. Retrieved from

Mallows, C. L. (1975). On some topics in robustness. *Unpublished memorandum, Bell Telephone Laboratories, Murray Hill, NJ*. Retrieved from

Mark, S., Philip, L., & Adrian, T. (Singer-songwriters). (2015). Research methods for business students. On: Prentice Hall.

Marshall, B., Cardon, P., Poddar, A., & Fontenot, R. (2013). Does Sample Size Matter in Qualitative Research?: A Review of Qualitative Interviews in is Research. *Journal of Computer Information Systems, 54*(1), 11-22. doi: 10.1080/08874417.2013.11645667

Martin Skitmore, R., & Thomas Ng, S. (2003). Forecast models for actual construction time and cost. *Building and Environment, 38*(8), 1075-1083. doi: https://doi.org/10.1016/S0360-1323(03)00067-2

Marzouk, M. M., & Ahmed, R. M. (2011). A case-based reasoning approach for estimating the costs of pump station projects. *Journal of Advanced Research, 2*(4), 289-295. doi: http://dx.doi.org/10.1016/j.jare.2011.01.007

Massie, S., Craw, S., & Wiratunga, N. (2005). Complexity-guided case discovery for case based reasoning. In *AAAI* (Vol. 5, pp. 216-221).

Massie, S., Craw, S., & Wiratunga, N. (2006). Complexity profiling for informed case-base editing. In *European Conference on Case-Based Reasoning* (pp. 325-339): Springer.

McKenna, E., & Smyth, B. (2000). Competence-guided case-base editing techniques. In *European Workshop on Advances in Case-Based Reasoning* (pp. 186-197): Springer.

Mckenna, E., & Smyth, B. J. A. I. (2001). An interactive visualisation tool for case-based reasoners. *14*(1), 95-114. Retrieved from

Merrow, E. W., McDonnell, L., & Argüden, R. Y. (1988). *Understanding the outcomes of megaprojects: A quantitative analysis of very large civilian projects*: Rand Corporation Santa Monica, CA.

Mills, A., Lawther, & P and Jones, D. (2006). *Standard Methods of Measurement - A comparative study of national and regional publications*. Australian Institute of Quantity Suveyors and Royal Institution Chartered Surveyors. Retrieved from

Minton, S. (1990). Quantitative results concerning the utility of explanation-based learning. *Artificial Intelligence, 42*(2-3), 363-391. Retrieved from

Mitchell, T. M. (Singer-songwriter). (1997). Machine learning. WCB. On: McGraw-Hill Boston, MA:.

Morcous, G., Rivard, H., & Hanna, A. (2002). Case-based reasoning system for modeling infrastructure deterioration. *Journal of computing in civil engineering, 16*(2), 104-114. Retrieved from

Moselhi, O., & Siqueira, I. J. A. I. T. (1998). Neural networks for cost estimating of structural steel buildings. IT22. Retrieved from

Motrenko, A., Strijov, V., & Weber, G.-W. (2014). Sample size determination for logistic regression. *Journal of Computational and Applied Mathematics, 255*, 743-752. doi: http://dx.doi.org/10.1016/j.cam.2013.06.031

Motwani, J., Kumar, A., & Novakoski, M. J. W. s. (1995). Measuring construction productivity: a practical approach. *44*(8), 18-20. Retrieved from

Mulla, S. S., & Waghmare, A. P. (2015). A Study of Factors Caused for Time & Cost Overruns in Construction Project & their Remedial Measures. *International Journal Of Engineering Research And Applications, 5*(1), 48-53. Retrieved from

Muse, A., Sullivan, J., & Smith, P. (2014). Improving certainty in construction: the need for international standards. In *World Congress on Cost Engineering, Project Management and Quantity Surveying, AACE American Association of Cost Engineers*: Federation of Scientific & Technical Associations (FAST).

Nakhjiri, N., Salamó, M., & Sànchez-marrè, M. (2019). Reputation-Based Maintenance in Case-Based Reasoning. *Knowledge-Based Systems*, 105283. doi: https://doi.org/10.1016/j.knosys.2019.105283

Nassar, K. M., Gunnarsson, H. G., & Hegab, M. Y. (2005). Using Weibull analysis for evaluation of cost and schedule performance. *Journal of Construction Engineering Management, 131*(12), 1257-1262. Retrieved from

Neter, J., Wasserman, W., & Kutner, M. H. (1989). Applied linear regression models. Retrieved from

Ng, T. (1996). *Case-based reasoning decision support for contractor prequalification* PhD. University of Manchester Institute of Science and Technology.

Ni, Z.-W., Liu, Y., Li, F.-G., & Yang, S.-L. (2005). Case base maintenance based on outlier data mining. In *2005 International Conference on Machine Learning and Cybernetics* (Vol. 5, pp. 2861-2864): IEEE.

Oberlender, G. D., & Trost, S. M. (2001a). Predicting Accuracy of Early Cost Estimates Based on Estimate Quality. *127*(3), 173-182. doi: doi:10.1061/(ASCE)0733-9364(2001)127:3(173)

Oberlender, G. D., & Trost, S. M. (2001b). Predicting accuracy of early cost estimates based on estimate quality. *Journal of construction engineering management, 127*(3), 173-182. Retrieved from

Odeck, J. (2004). Cost overruns in road construction—what are their sizes and determinants? *Transport Policy, 11*(1), 43-53. doi: https://doi.org/10.1016/S0967-070X(03)00017-9

Ofori, G. (2015). Nature of the construction industry, its needs and its development: A review of four decades of research. *Journal of construction in developing countries, 20*(2), 115. Retrieved from

Ogunsemi, D. R., & Jagboro, G. O. (2006). Time - cost model for building projects in Nigeria. *Construction Management and Economics, 24*(3), 253-258. doi: 10.1080/01446190500521041

Ökmen, Ö., Öztaş, A. J. C. M., & Economics. (2010). Construction cost analysis under uncertainty with correlated cost risk analysis model. *28*(2), 203-212. Retrieved from

Pal, B., Mhashilkar, A., Pandey, A., Nagphase, B., & Chandanshive, V. (2018). Cost Estimation Model (CEM) of Buildings by ANN (Artificial Neural Networks)– A Review. *Neural networks*
5(2). Retrieved from

Pal, S. K., & Shiu, S. C. (2004). *Foundations of soft case-based reasoning* (Vol. 8): John Wiley & Sons.

Pallant, J. (2013). *SPSS survival manual*: McGraw-Hill Education (UK).

Pan, R., Yang, Q., & Pan, S. J. (2007a). Mining competent case bases for case-based reasoning. *Artificial Intelligence, 171*(16), 1039-1068. doi: https://doi.org/10.1016/j.artint.2007.04.018

Pan, R., Yang, Q., & Pan, S. J. J. A. I. (2007b). Mining competent case bases for case-based reasoning. *171*(16-17), 1039-1068. Retrieved from

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *Ann. Math. Statist., 33*(3), 1065-1076. doi: 10.1214/aoms/1177704472

Perera, S., & Watson, I. (1998). Collaborative case-based estimating and design. *Advances in Engineering Software, 29*(10), 801-808. doi: http://dx.doi.org/10.1016/S0965-9978(97)00064-1

Peško, I., Mučenski, V., Šešlija, M., Radović, N., Vujkov, A., Bibić, D., & Krklješ, M. J. C. (2017). Estimation of costs and durations of construction of urban roads using ann and svm. *2017*. Retrieved from

Petroutsatou, K., Georgopoulos, E., Lambropoulos, S., & Pantouvakis, J. (2011). Early cost estimating of road tunnel construction using neural networks. *Journal of construction engineering management*
138(6), 679-687. Retrieved from

Phaobunjong, K. (2002). *Parametric cost estimating model for conceptual cost estimating of building construction projects*.

Pindyck, R. S., & Rubinfeld, D. L. (1998). *Econometric models and economic forecasts* (Vol. 4): Irwin/McGraw-Hill Boston.

Pohl, G., & Mihaljek, D. (1992). Project evaluation and uncertainty in practice: A statistical analysis of rate-of-return divergences of 1,015 World Bank projects. *The World Bank Economic Review*
6(2), 255-277. Retrieved from

Raisbeck, P., Duffield, C., Xu, M. J. C. M., & Economics. (2010). Comparative performance of PPPs and traditional procurement in Australia. *28*(4), 345-359. Retrieved from

Richter, M. M., & Weber, R. O. (2013). Case-based reasoning: A Textbook. *A Textbook*, 546. Retrieved from

Riesbeck, C. K., & Schank, R. C. (1989). *Inside case-based reasoning*: Psychology Press.

Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics applications, 8*(283-297), 37. Retrieved from

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and nonlinear time series analysis* (pp. 256-272): Springer.

Rui, Z., Metz, P. A., & Chen, G. (2012). An analysis of inaccuracy in pipeline construction cost estimation. *International Journal of Oil, Gas*
Coal Technology
5(1), 29-46. Retrieved from

Rui, Z., Metz, P. A., Reynolds, D. B., Chen, G., Zhou, X. J. O., & Journal, G. (2011). Regression models estimate pipeline construction costs. *109*(14), 120. Retrieved from

RunZhi, J., KyuMan, C., ChangTaek, H., & MyungJin, S. (2012). MRA-based revised CBR model for cost prediction in the early stage of construction projects. *Expert Systems with Applications, 39*(5), 5214-5222. doi: 10.1016/j.eswa.2011.11.018

Salamó, M., & Golobardes, E. (2003). Hybrid Deletion Policies for Case Base Maintenance. In *FLAIRS Conference* (pp. 150-154).

Salkind, N. J., & Rainwater, T. (2006). *Exploring research*: Pearson Prentice Hall Upper Saddle River, NJ.

Sanders, S. R., Maxwell, R. R., & Glagola, C. R. J. C. E. (1992). Preliminary estimating models for infrastructure projects. *34*(8), 7-13. Retrieved from

Sangyong, K., & Jae Heon, S. (2014). Combining case-based reasoning with genetic algorithm optimization for preliminary cost estimation in construction industry. *Canadian Journal of Civil Engineering, 41*(1), 65-73. doi: 10.1139/cjce-2013-0223

Saunders, M. N. (2011). *Research methods for business students, 5 ed*: Pearson Education India.

Saunders, M. N. (2012). *Research methods for business students, 5/e*: Pearson Education India.

Schall, O., Belyaev, A., & Seidel, H. (2005, 21-22 June 2005). Robust filtering of noisy scattered point data. In *Proceedings Eurographics/IEEE VGTC Symposium Point-Based Graphics, 2005.* (pp. 71-144).

Schank, R. C. (1982). *Dynamic memory revisited*: Cambridge University Press.

Schank, R. C. (1983). *Dynamic memory: A theory of reminding and learning in computers and people*: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*: Psychology Press.

Scheffer, J. (2002). Dealing with missing data. Retrieved from

Senouci, A., Ismail, A., & Eldin, N. (2016). Time Delay and Cost Overrun in Qatari Public Construction Projects. *Procedia Engineering, 164*, 368-375. doi: https://doi.org/10.1016/j.proeng.2016.11.632

She, Y., & Owen, A. B. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association*
106(494), 626-639. Retrieved from

Shehu, Z., Endut, I. R., & Akintoye, A. (2014). Factors contributing to project time and hence cost overrun in the Malaysian construction industry. *Journal of Financial Management of Property and Construction, 19*(1), 55-75. Retrieved from

Sheskin, D. J. (2010). Outlier. In N. J. Salkind (Ed.), *Encyclopedia of Research Design*: SAGE Publications, Inc.

Shin, K.-s., & Han, I. (1999). Case-based reasoning supported by genetic algorithms for corporate bond rating. *Expert Systems with Applications, 16*(2), 85-95. doi: https://doi.org/10.1016/S0957-4174(98)00063-3

Shin, Y. (2015). Application of boosting regression trees to preliminary cost estimation in building construction projects. *Computational intelligence*
neuroscience, 2015, 1. Retrieved from

Shiu, S. C., Yeung, D. S., Sun, C. H., & Wang, X. Z. J. C. I. (2001). Transferring Case Knowledge To Adaptation Knowledge: An Approach for Case - Base Maintenance. *17*(2), 295-314. Retrieved from

Shiu, S. C. K., Sun, C. H., Wang, X. Z., & Yeung, D. S. (2000). Maintaining case-based reasoning systems using fuzzy decision trees. In *European Workshop on Advances in Case-Based Reasoning* (pp. 285-296): Springer.

Siegel, A. F. J. B. (1982). Robust regression using repeated medians. *69*(1), 242-244. Retrieved from

Simpson, R. L. (1985). *A computer model of case-based reasoning in problem solving: an investigation in the domain of dispute mediation (analogy, machine learning, conceptual memory)* PhD(Doctoral Dissertation). Georgia Institute of Technology Atlanta.

Skitmore, M. (1985). *The influence of professional expertise in construction price forecasts*: The University of Salford.

Skitmore, M. (1987). Mobile cost estimating. *The Chartered Quantity Surveyor, 9*(9), 31-32. Retrieved from http://eprints.qut.edu.au/59689/

Skitmore, M. (1988). An expert system for the strategic planning of construction projects. In *Proceedings 1st Joint Conference on the Computerisation of Quantity Surveying* (pp. 172-199): Department of Construction and Surveying, Hong Kong Polytechnic, Hung Hom ….

Skitmore, M. (1990). *Which estimating technique?* Paper presented at 11th International Cost Engineering Congress and 6th Association Francais des Ingenieurs et Techniciens D'estimation de Planification de Projets Annual Meeting, Paris, France. Retrieved from http://eprints.qut.edu.au/9447/

Skitmore, M. (2001). Raftery curves for tender price forecasting: Empirical probabilities and pooling. *Financial Management of Property and Construction, 6*(3), 141-154. Retrieved from

Skitmore, M., & Marston, V. (2005). *Cost modelling*: Routledge.

Skitmore, R. M., & Ng, S. T. (2003). Forecast models for actual construction time and cost. *Building Environment*
38(8), 1075-1083. Retrieved from

Slonim, T., & Schneider, M. (2001). Design issues in fuzzy case-based reasoning. *Fuzzy Sets and Systems, 117*(2), 251-267. Retrieved from

Smith, A. E., & Mason, A. K. J. T. E. E. (1997). Cost estimation predictive modeling: Regression versus neural network. *42*(2), 137-161. Retrieved from

Smith, P. (2016). Global professional standards for project cost management. In *PROCEEDINGS OF THE 29TH IPMA WORLD CONGRESS WC2015*: ELSEVIER SCIENCE BV.

Smiti, A., & Elouedi, Z. (2010). Coid: Maintaining case method based on clustering, outliers and internal detection. In *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing 2010* (pp. 39-52): Springer.

Smiti, A., & Elouedi, Z. (2011a). Overview ofMaintenance for Case based Reasoning Systems. *International Journal of Computer Applications*
975, 8887. Retrieved from

Smiti, A., & Elouedi, Z. (2011b). WCOID: Maintaining case-based reasoning systems using Weighting, Clustering, Outliers and Internal cases Detection. In *2011 11th International Conference on Intelligent Systems Design and Applications* (pp. 356-361): IEEE.

Smiti, A., & Elouedi, Z. (2014). WCOID-DG: An approach for case base maintenance based on Weighting, Clustering, Outliers, Internal Detection and Dbsan-Gmeans. *Journal of computer system sciences*
80(1), 27-38. Retrieved from

Smiti, A., & Elouedi, Z. (2016). Fuzzy density based clustering method: Soft DBSCAN-GM. In *2016 IEEE 8th International Conference on Intelligent Systems (IS)* (pp. 443-448): IEEE.

Smiti, A., & Elouedi, Z. (2018a). SCBM: soft case base maintenance method based on competence model. *Journal of Computational Science, 25*, 221-227. doi: https://doi.org/10.1016/j.jocs.2017.09.013

Smiti, A., & Elouedi, Z. (2018b). SCBM: soft case base maintenance method based on competence model. *Journal of Computational Science*
25, 221-227. Retrieved from

Smiti, A., & Elouedi, Z. (2019). Dynamic maintenance case base using knowledge discovery techniques for case based reasoning systems. *Theoretical Computer Science*. doi: https://doi.org/10.1016/j.tcs.2019.06.026

Smiti, A., & Elouedi, Z. J. I. J. o. C. A. (2011c). Overview ofMaintenance for Case based Reasoning Systems. *975*, 8887. Retrieved from

Smyt, B., & McKenna, E. (1999). Footprint-Based Retrieval. In (pp. 343-357): Springer Berlin Heidelberg.

Smyth, B. (1998). Case-base maintenance. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 507-516): Springer.

Smyth, B., & Cunningham, P. (1996). The utility problem analysed. In *European Workshop on Advances in Case-Based Reasoning* (pp. 392-399): Springer.

Smyth, B., & Keane, M. T. (1995). Remembering to forget. In *Proceedings of the 14th international joint conference on Artificial intelligence* (pp. 377-382).

Smyth, B., & McKenna, E. (1998). Modelling the competence of case-bases. In *European Workshop on Advances in Case-Based Reasoning* (pp. 208-220): Springer.

Smyth, B., & McKenna, E. (1999). Building compact competent case-bases. In *International Conference on Case-Based Reasoning* (pp. 329-342): Springer.

Sonmez, R. (2008). Parametric range estimating of building costs using regression models and bootstrap. *Journal of construction Engineering Management* *134*(12), 1011-1016. Retrieved from

Sonmez, R. (2011). Range estimation of construction costs using neural networks with bootstrap prediction intervals. *Expert Systems with Applications, 38*(8), 9913-9917. doi: https://doi.org/10.1016/j.eswa.2011.02.042

Sonmez, R. J. C. J. o. C. E. (2004). Conceptual cost estimation of building projects with regression analysis and neural networks. *31*(4), 677-683. Retrieved from

Stackpole, C. (2010). *A user's guide to the PMBOK guide*. Hoboken, N.J: Wiley.

(2019). *ICMS: Global Consistency in Presenting Construction and Other Life Cycle Costs*. Retrieved from https://icmscblog.files.wordpress.com/2019/10/international-construction-measurement-standards-2nd-edition.pdf

Swei, O., Gregory, J., & Kirchain, R. J. T. R. P. B. M. (2017). Construction cost estimation: A parametric approach for better estimates of expected cost and variation. *101*, 295-305. Retrieved from

Tahir, M., Haron, N., Alias, A., Harun, A., Muhammad, I., & Baba, D. (2018). Improving Cost and Time Control in Construction Using Building Information Model (BIM): A Review. *Pertanika Journal of Science Technology, 26*(1). Retrieved from

Taleb, I., Kassabi, H. T. E., Serhani, M. A., Dssouli, R., & Bouhaddioui, C. (2016, 18-21 July 2016). Big Data Quality: A Quality Dimensions Evaluation. In *2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld)* (pp. 759-765).

Tatiya, A., Zhao, D., Syal, M., Berghorn, G. H., & LaMore, R. (2018). Cost prediction model for building deconstruction in urban areas. *Journal of Cleaner Production, 195*, 1572-1580. doi: https://doi.org/10.1016/j.jclepro.2017.08.084

Tayi, G. K., & Ballou, D. P. (1998). Examining data quality. *Communications of the ACM*
41(2), 54-57. Retrieved from

Themsen, T. N. (2019). The processes of public megaproject cost estimation: The inaccuracy of reference class forecasting. *35*(4), 337-352. doi: 10.1111/faam.12210

Thomas Ng, S., Mak, M. M. Y., Martin Skitmore, R., Lam, K. C., & Varnam, M. (2001). The predictive ability of Bromilow's time–cost model. *Construction Management and Economics, 19*(2), 165-173. doi: 10.1080/01446190150505090

Tomek, I. (1976). An experiment with the edited nearest-neighbor rule. *Transactions on systems, Man, Cybernetics*
6(6), 448-452. Retrieved from

Torrent-Fontbona, F., Massana, J., & López, B. (2019). Case-base maintenance of a personalised and adaptive CBR bolus insulin recommender system for type 1 diabetes. *Expert Systems with Applications*
121, 338-346. Retrieved from

Touran, A., & Lopez, R. (2006). Modeling cost escalation in large infrastructure projects. *Journal of construction engineering management*
132(8), 853-860. Retrieved from

Touran, A. J. J. o. C. E., & Management. (1993). Probabilistic cost estimating with subjective correlations. *119*(1), 58-71. Retrieved from

Towey, D. (2013). *Cost management of construction projects*: John Wiley & Sons.

Trost, S. M., & Oberlender, G. D. (2003a). Predicting accuracy of early cost estimates using factor analysis and multivariate regression. *Journal of construction Engineering Management*
129(2), 198-204. Retrieved from

Trost, S. M., & Oberlender, G. D. (2003b). Predicting Accuracy of Early Cost Estimates Using Factor Analysis and Multivariate Regression. *129*(2), 198-204. doi: doi:10.1061/(ASCE)0733-9364(2003)129:2(198)

Trost, S. M., & Oberlender, G. D. (2003c). Predicting accuracy of early cost estimates using factor analysis and multivariate regression. *Journal of construction Engineering Management, 129*(2), 198-204. Retrieved from

Wang, H.-J., Chiou, C.-W., & Juan, Y.-K. (2008). Decision support model based on case-based reasoning approach for estimating the restoration budget of historical buildings. *Expert Systems with Applications, 35*(4), 1601-1610. doi: https://doi.org/10.1016/j.eswa.2007.08.095

Wang, W.-c., Chau, K.-w., Xu, D.-m., & Chen, X.-Y. J. W. R. M. (2015). Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. *29*(8), 2655-2675. Retrieved from

Wang, X., Chen, H., Cai, W., Shen, D., & Huang, H. (2017). Regularized modal regression with applications in cognitive impairment prediction. In *Advances in neural information processing systems* (pp. 1448-1458).

Watson, I. (1999). Case-based reasoning is a methodology not a technology. *Knowledge-Based Systems, 12*(5–6), 303-308. doi: http://dx.doi.org/10.1016/S0950-7051(99)00020-9

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics computing 4*(2), 65-85. Retrieved from

Wilcox, R. R. (1996). A review of some recent developments in robust regression. *British Journal of Mathematical Statistical Psychology, 49*(2), 253-274. Retrieved from

Wilmot, C. G., Mei, B. J. J. o. c. e., & management. (2005). Neural network modeling of highway construction costs. *131*(7), 765-771. Retrieved from

Wilson, D. C. (2001). *Case-base maintenance: the husbandry of experience*: Indiana University.

Wilson, D. C., & Leake, D. B. (2001a). Maintaining Case-Based Reasoners: Dimensions and Directions. *17*(2), 196-213. doi: 10.1111/0824-7935.00140

Wilson, D. C., & Leake, D. B. J. C. I. (2001b). Maintaining Case - Based Reasoners: Dimensions and Directions. *17*(2), 196-213. Retrieved from

Wilson, D. L. (1972). Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, Cybernetics* (3), 408-421. Retrieved from

Wilson, D. R., & Martinez, T. R. (2000). Reduction techniques for instance-based learning algorithms. *Machine learning, 38*(3), 257-286. Retrieved from

Wing Chau, K. (1995). The validity of the triangular distribution assumption in Monte Carlo simulation of construction costs: empirical evidence from Hong Kong. *Construction Management and Economics, 13*(1), 15-21. Retrieved from

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample Size Requirements for Structural Equation Models. *Educational and Psychological Measurement, 73*(6), 913-934. doi: 10.1177/0013164413495237

Wu, G., Wang, H., & Chang, R. J. A. i. C. E. (2018). A Decision Model Assessing the Owner and Contractor's Conflict Behaviors in Construction Projects. *2018*. Retrieved from

Xia, B., Chan, A., Molenaar, K., & Skitmore, M. (2012). Determining the appropriate proportion of owner-provided design in design-build contracts - a content analysis approach. *Journal of Construction Engineering and Management, 138*(9), 1017-1022. doi: 10.1061/(asce)co.1943-7862.0000522

Xiao, X., Wang, F., Li, H., & Skitmore, M. (2018). Modelling the stochastic dependence underlying construction cost and duration. *Journal of Civil Engineering Management*
24(6), 444-456. Retrieved from

Yaman, H., & Tas, E. I. A. (2007). A building cost estimation model based on functional elements. *ITU A, 4*(1), 73-87. Retrieved from

Yang, Q., & Wu, J. (2000). Keep it simple: A case-base maintenance policy based on clustering and information theory. In *Conference of the Canadian Society for Computational Studies of Intelligence* (pp. 102-114): Springer.

Yang, Q., & Zhu, J. (2001). A Case - Addition Policy for Case - Base Maintenance. *Computational Intelligence, 17*(2), 250-262. Retrieved from

Yao, W., & Li, L. (2014). A New Regression Model: Modal Linear Regression. *Scandinavian Journal of Statistics, 41*(3), 656-671. doi: https://doi.org/10.1111/sjos.12054

Yau, N. J., & Yang, J. B. (1998). Case - based reasoning in construction management. *Computer - Aided Civil and Infrastructure Engineering, 13*(2), 143-150. Retrieved from

Yeong, C. M. (1994). *Time and cost performance of building contracts in Australia and Malaysia*. University of South Australia.

Yeung, D., & Skitmore, M. (2012). A method for systematically pooling data in very early stage construction price forecasting. *Construction Management and Economics, 30*(11), 929-939. Retrieved from

Yeung, D. K., & Skitmore, M. (2005). Data Pooling for Early-Stage Price Forecasts. In *21st Annual Association of Researchers in Construction Management (ARCOM) Conference"(F. Khosrowshahi, ed.)* (Vol. 1, pp. 269-276).

Yildiz, A. E., Dikmen, I., Birgonul, M. T., Ercoskun, K., & Alten, S. (2014). A knowledge-based risk mapping tool for cost estimation of international construction projects. *Automation in Construction, 43*, 144-155. doi: https://doi.org/10.1016/j.autcon.2014.03.010

Yohai, V. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*
15(2), 642-656. Retrieved from

You, J. (1999). A Monte Carlo comparison of several high breakdown and efficient estimators. *Computational statistics data analysis*
30(2), 205-219. Retrieved from

Yu, C., Yao, W. J. C. i. S.-S., & Computation. (2017). Robust linear regression: A review and comparison. *46*(8), 6261-6282. Retrieved from

Yu, W.-d., & Skibniewski, M. (2009). Integrating neurofuzzy system with conceptual cost estimation to discover cost-related knowledge from residential construction projects. *Journal of Computing in Civil Engineering*
24(1), 35-44. Retrieved from

Zhu, J., & Yang, Q. (1999). Remembering to add: competence-preserving case-addition policies for case-base maintenance. In *IJCAI* (Vol. 99, pp. 234-241).

Žujo, V., Car-Pušić, D., Žileska-Pančovska, V., & Ćećez, M. (2017). Time and cost interdependence in water supply system construction projects. *Technological and Economic Development of Economy, 23*(6), 895-914. doi: 10.3846/20294913.2015.1071292

Appendices

APPENDIX A

Conversion of the index from 2006 to 2014 to 2015

Year	Transfer year	Region 1	Region 2	Region 3	Region 4	Region 5	Region 6
2006	2007	1.0410	1.0460	1.0630	1.0380	1.0600	1.0560
	2008	1.1638	1.1726	1.1916	1.1646	1.2052	1.1964
	2009	1.0975	1.1116	1.1273	1.1099	1.1691	1.1857
	2010	1.1414	1.1794	1.1825	1.1576	1.2006	1.2485
	2011	1.2521	1.3044	1.3019	1.2502	1.2943	1.3472
	2012	1.2396	1.2875	1.3202	1.2740	1.3215	1.3930
	2013	1.2061	1.2849	1.3175	1.2982	1.3281	1.4250
	2014	1.1880	1.2887	1.3188	1.3242	1.3334	1.4421
	2015	1.1215	1.2230	1.2727	1.3030	1.3000	1.4190
2007	2008	1.1180	1.1210	1.1210	1.1220	1.1370	1.1330
	2009	1.0543	1.0627	1.0605	1.0693	1.1029	1.1228
	2010	1.0964	1.1275	1.1124	1.1152	1.1327	1.1823
	2011	1.2028	1.2471	1.2248	1.2045	1.2210	1.2757
	2012	1.1908	1.2308	1.2419	1.2273	1.2467	1.3191
	2013	1.1586	1.2284	1.2394	1.2507	1.2529	1.3494
	2014	1.1412	1.2321	1.2407	1.2757	1.2579	1.3656
	2015	1.0773	1.1692	1.1973	1.2553	1.2265	1.3438
2008	2009	0.9430	0.9480	0.9460	0.9530	0.9700	0.9910
	2010	0.9807	1.0058	0.9924	0.9940	0.9962	1.0435
	2011	1.0758	1.1124	1.0926	1.0735	1.0739	1.1260
	2012	1.0651	1.0980	1.1079	1.0939	1.0964	1.1642
	2013	1.0363	1.0958	1.1057	1.1147	1.1019	1.1910
	2014	1.0208	1.0991	1.1068	1.1370	1.1063	1.2053
	2015	0.9636	1.0430	1.0680	1.1188	1.0787	1.1860
2009	2010	1.0400	1.0610	1.0490	1.0430	1.0270	1.0530
	2011	1.1409	1.1735	1.1549	1.1264	1.1071	1.1362
	2012	1.1295	1.1582	1.1711	1.1478	1.1304	1.1748
	2013	1.0990	1.1559	1.1688	1.1697	1.1360	1.2018
	2014	1.0825	1.1594	1.1699	1.1930	1.1406	1.2163

	2015	1.0219	1.1002	1.1290	1.1740	1.1120	1.1968
2010	2011	1.0970	1.1060	1.1010	1.0800	1.0780	1.0790
	2012	1.0860	1.0916	1.1164	1.1005	1.1006	1.1157
	2013	1.0567	1.0894	1.1142	1.1214	1.1061	1.1413
	2014	1.0409	1.0927	1.1153	1.1439	1.1106	1.1550
	2015	0.9826	1.0370	1.0763	1.1256	1.0828	1.1366
2011	2012	0.9900	0.9870	1.0140	1.0190	1.0210	1.0340
	2013	0.9633	0.9850	1.0120	1.0384	1.0261	1.0578
	2014	0.9488	0.9880	1.0130	1.0591	1.0302	1.0705
	2015	0.8957	0.9376	0.9775	1.0422	1.0045	1.0533
2012	2013	0.9730	0.9980	0.9980	1.0190	1.0050	1.0230
	2014	0.9584	1.0010	0.9990	1.0394	1.0090	1.0353
	2015	0.9047	0.9499	0.9640	1.0227	0.9838	1.0187
2013	2014	0.9850	1.0030	1.0010	1.0200	1.0040	1.0120
	2015	92.9840	0.9518	0.9660	1.0037	0.9789	0.9958
2014	2015	0.9440	0.9490	0.9650	0.9840	0.9750	0.9840

APPENDIX B

Table B-1 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2, CBR-W3S2 (100-size sample and K=5)

MAPE

	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	35.18%	27.37%	48.77%	28.15%	26.61%	67.02%
Round2	24.58%	25.07%	39.27%	22.32%	22.28%	34.09%
Round3	29.54%	29.87%	36.85%	22.56%	20.83%	27.07%
Round4	38.54%	30.27%	48.77%	25.80%	22.64%	33.68%
Round5	34.13%	41.29%	36.40%	27.09%	24.30%	43.30%
Round6	19.55%	18.28%	25.21%	17.99%	15.94%	26.01%
Round7	25.06%	24.11%	23.19%	26.51%	19.04%	36.72%
Round8	15.81%	14.68%	31.27%	17.10%	13.10%	38.60%
Round9	35.77%	34.77%	38.67%	35.22%	30.43%	48.47%
Round10	22.44%	23.45%	30.15%	18.97%	14.57%	25.65%
Average	28.06%	26.92%	35.85%	24.17%	20.97%	38.06%

RMSE

	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	35.22%	29.55%	46.21%	31.20%	29.17%	58.65%
Round2	27.22%	27.45%	46.99%	26.57%	26.27%	41.30%
Round3	34.19%	34.20%	44.30%	26.55%	27.16%	30.93%
Round4	38.13%	34.53%	45.20%	31.86%	27.94%	34.76%
Round5	34.61%	41.84%	37.08%	31.28%	29.90%	43.48%
Round6	23.96%	22.08%	32.80%	23.14%	20.97%	32.30%
Round7	26.56%	25.79%	25.02%	28.94%	22.07%	39.68%
Round8	19.77%	18.26%	52.74%	25.35%	17.95%	45.97%
Round9	39.89%	40.44%	45.60%	41.90%	36.97%	52.61%
Round10	29.07%	30.55%	48.09%	23.74%	19.38%	33.20%
Average	30.86%	30.47%	42.40%	29.05%	25.78%	41.29%

Table B-2 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2, CBR-W3S2 (200-size sample and K=5)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	21.03%	21.32%	28.93%	18.03%	18.35%	21.73%
Round2	24.66%	24.13%	34.74%	25.17%	22.10%	32.87%
Round3	24.62%	21.99%	34.86%	21.91%	20.91%	31.50%
Round4	22.67%	21.69%	27.56%	19.08%	19.11%	22.95%
Round5	25.79%	26.86%	29.48%	22.99%	19.58%	27.02%
Round6	27.33%	27.12%	34.46%	24.44%	22.94%	31.24%
Round7	21.85%	21.15%	30.35%	19.62%	16.89%	22.80%
Round8	21.83%	20.35%	29.36%	17.37%	17.91%	22.38%
Round9	28.08%	26.68%	33.79%	21.00%	22.02%	38.50%
Round10	25.59%	19.78%	30.39%	22.66%	19.31%	29.81%
Average	24.34%	23.11%	31.39%	21.23%	19.91%	28.08%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	24.35%	24.53%	40.99%	21.52%	21.91%	25.59%
Round2	30.14%	29.32%	58.45%	31.81%	28.36%	40.99%
Round3	27.11%	24.16%	46.38%	26.50%	25.05%	36.85%
Round4	24.81%	23.65%	32.54%	22.23%	22.18%	27.38%
Round5	32.36%	28.13%	39.88%	25.61%	22.23%	31.60%
Round6	30.82%	30.30%	41.05%	29.86%	26.87%	34.43%
Round7	27.21%	25.43%	41.91%	23.75%	21.18%	28.23%
Round8	24.45%	24.15%	36.86%	22.75%	23.14%	29.00%
Round9	30.83%	28.23%	38.63%	25.01%	27.22%	41.94%
Round10	37.61%	23.48%	48.28%	26.62%	22.43%	34.14%
Average	28.97%	26.14%	42.50%	25.57%	24.06%	33.02%

Table B-3 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2，CBR-W3S2 (400-size sample and K=5)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	24.33%	22.50%	30.27%	19.98%	19.22%	24.69%
Round2	25.53%	25.93%	31.23%	23.28%	21.41%	27.51%
Round3	23.27%	21.85%	30.85%	20.59%	17.94%	20.14%
Round4	25.29%	25.15%	28.31%	20.88%	20.83%	23.56%
Round5	23.09%	22.37%	30.52%	18.62%	19.80%	23.19%
Round6	19.97%	20.60%	27.94%	18.45%	18.27%	24.59%
Round7	21.21%	21.02%	30.00%	15.69%	16.29%	23.29%
Round8	19.50%	19.28%	31.64%	17.67%	18.09%	25.41%
Round9	22.42%	22.00%	27.77%	18.66%	19.47%	24.56%
Round10	23.39%	22.07%	28.45%	19.25%	18.71%	23.71%
Average	22.80%	22.28%	29.70%	19.31%	19.00%	24.06%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	25.79%	24.93%	44.17%	23.26%	22.17%	28.55%
Round2	28.65%	28.43%	36.24%	27.16%	24.67%	31.37%
Round3	27.06%	25.50%	46.73%	25.56%	22.46%	26.54%
Round4	27.98%	27.48%	36.67%	26.45%	26.94%	29.37%
Round5	25.70%	25.30%	40.92%	22.74%	24.21%	27.44%
Round6	23.44%	24.14%	38.41%	22.85%	21.92%	28.67%
Round7	24.68%	24.72%	37.59%	19.57%	19.72%	27.11%
Round8	23.33%	22.96%	47.50%	22.91%	22.50%	33.23%
Round9	25.41%	24.98%	43.87%	22.08%	23.65%	28.96%
Round10	25.34%	24.64%	38.81%	22.23%	21.82%	26.76%
Average	25.74%	25.31%	41.09%	23.48%	23.01%	28.80%

Table B-4 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2, CBR-W3S2 (600-size sample and K=5)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	25.88%	25.45%	34.08%	18.98%	19.63%	22.95%
Round2	24.13%	23.82%	33.24%	17.92%	18.69%	22.39%
Round3	25.47%	22.86%	33.47%	17.85%	17.78%	21.63%
Round4	21.43%	21.91%	30.04%	19.12%	19.58%	22.87%
Round5	22.76%	22.88%	30.30%	19.53%	19.34%	26.63%
Round6	23.40%	21.55%	26.90%	18.20%	15.75%	20.92%
Round7	23.16%	23.81%	30.12%	19.18%	19.49%	25.97%
Round8	22.25%	21.47%	27.53%	17.63%	18.14%	25.81%
Round9	22.03%	23.12%	29.25%	19.93%	21.31%	24.49%
Round10	21.19%	21.09%	27.08%	15.54%	15.57%	18.73%
Average	23.17%	22.80%	30.20%	18.39%	18.53%	23.24%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	28.25%	27.95%	52.79%	22.79%	23.61%	27.32%
Round2	26.56%	25.95%	40.58%	22.35%	22.61%	27.65%
Round3	28.76%	26.05%	45.94%	22.44%	22.27%	26.07%
Round4	25.20%	25.28%	46.50%	24.17%	24.26%	29.47%
Round5	26.10%	26.07%	45.13%	25.15%	24.42%	32.98%
Round6	25.87%	24.31%	38.93%	24.47%	19.92%	28.15%
Round7	26.56%	26.67%	45.60%	24.15%	24.73%	30.73%
Round8	25.61%	25.09%	46.76%	22.82%	22.65%	30.99%
Round9	25.47%	26.45%	39.26%	24.43%	25.40%	29.86%
Round10	24.12%	24.14%	42.73%	19.99%	19.94%	26.48%
Average	26.25%	25.80%	44.42%	23.28%	22.98%	28.97%

Table B-5 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2, CBR-W3S2 (800-size sample and K=5)

	ARE					
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	23.57%	21.31%	32.46%	18.54%	18.98%	21.95%
Round2	24.75%	24.56%	29.01%	19.22%	19.85%	20.55%
Round3	25.24%	24.19%	33.29%	20.75%	18.90%	25.78%
Round4	22.49%	21.88%	28.63%	17.27%	18.01%	19.80%
Round5	22.83%	22.21%	28.97%	18.82%	19.24%	23.36%
Round6	25.58%	25.43%	28.88%	19.03%	19.42%	21.48%
Round7	22.58%	20.71%	26.73%	16.84%	16.46%	21.39%
Round8	23.05%	23.53%	29.14%	16.96%	16.77%	19.44%
Round9	25.59%	24.99%	33.11%	19.99%	20.23%	24.15%
Round10	24.77%	24.59%	27.95%	18.92%	18.23%	23.08%
Average	24.04%	23.34%	29.81%	18.63%	18.61%	22.10%
	RMSE					
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	26.55%	24.20%	45.36%	23.74%	23.24%	27.14%
Round2	27.37%	27.58%	39.39%	23.29%	23.68%	25.28%
Round3	28.41%	27.41%	41.45%	26.00%	23.32%	30.94%
Round4	25.15%	25.14%	38.92%	22.70%	23.16%	26.66%
Round5	25.42%	25.00%	43.23%	22.71%	23.08%	28.81%
Round6	27.92%	28.11%	38.36%	23.33%	23.42%	26.14%
Round7	25.24%	23.57%	39.42%	21.49%	21.05%	26.78%
Round8	26.07%	26.33%	43.57%	20.92%	21.08%	26.35%
Round9	28.62%	28.42%	44.55%	23.87%	24.53%	28.11%
Round10	27.81%	27.57%	41.73%	24.30%	23.05%	28.60%
Average	26.86%	26.33%	41.60%	23.23%	22.96%	27.48%

Table B-6 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2，CBR-W3S2 (1000-size sample and K=5)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	23.74%	22.96%	26.36%	19.17%	18.42%	23.50%
Round2	25.90%	24.40%	28.65%	18.49%	18.27%	19.62%
Round3	23.65%	24.26%	26.20%	19.09%	18.25%	20.67%
Round4	24.67%	23.73%	30.68%	18.64%	18.31%	23.17%
Round5	26.07%	25.08%	31.70%	21.56%	20.74%	24.08%
Round6	24.18%	23.51%	31.13%	18.27%	17.13%	21.51%
Round7	23.47%	23.82%	27.84%	17.70%	17.53%	20.05%
Round8	24.97%	24.42%	31.12%	20.04%	20.86%	23.16%
Round9	23.13%	22.32%	26.65%	18.71%	18.64%	20.73%
Round10	23.21%	22.79%	28.52%	18.57%	18.49%	20.72%
Average	24.30%	23.73%	28.88%	19.02%	18.66%	21.72%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	26.51%	25.64%	40.59%	24.77%	23.07%	29.92%
Round2	28.03%	26.71%	40.68%	22.43%	21.93%	23.50%
Round3	26.67%	27.10%	36.65%	24.10%	22.93%	26.10%
Round4	27.79%	26.89%	41.50%	24.13%	23.55%	28.61%
Round5	28.55%	27.82%	43.89%	26.41%	25.51%	28.98%
Round6	26.69%	26.19%	43.31%	23.30%	21.46%	27.43%
Round7	26.09%	26.41%	39.52%	22.19%	22.09%	24.61%
Round8	27.67%	27.24%	44.00%	24.55%	24.58%	27.59%
Round9	25.99%	25.62%	38.84%	23.69%	22.98%	25.77%
Round10	26.30%	25.71%	39.36%	22.73%	22.32%	25.25%
Average	27.03%	26.53%	40.83%	23.83%	23.04%	26.78%

Table B-7 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2，CBR-W3S2 (50-size sample and K=3)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	30.95%	23.93%	34.29%	38.82%	20.79%	23.06%
Round2	48.37%	34.58%	56.84%	34.77%	32.35%	69.38%
Round3	42.07%	42.20%	38.86%	33.62%	28.61%	41.64%
Round4	52.84%	18.11%	45.89%	32.64%	17.79%	36.41%
Round5	32.85%	43.90%	38.52%	36.09%	27.47%	51.48%
Round6	39.21%	27.01%	38.25%	31.58%	14.64%	42.75%
Round7	28.76%	24.79%	33.88%	23.36%	27.15%	37.57%
Round8	39.56%	26.40%	46.52%	30.54%	24.37%	34.52%
Round9	25.65%	24.19%	39.76%	39.39%	30.58%	45.08%
Round10	17.85%	18.70%	20.09%	24.54%	25.28%	10.10%
Average	35.81%	28.38%	39.29%	32.53%	24.90%	39.20%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	31.89%	24.89%	32.93%	36.91%	22.75%	27.20%
Round2	67.75%	38.38%	67.38%	37.96%	34.42%	65.96%
Round3	42.95%	41.52%	40.89%	33.57%	30.93%	40.28%
Round4	46.15%	20.67%	41.51%	32.22%	20.11%	45.08%
Round5	32.37%	42.07%	48.34%	39.32%	28.20%	51.45%
Round6	41.00%	27.51%	45.54%	40.54%	19.54%	51.88%
Round7	29.15%	25.69%	34.30%	25.20%	29.02%	37.61%
Round8	60.90%	35.04%	73.04%	46.81%	30.18%	50.08%
Round9	31.29%	28.40%	45.18%	43.83%	33.03%	50.02%
Round10	20.72%	21.29%	21.98%	31.43%	31.38%	12.12%
Average	40.42%	30.55%	45.11%	36.78%	27.96%	43.17%

Table B-8 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2, CBR-W3S2 (100-size sample and K=3)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	33.36%	29.57%	48.40%	33.20%	24.77%	56.82%
Round2	23.95%	22.63%	31.22%	19.17%	22.39%	38.39%
Round3	27.84%	29.57%	36.67%	22.55%	21.89%	26.81%
Round4	38.63%	29.68%	50.40%	27.69%	18.28%	27.16%
Round5	31.05%	35.62%	42.94%	25.78%	24.74%	32.72%
Round6	15.20%	16.86%	20.73%	21.55%	15.14%	25.30%
Round7	19.75%	15.75%	28.87%	22.64%	19.35%	31.49%
Round8	15.17%	16.33%	32.39%	16.22%	15.25%	35.22%
Round9	32.34%	31.10%	38.77%	32.51%	31.16%	52.15%
Round10	23.35%	14.17%	23.87%	15.75%	13.52%	22.60%
Average	26.06%	24.13%	35.43%	23.71%	20.65%	34.87%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	34.58%	32.21%	47.15%	36.46%	29.27%	52.88%
Round2	27.84%	26.05%	42.10%	22.18%	26.82%	43.87%
Round3	33.16%	34.01%	48.91%	28.59%	29.19%	31.61%
Round4	39.57%	34.73%	47.22%	32.03%	24.91%	30.14%
Round5	32.21%	37.18%	41.80%	29.97%	28.59%	34.32%
Round6	19.13%	19.59%	28.67%	26.55%	21.26%	31.32%
Round7	22.26%	19.76%	30.51%	27.43%	22.66%	36.28%
Round8	20.03%	20.82%	56.00%	22.13%	20.09%	43.98%
Round9	37.67%	37.18%	47.00%	39.77%	39.23%	53.76%
Round10	29.12%	18.57%	43.48%	21.69%	18.20%	28.20%
Average	29.56%	28.01%	43.28%	28.68%	26.02%	38.63%

Table B-9 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2，CBR-W3S2 (200-size sample and K=3)

ARE

	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	22.46%	21.64%	28.48%	18.80%	18.80%	20.81%
Round2	20.66%	20.69%	34.41%	25.35%	21.19%	34.07%
Round3	26.99%	27.12%	31.25%	23.83%	20.64%	31.79%
Round4	23.88%	21.62%	28.79%	18.83%	18.92%	23.92%
Round5	29.95%	28.17%	29.14%	22.87%	21.54%	28.14%
Round6	28.88%	29.90%	35.47%	23.33%	21.90%	30.36%
Round7	23.76%	23.23%	30.39%	18.57%	18.60%	24.68%
Round8	20.97%	23.15%	29.58%	15.89%	18.20%	23.54%
Round9	28.18%	26.03%	35.23%	23.76%	23.02%	38.98%
Round10	24.98%	21.04%	31.52%	22.85%	20.95%	27.78%
Average	25.07%	24.26%	31.43%	21.41%	20.38%	28.41%

RMSE

	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	25.54%	24.89%	43.56%	23.47%	24.20%	26.53%
Round2	25.55%	25.73%	59.04%	33.09%	26.77%	41.70%
Round3	30.19%	29.29%	44.06%	28.25%	25.93%	37.21%
Round4	26.29%	23.77%	32.83%	23.61%	23.35%	28.28%
Round5	36.02%	29.57%	38.75%	26.46%	23.42%	31.50%
Round6	34.09%	33.99%	41.93%	28.42%	26.72%	35.70%
Round7	28.87%	27.71%	46.58%	23.33%	23.54%	29.16%
Round8	24.29%	26.10%	39.67%	20.65%	22.89%	30.44%
Round9	31.44%	29.36%	39.98%	28.96%	28.37%	42.38%
Round10	37.32%	24.68%	53.34%	26.08%	23.88%	33.29%
Average	29.96%	27.51%	43.97%	26.23%	24.91%	33.62%

Table B-10 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2 and CBR-W2S2，CBR-W3S2 (400-size sample and K=3)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	24.72%	24.92%	32.30%	20.01%	19.69%	19.84%
Round2	25.36%	24.37%	30.86%	21.80%	21.35%	24.89%
Round3	23.70%	22.40%	32.72%	21.73%	21.36%	19.59%
Round4	28.13%	28.41%	27.70%	21.19%	22.89%	22.78%
Round5	23.32%	22.87%	32.72%	19.18%	19.83%	22.32%
Round6	21.01%	21.71%	29.24%	17.99%	19.73%	24.07%
Round7	24.17%	24.11%	29.98%	16.79%	15.99%	21.83%
Round8	21.23%	20.58%	29.27%	19.25%	18.85%	25.27%
Round9	24.12%	23.25%	28.10%	19.90%	20.42%	22.14%
Round10	24.41%	21.29%	28.91%	18.49%	17.22%	24.21%
Average	24.02%	23.39%	30.18%	19.64%	19.73%	22.69%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	27.26%	27.44%	46.70%	23.43%	23.15%	24.08%
Round2	29.38%	27.99%	36.42%	25.66%	25.48%	28.89%
Round3	27.13%	26.11%	46.59%	27.59%	25.76%	26.73%
Round4	30.55%	30.36%	37.14%	26.94%	29.06%	28.91%
Round5	26.48%	26.27%	42.97%	23.00%	23.41%	26.64%
Round6	24.58%	25.18%	40.09%	21.73%	23.40%	29.05%
Round7	27.08%	27.30%	37.18%	21.13%	19.27%	26.42%
Round8	25.05%	24.36%	47.24%	24.18%	22.82%	32.07%
Round9	27.99%	26.31%	45.04%	23.50%	24.23%	27.50%
Round10	26.26%	23.78%	40.58%	22.51%	21.41%	27.78%
Average	27.18%	26.51%	41.99%	23.97%	23.80%	27.81%

Table B-8.11 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2, CBR-W3S2 (600-size sample and K=3)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	27.21%	27.56%	36.06%	19.55%	20.08%	24.12%
Round2	26.08%	24.27%	32.34%	18.26%	19.32%	23.04%
Round3	28.84%	26.51%	35.85%	17.45%	16.89%	21.38%
Round4	23.37%	22.82%	31.13%	18.97%	19.97%	21.40%
Round5	24.43%	24.88%	30.05%	19.11%	20.01%	25.66%
Round6	25.94%	23.95%	27.51%	18.18%	16.79%	19.95%
Round7	24.24%	24.89%	29.35%	19.13%	19.97%	24.38%
Round8	24.45%	22.79%	29.73%	16.70%	18.32%	24.35%
Round9	23.39%	24.41%	28.70%	19.92%	21.13%	23.12%
Round10	22.24%	21.62%	30.00%	15.20%	15.14%	19.04%
Average	25.02%	24.37%	31.07%	18.25%	18.76%	22.64%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	29.92%	30.35%	54.12%	23.66%	24.47%	30.16%
Round2	28.46%	26.37%	41.74%	22.81%	23.73%	28.35%
Round3	31.43%	29.15%	50.48%	22.03%	21.93%	27.24%
Round4	27.34%	27.08%	49.56%	24.17%	24.74%	27.84%
Round5	27.74%	27.68%	45.73%	24.88%	25.09%	32.23%
Round6	28.03%	27.13%	39.53%	25.24%	22.26%	27.66%
Round7	27.72%	27.85%	46.49%	24.31%	25.63%	29.92%
Round8	27.33%	25.77%	49.20%	22.66%	22.86%	30.05%
Round9	27.33%	28.17%	39.31%	25.34%	25.74%	29.33%
Round10	25.19%	24.92%	46.56%	19.97%	20.32%	27.36%
Average	28.05%	27.45%	46.27%	23.51%	23.68%	29.01%

Table B-8.22 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2, CBR-W3S2 (800-size sample and K=3)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	25.31%	22.17%	31.26%	18.09%	20.08%	21.91%
Round2	25.36%	25.71%	30.09%	19.77%	20.67%	21.58%
Round3	24.16%	23.81%	33.90%	19.69%	19.74%	23.41%
Round4	23.58%	23.53%	30.37%	18.39%	19.46%	21.43%
Round5	24.83%	23.94%	29.37%	17.84%	18.68%	22.13%
Round6	25.74%	25.78%	29.16%	19.84%	19.36%	22.28%
Round7	22.31%	22.39%	28.75%	17.15%	15.82%	20.23%
Round8	23.49%	24.01%	29.17%	16.00%	16.45%	19.49%
Round9	27.72%	25.85%	34.91%	20.38%	20.52%	23.56%
Round10	26.00%	25.18%	30.27%	19.46%	18.86%	22.43%
Average	24.85%	24.24%	30.73%	18.66%	18.96%	21.85%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	28.25%	25.53%	45.84%	23.90%	24.10%	27.73%
Round2	27.79%	28.29%	40.18%	24.13%	25.12%	26.25%
Round3	27.88%	27.35%	43.31%	24.65%	23.81%	28.68%
Round4	27.03%	26.89%	43.08%	24.20%	25.04%	28.43%
Round5	27.63%	26.95%	44.22%	22.45%	23.70%	28.42%
Round6	28.18%	28.20%	37.91%	24.26%	24.08%	27.15%
Round7	25.21%	25.40%	42.70%	21.96%	20.53%	26.27%
Round8	26.68%	27.07%	44.90%	20.04%	20.59%	25.53%
Round9	30.50%	28.99%	45.99%	24.99%	25.21%	27.95%
Round10	28.77%	28.32%	45.08%	25.51%	23.46%	28.40%
Average	27.79%	27.30%	43.32%	23.61%	23.56%	27.48%

Table B-8.33 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2, CBR-W3S2 (1000-size sample and K=3)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	24.64%	23.66%	28.10%	20.02%	17.89%	22.10%
Round2	26.47%	26.99%	30.53%	19.49%	18.56%	20.09%
Round3	24.45%	23.74%	27.52%	19.90%	18.51%	19.91%
Round4	26.88%	26.62%	32.25%	20.19%	19.72%	23.58%
Round5	27.64%	26.50%	31.55%	21.38%	21.06%	24.77%
Round6	26.15%	26.12%	32.12%	18.23%	18.65%	21.17%
Round7	23.64%	23.55%	29.24%	19.04%	18.18%	19.74%
Round8	25.16%	24.61%	31.07%	20.26%	20.90%	24.60%
Round9	24.27%	24.07%	28.02%	19.02%	19.24%	21.13%
Round10	23.77%	23.37%	30.11%	17.61%	18.18%	20.42%
Average	25.31%	24.92%	30.05%	19.51%	19.09%	21.75%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	27.73%	26.62%	42.45%	25.86%	23.00%	28.42%
Round2	28.45%	29.19%	43.54%	23.49%	22.17%	24.78%
Round3	27.49%	26.92%	40.14%	25.48%	23.21%	25.45%
Round4	29.25%	29.11%	43.15%	25.66%	25.48%	29.31%
Round5	30.23%	29.09%	45.03%	26.21%	25.98%	29.70%
Round6	28.72%	28.86%	44.78%	24.29%	22.88%	27.32%
Round7	26.53%	26.27%	42.14%	23.15%	22.79%	25.66%
Round8	28.27%	27.76%	45.70%	25.37%	25.32%	29.81%
Round9	27.59%	27.42%	40.86%	24.10%	23.98%	26.46%
Round10	26.94%	26.79%	41.66%	22.05%	22.39%	25.10%
Average	28.12%	27.80%	42.95%	24.56%	23.72%	27.20%

Table B-8.44 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2，CBR-W3S2 (50-size sample and K=1)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	33.98%	36.46%	52.02%	29.18%	31.75%	38.54%
Round2	48.10%	40.00%	49.73%	25.25%	21.10%	51.08%
Round3	39.67%	44.59%	32.16%	42.23%	36.34%	60.52%
Round4	57.29%	24.88%	45.06%	39.05%	36.89%	33.67%
Round5	41.75%	37.70%	42.46%	31.29%	28.44%	37.65%
Round6	36.70%	34.26%	36.56%	29.82%	35.08%	57.14%
Round7	22.04%	19.76%	38.39%	15.37%	17.44%	14.08%
Round8	53.64%	26.67%	53.95%	25.62%	22.33%	28.05%
Round9	22.44%	24.78%	29.52%	31.98%	26.04%	42.07%
Round10	28.90%	24.07%	19.76%	27.38%	29.27%	33.79%
Average	38.45%	31.32%	39.96%	29.72%	28.47%	39.66%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	21.88%	38.44%	47.19%	32.78%	33.82%	40.30%
Round2	74.88%	109.45%	73.06%	36.97%	28.07%	58.60%
Round3	28.48%	45.98%	45.05%	42.46%	36.73%	55.51%
Round4	29.82%	27.28%	45.74%	41.55%	44.42%	41.97%
Round5	20.33%	38.20%	51.65%	34.94%	31.88%	42.79%
Round6	19.69%	31.74%	44.20%	39.12%	39.22%	59.86%
Round7	41.93%	26.50%	39.87%	19.95%	28.06%	25.02%
Round8	62.32%	42.79%	92.57%	34.15%	32.75%	34.48%
Round9	38.25%	32.93%	43.81%	35.76%	28.96%	51.90%
Round10	39.03%	34.05%	26.47%	35.02%	37.33%	45.83%
Average	37.66%	42.74%	50.96%	35.27%	34.12%	45.63%

Table B-8.55 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2, CBR-W3S2 (100-size sample and K=1)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	34.45%	30.43%	52.58%	39.17%	35.57%	50.35%
Round2	21.22%	23.11%	35.43%	18.80%	28.69%	35.43%
Round3	28.05%	33.21%	33.49%	25.53%	28.89%	31.78%
Round4	36.07%	36.99%	47.67%	34.30%	29.95%	28.58%
Round5	27.69%	29.30%	46.98%	24.86%	28.86%	34.96%
Round6	23.87%	24.17%	21.77%	19.42%	12.93%	28.68%
Round7	28.65%	29.03%	37.10%	32.73%	21.02%	42.22%
Round8	22.01%	24.37%	41.43%	24.00%	17.44%	45.29%
Round9	30.46%	29.08%	30.24%	36.17%	30.69%	34.17%
Round10	17.65%	22.83%	29.60%	20.64%	17.63%	19.48%
Average	27.01%	28.25%	37.63%	27.56%	25.17%	35.09%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	29.00%	34.44%	49.46%	39.94%	35.68%	49.34%
Round2	23.49%	26.90%	57.64%	19.96%	38.16%	46.70%
Round3	27.40%	38.21%	51.88%	31.82%	33.23%	35.58%
Round4	20.51%	38.72%	50.54%	41.34%	44.84%	36.37%
Round5	28.34%	30.52%	46.79%	30.47%	34.21%	35.20%
Round6	22.71%	29.05%	33.03%	28.44%	19.15%	40.04%
Round7	28.56%	32.85%	40.56%	38.28%	25.67%	46.23%
Round8	32.76%	29.23%	62.64%	33.41%	29.11%	50.54%
Round9	32.70%	35.34%	35.74%	47.88%	43.01%	44.45%
Round10	31.27%	29.54%	47.82%	28.08%	24.97%	27.64%
Average	27.67%	32.48%	47.61%	33.96%	32.80%	41.21%

Table B-16 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2, CBR-W3S2 (200-size sample and K=1)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	29.43%	29.69%	30.27%	22.84%	22.55%	23.83%
Round2	25.29%	19.30%	35.35%	25.50%	23.32%	28.96%
Round3	29.85%	29.45%	30.78%	26.14%	24.39%	38.78%
Round4	29.28%	25.80%	33.78%	22.57%	19.34%	27.91%
Round5	32.33%	33.16%	34.47%	20.39%	21.96%	29.91%
Round6	28.16%	24.28%	33.92%	21.65%	24.16%	28.39%
Round7	31.42%	32.41%	36.09%	18.84%	18.20%	19.92%
Round8	25.61%	22.64%	32.08%	18.62%	19.20%	29.14%
Round9	31.16%	31.18%	35.68%	26.36%	25.92%	28.97%
Round10	28.66%	24.06%	31.62%	20.06%	23.54%	32.90%
Average	29.12%	27.20%	33.40%	22.30%	22.26%	28.87%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	22.95%	32.17%	52.84%	32.83%	32.31%	31.37%
Round2	40.45%	24.60%	58.42%	34.21%	31.32%	42.12%
Round3	27.59%	31.99%	48.55%	31.78%	29.29%	43.04%
Round4	31.91%	27.79%	41.85%	27.33%	24.16%	33.33%
Round5	34.24%	52.71%	53.27%	25.70%	25.13%	38.88%
Round6	26.90%	29.74%	41.65%	28.89%	31.63%	39.39%
Round7	43.69%	37.28%	54.13%	25.46%	25.59%	28.24%
Round8	26.56%	27.69%	46.84%	27.36%	26.90%	38.22%
Round9	35.69%	33.99%	46.31%	30.60%	29.50%	34.44%
Round10	28.23%	25.32%	55.77%	24.31%	27.37%	40.52%
Average	31.82%	32.33%	49.96%	28.85%	28.32%	36.95%

Table B-17 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2，CBR-W3S2 (400-size sample and K=1)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	26.60%	27.83%	35.01%	19.10%	20.89%	23.34%
Round2	31.03%	27.86%	31.98%	25.32%	24.83%	25.36%
Round3	29.18%	26.74%	38.08%	21.99%	22.56%	26.47%
Round4	34.51%	35.66%	30.71%	29.72%	25.80%	30.37%
Round5	27.80%	26.67%	38.44%	24.53%	24.45%	28.66%
Round6	25.51%	27.09%	33.24%	20.59%	22.33%	25.14%
Round7	27.60%	28.13%	29.12%	20.88%	20.17%	25.60%
Round8	26.22%	25.82%	29.31%	24.13%	20.70%	27.82%
Round9	30.71%	34.12%	29.70%	21.50%	22.54%	19.74%
Round10	26.28%	24.12%	34.86%	18.07%	21.38%	24.15%
Average	28.54%	28.40%	33.04%	22.58%	22.56%	25.66%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	30.09%	31.93%	53.44%	24.34%	27.98%	30.50%
Round2	28.14%	31.70%	38.44%	31.51%	30.62%	31.06%
Round3	31.52%	30.80%	56.55%	30.49%	30.15%	33.49%
Round4	27.33%	38.58%	41.03%	35.60%	32.37%	37.67%
Round5	24.79%	30.59%	47.19%	31.35%	31.29%	34.04%
Round6	29.73%	31.59%	44.81%	27.59%	30.39%	31.07%
Round7	25.51%	33.07%	40.17%	26.12%	25.22%	32.87%
Round8	27.49%	29.82%	46.98%	31.32%	25.73%	37.74%
Round9	25.09%	37.62%	47.47%	27.73%	29.09%	26.10%
Round10	32.05%	28.56%	52.59%	24.37%	27.37%	31.54%
Average	28.18%	32.42%	46.87%	29.04%	29.02%	32.61%

Table B-8.68 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2, CBR-W3S2 (600-size sample and K=1)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	29.35%	29.34%	34.48%	22.86%	23.23%	27.25%
Round2	26.69%	27.63%	33.86%	22.17%	22.96%	26.80%
Round3	34.09%	31.09%	39.73%	19.02%	19.17%	22.16%
Round4	26.85%	27.36%	34.25%	23.81%	22.66%	23.15%
Round5	28.87%	28.66%	33.15%	23.57%	25.02%	31.95%
Round6	31.74%	30.49%	32.39%	19.73%	19.60%	21.00%
Round7	24.97%	25.43%	32.23%	19.27%	20.61%	22.80%
Round8	27.65%	26.27%	38.21%	19.70%	21.31%	23.95%
Round9	26.23%	29.74%	33.88%	22.10%	22.75%	27.54%
Round10	26.95%	26.45%	31.59%	16.91%	16.43%	21.25%
Average	28.34%	28.24%	34.38%	20.91%	21.37%	24.78%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	25.17%	31.92%	63.29%	28.78%	29.12%	35.10%
Round2	29.56%	30.91%	45.01%	26.75%	27.94%	32.28%
Round3	26.54%	33.78%	60.38%	25.93%	24.72%	30.23%
Round4	27.81%	32.83%	55.71%	30.44%	29.33%	30.25%
Round5	27.57%	32.28%	50.07%	29.24%	30.66%	38.40%
Round6	28.43%	34.34%	46.29%	29.85%	29.69%	31.95%
Round7	28.12%	30.64%	52.54%	27.96%	28.26%	31.64%
Round8	27.05%	30.38%	58.01%	25.16%	27.63%	30.22%
Round9	28.00%	34.48%	44.72%	27.81%	28.67%	34.42%
Round10	26.83%	30.23%	51.84%	23.41%	23.16%	32.44%
Average	27.51%	32.18%	52.79%	27.53%	27.92%	32.69%

Table B-198.7 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2，CBR-W3S2 (800-size sample and K=1)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	29.92%	28.05%	36.93%	19.86%	22.18%	22.42%
Round2	30.56%	30.25%	36.69%	21.93%	25.19%	25.21%
Round3	28.56%	27.09%	36.87%	21.82%	23.26%	22.81%
Round4	25.54%	24.96%	31.91%	20.10%	22.26%	25.01%
Round5	27.94%	30.34%	36.87%	19.77%	19.75%	21.98%
Round6	28.93%	27.45%	31.99%	24.06%	22.30%	25.90%
Round7	26.70%	25.15%	30.43%	20.48%	21.43%	22.93%
Round8	25.15%	25.95%	33.55%	20.76%	18.95%	22.71%
Round9	30.66%	29.96%	40.77%	22.85%	21.63%	24.39%
Round10	28.25%	26.88%	33.49%	23.45%	22.61%	24.24%
Average	28.22%	27.61%	34.95%	21.51%	21.96%	23.76%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	29.90%	33.50%	55.20%	26.47%	28.53%	30.06%
Round2	26.50%	33.65%	49.83%	28.78%	30.87%	34.81%
Round3	24.23%	31.76%	49.21%	29.04%	30.49%	29.62%
Round4	29.36%	29.44%	47.01%	27.68%	29.22%	36.37%
Round5	30.11%	34.25%	53.29%	25.85%	25.68%	28.48%
Round6	30.76%	30.92%	41.02%	29.67%	27.92%	31.93%
Round7	27.85%	29.95%	49.22%	27.22%	27.76%	30.43%
Round8	28.34%	30.27%	49.95%	26.95%	25.39%	30.35%
Round9	27.77%	33.84%	53.33%	28.99%	27.60%	30.69%
Round10	28.87%	31.04%	49.85%	31.02%	28.67%	31.27%
Average	28.37%	31.86%	49.79%	28.17%	28.21%	31.40%

Table B-8.80 Comparative results of CBR-W1S1, CBR-W2S1, CBR-W3S1, CBR-W1S2, and CBR-W2S2，CBR-W3S2 (1000-size sample and K=1)

ARE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	29.12%	27.47%	29.59%	21.23%	19.25%	24.47%
Round2	30.88%	30.42%	37.55%	19.51%	20.32%	21.91%
Round3	25.79%	25.07%	32.19%	23.05%	21.19%	22.51%
Round4	30.55%	29.26%	38.97%	23.65%	21.94%	27.14%
Round5	33.64%	33.91%	34.84%	23.59%	22.95%	26.14%
Round6	28.24%	28.16%	34.13%	20.39%	18.33%	22.47%
Round7	26.84%	27.21%	30.73%	21.54%	22.70%	23.40%
Round8	26.75%	27.68%	35.91%	22.55%	22.22%	26.23%
Round9	26.69%	26.48%	34.06%	21.57%	21.93%	24.29%
Round10	25.72%	26.74%	31.87%	19.69%	21.20%	21.61%
Average	28.42%	28.24%	33.99%	21.68%	21.20%	24.02%
RMSE						
	CBR-W1S1	CBR-W2S1	CBR-W3S1	CBR-W1S2	CBR-W2S2	CBR-W3S2
Round1	30.20%	31.81%	45.22%	27.41%	25.68%	33.54%
Round2	26.71%	33.87%	55.14%	25.05%	25.53%	32.07%
Round3	28.55%	29.19%	47.66%	30.70%	28.35%	30.11%
Round4	28.68%	33.35%	53.19%	31.09%	29.10%	34.50%
Round5	26.28%	35.91%	47.85%	28.86%	28.28%	33.01%
Round6	27.64%	32.68%	48.26%	26.58%	24.92%	30.41%
Round7	28.20%	30.66%	45.49%	27.75%	28.68%	29.80%
Round8	29.08%	31.79%	54.18%	29.99%	28.53%	34.27%
Round9	26.49%	30.64%	50.14%	27.76%	28.46%	33.02%
Round10	27.27%	30.70%	49.85%	25.35%	26.52%	27.59%
Average	27.91%	32.06%	49.70%	28.05%	27.40%	31.83%

APPENDIX C

Table C-1 The error rate of OLS-CBR after case-base editing (K=3)

OLS-CBR (K=3)								
Threshold	No. of cases	CR	Strategy 1		Strategy 2		Strategy 3	
			MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	1315	0.9074	0.1789	0.2287	0.1832	0.2332	0.1764	0.2253
2	1242	0.8571	0.1771	0.2246	0.1835	0.2329	0.1733	0.2203
3	1079	0.7443	0.1814	0.2326	0.1837	0.2345	0.1768	0.2263
4	948	0.6545	0.1819	0.2323	0.1854	0.2363	0.1758	0.2250
5	792	0.5467	0.1857	0.2343	0.1896	0.2397	0.1730	0.2219
6	631	0.4353	0.1819	0.2343	0.1925	0.2441	0.1628	0.2088
7	476	0.3288	0.1889	0.2396	0.1962	0.2475	0.1645	0.2110
8	340	0.2346	0.1967	0.2513	0.2001	0.2552	0.1710	0.2186
9	226	0.1563	0.2098	0.2667	0.2087	0.2634	0.1706	0.2181
10	138	0.0954	0.2220	0.2763	0.2169	0.2707	0.1709	0.2199
11	65	0.0446	0.2264	0.2903	0.2247	0.2854	0.1676	0.2176
12	37	0.0253	0.2748	0.3408	0.2551	0.3080	0.1584	0.2073
13	18	0.0124	0.3556	0.4214	0.3363	0.3905	0.1778	0.2296
14	7	0.0047	0.6828	0.7063	0.6078	0.6454	0.1943	0.2556
15	6	0.0041	1.3851	0.9831	1.3202	0.9431	0.1947	0.2473

CR is the compression rate.

Table C-2 The error rate of GA -CBR after case-base editing (K=3)

GA-CBR (K=3)								
Threshold	No. of cases	CR	Strategy 1		Strategy 2		Strategy 3	
			MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	1309.3655	0.9036	0.1779	0.2260	0.1812	0.2292	0.1773	0.2248
2	1234.8497	0.8522	0.1811	0.2283	0.1817	0.2290	0.1771	0.2240
3	1073.6007	0.7409	0.1811	0.2267	0.1818	0.2282	0.1773	0.2237
4	944.2545	0.6517	0.1798	0.2272	0.1829	0.2288	0.1760	0.2214
5	789.8828	0.5451	0.1830	0.2315	0.1851	0.2323	0.1753	0.2228
6	621.8269	0.4291	0.1876	0.2364	0.1897	0.2384	0.1777	0.2239
7	471.0441	0.3251	0.1883	0.2350	0.1907	0.2383	0.1735	0.2189
8	339.9876	0.2346	0.1930	0.2396	0.1933	0.2435	0.1722	0.2185
9	229.2993	0.1582	0.1978	0.2488	0.1981	0.2486	0.1754	0.2235
10	142.0386	0.0980	0.2021	0.2554	0.2043	0.2573	0.1746	0.2219
11	82.7655	0.0571	0.2150	0.2679	0.2133	0.2658	0.1745	0.2241
12	44.9952	0.0311	0.2479	0.3025	0.2436	0.2998	0.1767	0.2261
13	22.6372	0.0156	0.3243	0.3756	0.3171	0.3698	0.1751	0.2244
14	10.7738	0.0074	0.6819	0.5980	0.6613	0.5874	0.1757	0.2239
15	5.5152	0.0038	1.7221	0.9669	1.7136	0.9642	0.1868	0.2405

CR is the compression rate.

Table C-3 The error rate of MODAL-CBR after case-base editing (K=3)

MODAL-CBR (K=3)								
Threshold	No. of cases	CR	Strategy 1		Strategy 2		Strategy 3	
			MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	1317	0.9086	0.1523	0.1986	0.1478	0.1944	0.1489	0.1940
2	1237	0.8537	0.1532	0.1979	0.1493	0.1945	0.1479	0.1934
3	1085	0.7487	0.1590	0.2067	0.1558	0.2036	0.1467	0.1920
4	946	0.6527	0.1664	0.2185	0.1632	0.2127	0.1472	0.1946
5	771	0.5320	0.1738	0.2256	0.1710	0.2213	0.1489	0.1969
6	598	0.4129	0.1819	0.2338	0.1788	0.2294	0.1499	0.1974
7	454	0.3133	0.1933	0.2453	0.1874	0.2395	0.1569	0.2034
8	344	0.2376	0.1987	0.2501	0.1936	0.2468	0.1666	0.2148
9	236	0.1631	0.2059	0.2588	0.1976	0.2475	0.1662	0.2157
10	142	0.0980	0.2122	0.2678	0.2117	0.2671	0.1654	0.2101
11	81	0.0559	0.2237	0.2782	0.2281	0.2826	0.1672	0.2140
12	48	0.0334	0.2878	0.3480	0.2651	0.3215	0.1653	0.2130
13	30	0.0210	0.3418	0.4103	0.3048	0.3600	0.1698	0.2172
14	16	0.0107	0.3733	0.4786	0.3456	0.4304	0.1758	0.2238
15	9	0.0062	0.4273	0.5122	0.3931	0.4542	0.1905	0.2469

CR is the compression rate.

Table C-4 The error rate of OLS-CBR after case-base editing (K=1)

OLS-CBR (K=1)								
Threshold	No. of cases	CR	Strategy 1		Strategy 2		Strategy 3	
			MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	369	0.2546	0.2260	0.2977	0.2294	0.3004	0.1921	0.2563
2	68	0.0472	0.3150	0.3811	0.2870	0.3526	0.1874	0.2524
3	13	0.0089	0.4532	0.5615	0.3292	0.4086	0.2148	0.2865

CR is the compression rate.

Table C-5 The error rate of GA -CBR after case-base editing (K=1)

GA-CBR (K=1)								
Threshold	No. of cases	CR	Strategy 1		Strategy 2		Strategy 3	
			MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	376	0.2595	0.2264	0.2837	0.2202	0.2820	0.1958	0.2620
2	73	0.0500	0.2577	0.3195	0.2577	0.3175	0.1940	0.2565
3	9	0.0064	0.4591	0.5006	0.4467	0.4928	0.2016	0.2661

CR is the compression rate.

Table C-6 The error rate of MODAL-CBR after case-base editing (K=1)

MODAL-CBR (K=1)								
Threshold	No. of cases	Strategy 1			Strategy 2		Strategy 3	
		CR	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE
1	364	0.2513	0.2120	0.2780	0.2040	0.2677	0.1808	0.2423
2	82	0.0565	0.2723	0.3273	0.2600	0.3185	0.1925	0.2553
3	7	0.0051	0.4674	0.6801	0.4019	0.5566	0.2157	0.2943

CR is the compression rate.