Doctor of Philosophy

Conceptual Estimation of Construction Cost by Neural Network Ensemble Learning incorporating Factor Analysis

February, 2018

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

Jin Gang Lee

Abstract

Conceptual Estimation of Construction Cost by Neural Network Ensemble Learning incorporating Factor Analysis

Jin Gang Lee

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

Conceptual cost estimation at the planning stage of project development is critical for the successful execution of the construction project as the estimated cost is an important resource for decision making for all stakeholders. Despite the importance of conceptual cost estimation, accurate estimation of cost budgets is a difficult task due to the increasing complexity of the project and limited information availability in the early phase of the project. To this end, estimators and researchers developed a number of techniques forecasting project cost to achieve the benefits of utilizing past project information.

In order to improve cost estimate accuracy and reliability, researchers have been applying a data-driven approach to cost forecasting model that maximize the value of past project data. The data-driven approach requires project data as much as possible to be collected to ensure the ability of that data to fully reflect the project and the accuracy of forecasting. However, the previous models are too often developed without due consideration given to the effect that the chracteristics of input variables have on model complexity and performance of the subsequently trained forecasting model. In-depth analysis of input data chractetristics is necessary as the performance of the forecasting model is profoundly affected by the characteristics of input data. In this aspect, this research focuses on the complexity of forecasting model to be addressed for better performance of the conceptual cost forecasting model.

The main aim of this research is to investigate the effects of model complexity on the accuracy of conceptual cost forecasting. The first objective is to reexamine the current conceptual estimation practices and model complexity issues when developing data-driven conceptual cost estimation model. And next objective is to develop conceptual cost forecasting model incorporating artificial neural network, ensemble modeling and factor analysis that could help in providing improved accuracy of conceptual cost forecasting at the early stage of the project development.

The proposed conceptual cost forecasting model is verified and validated by several experiments and case studies. Three types of construction project data including combined cycle power plant, high-rise building and government office building are utilized for the case studies. Under limited numbers of case project, the proposed conceptual cost forecasting model compensate for the limitations of project data utilization by showing improved accuracy. The results of case studies confirm the ideas of ensemble methods application being able to improve the accuracy of estimates and broaden the possible application of the proposed cost forecasting model.

The findings from this research show an opportunity to improve the accuracy and stability of conceptual cost estimation by proposing more flexible methods incorporating artificial neural network, ensemble methods, and factor analysis. Further this research can contribute to making a shift from traditional uses of project data to more enhanced resources for performance prediction support by exploring the usefulness of project data.

Keywords: Project Cost Estimation, Conceptual Cost Estimation, Artificial Neural Network, Ensemble Method, Factor Analysis, Model Complexity

Student Number: 2013-30174

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. Introduction

1.1 Research Background

The planning phase is particularly vital as decisions made during the early stages of the project development can result in more significant consequences than the decisions which can be made later in the project life-cycle (Dominic D A. 2012). As project cost is considered as one of the most critical management element for successful construction project including schedule and quality (Dursun, O. and Stoy, C. 2016), completing the project within the budget expected in the planning phase is a common and essential determinant for their decision making for most clients and contractors. Economic feasibility studies may be tried at this stage to analyze initial costs and project benefit including internal investment studies. This initial process may also affect the client's and contractor's decision on whether or not to go with the project.

These initial preparation and economic feasibility studies for a construction project are based on the cost estimates prepared at the early stages of the project. In this aspect, conceptual cost estimation is of great importance in project management as it provides essential information for

decision making at the planning phase of the project. Therefore, reliable estimation of project cost at the planning stage is critical for the successful planning and execution of construction project as the estimated cost is an essential information for decision making for all stakeholders. (Skitmore and Ng, 2003; Savas and Bayrum, 2016). Despite the importance of conceptual cost estimation, accurate estimation of cost budgets is a difficult task due to the increasing complexity of work scope among the various stakeholders including project owners, contractors, and facility managers in modern construction projects. Competitive environment in the construction industry, limited information availability, and inherent complex nature of project development may result in inaccurate budgeting and cost overrun at the completion of projects.

Project participants identified increasing estimating accuracy as a key strategic issue in their organizations and cost estimators need effective estimation strategies in the early stages of the project. The reliable comprehensive conceptual estimating of the cost of a project is fundamental and vital to any organization, and this skill set gives a “competitive advantage” both owners and contractors including EPC companies. This is why accurate and timely conceptual estimating is so crucial in today’s competitive global marketplace.

1.2 Problem Statement

Even though contractors identified increasing estimating accuracy as a key strategic issue in their project development, the actual final costs of construction projects still show a significant difference from their forecasted initially costs. The consequence cost overrun is a critical problem across the construction industry (Hemanta Kumar Doloi, 2011). Since a huge amount of cost is required for the project development, the cost overrun can reduce the company's and industry's competitiveness (Kim, 2005; Kirkham 2014). Prepared with limited information availability, the estimated cost at the conceptual phase is known as having a low level of accuracy. Although detailed estimation work can be started for accurate forecasting, preparing a detailed estimate is time-consuming and therefore more expensive. As the bidding phase progresses, more information will be gathered, and more detailed estimates will be produced, but it is not suitable for quick decisions early in the bidding phase. Due to the purpose and requirements of conceptual cost estimation, the estimates must be prepared with reliable accuracy within minimal amount time and efforts.

To this end, for accurate and quick conceptual estimation, estimators and researchers developed many techniques forecasting project cost to achieve

the benefits of utilizing past project information. To improve cost estimate accuracy and reliability, researchers have been working to apply a data-driven approach, machine learning algorithm, to cost forecasting model that maximize the practical value of historical information available. However, the data-driven approach requires data as much as possible to be collected to ensure the ability of that data to fully reflect the project and the accuracy of forecasting. Machine learning algorithms are often black box models, which cannot explain the process of producing forecasts. Therefore, the performance of data-driven conceptual cost forecasting models is subject to variations in data characteristics; this includes the amount of project data and the number and types of attributes used for forecasting. As poor quality data can cause a major competitive disadvantage due to unreasonable decision making in the business (Redman, 1998), the unreliable information input cannot validate the result of data-driven research (De Veaux and Hand, 2005). In particular, the previous models are too often developed without due consideration given to the effect that the choice of input variables has on model complexity and performance of the subsequently trained forecasting model. In-depth analysis of input data is necessary as the performance of the forecasting model is profoundly affected by the characteristics of input data. In this aspect, this research focuses on the complexity of forecasting model, and discuss the issues that need to be addressed to improve the performance of the conceptual cost forecasting model.

1.3 Research Objective and Scope

The primary objective of this dissertation is improving performance and applicability of conceptual cost forecasting model. To achieve this objective, three specific objectives were performed as follows.

1) Examinations of the conceptual cost estimation practices and researches: in order to define the conceptual cost estimation model requirements and challenging issues and its development way sets.

It is necessary to investigate the current problems in estimating project cost at the early stages of the construction project and issues of data-driven cost forecasting research for the establishment of model development.

2) Development of a conceptual cost forecasting model by combining artificial neural networks, ensemble modeling, and factor analysis to better forecast project costs with project-level information.

Based on the examined problems in conceptual cost estimation practices and issues of data-driven cost forecasting research, a conceptual cost forecasting model that have availability for improving forecasting performance is developed.

3) Investigation of the effects of model complexity and the feasibility of methodology proposed in the conceptual cost estimation through the case studies using three types of project data; combined cycle power plant, high-rise building and government office building.

In addition to verification and validation the applicability of the proposed model, the effectiveness of the proposed methodologies to mitigate the problems in developing a conceptual cost forecasting model are investigated when testing the developed cost forecasting model.

Hence, this research scope is mainly focused on conceptual cost estimation of the construction project. And the author focused on developing a mathematical cost forecasting model based on the data collected at the early stage of the project. After investigating the current problems in estimating project cost at the early stages of construction project and issues of data-driven cost forecasting research, conceptual cost forecasting model incorporating artificial neural network, ensemble method, and factor analysis are proposed and developed. And to verify and validate the proposed model in this research, the project data from the combined cycle power plant, high-rise building, and government office buildings are used. The overall framework for this research is summarized in figure 1-1.

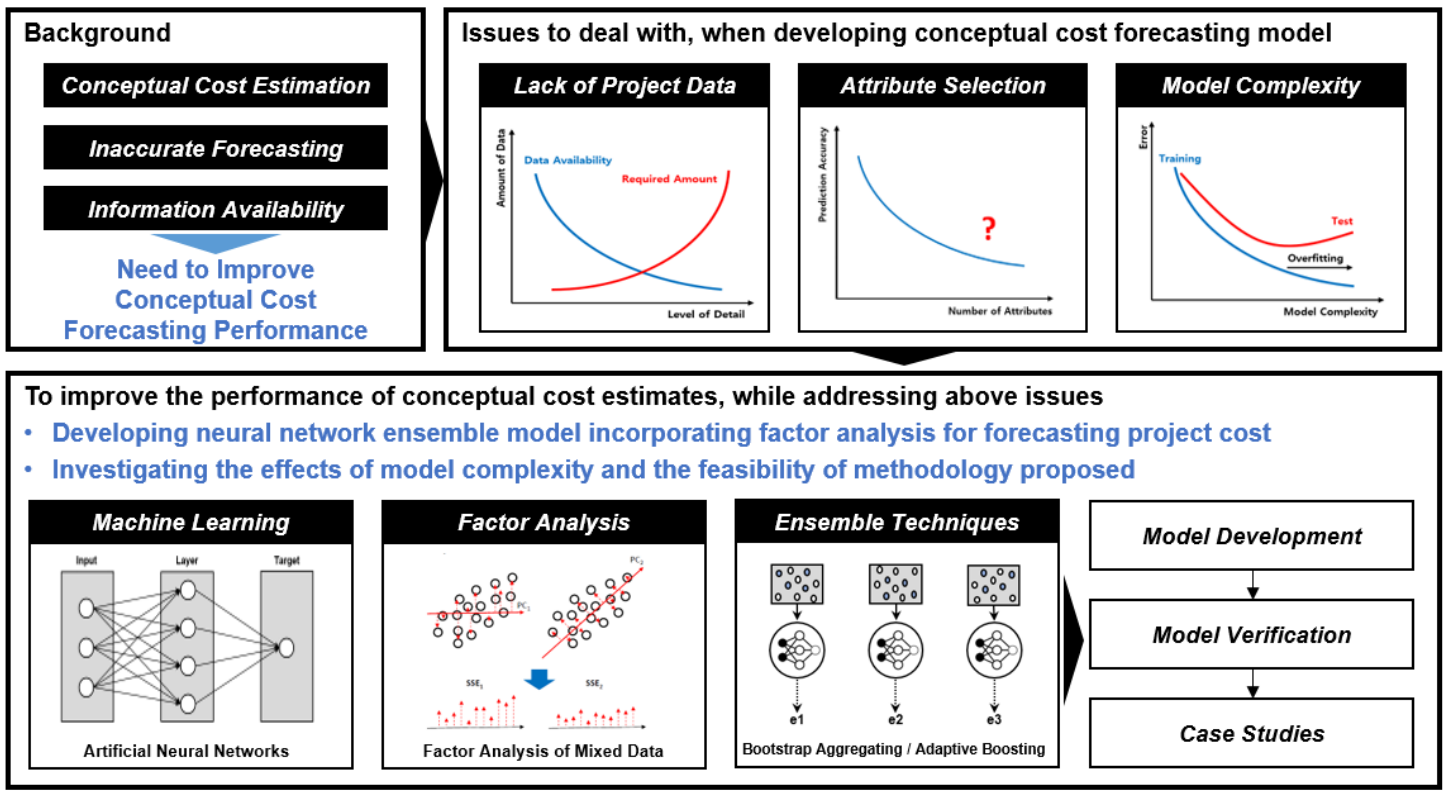


Figure 1-1. Research Framework

1.4 Organization of Dissertation

This dissertation consists of six chapters including this introduction. Descriptions of the following chapters are summarized below.

Chapter 1. Introduction

This research begins with addressing research backgrounds, problem statement, research objective, and research scope and process.

Chapter 2. Theoretical Background

This chapter introduces the essential basics of conceptual cost estimation and previous literature on developing conceptual cost estimation model using machine learning to form the theoretical foundation of this research. The subjects of a literature review of conceptual cost estimation using a machine learning approach are presented. And challenging issues of developing cost forecasting model are clarified.

Chapter 3. Model Development

This chapter first explains the overall concept of model development framework based on the theoretical backgrounds. The proposed model addresses the issues as mentioned earlier in the previous chapter. The

conceptual cost forecasting model developed in this research is explained with its purpose, process, functions and methodologies. Also, the detailed description of applied methodologies including artificial neural networks, ensemble modeling, and factor analysis of mixed data is explained.

Chapter 4. Model Verification

To obtain verification of the proposed model, this chapter conducts comparative experiments with several types of project data that test the proposed concept used for developing conceptual cost estimation model. The data set include power plant project and high-rise building project.

Chapter 5. Case Studies

This chapter describes case studies using the proposed conceptual cost forecasting model. The case studies are designed to demonstrate and validate the effectiveness and applicability of the developed model. Total three types of project data including combined cycle power plant, high-rise building and government office building are utilized for case studies.

Chapter 6. Conclusions

This chapter provides an overall review of the results of this research, expected contributions, limitations, and required future works to overcome the limitations.

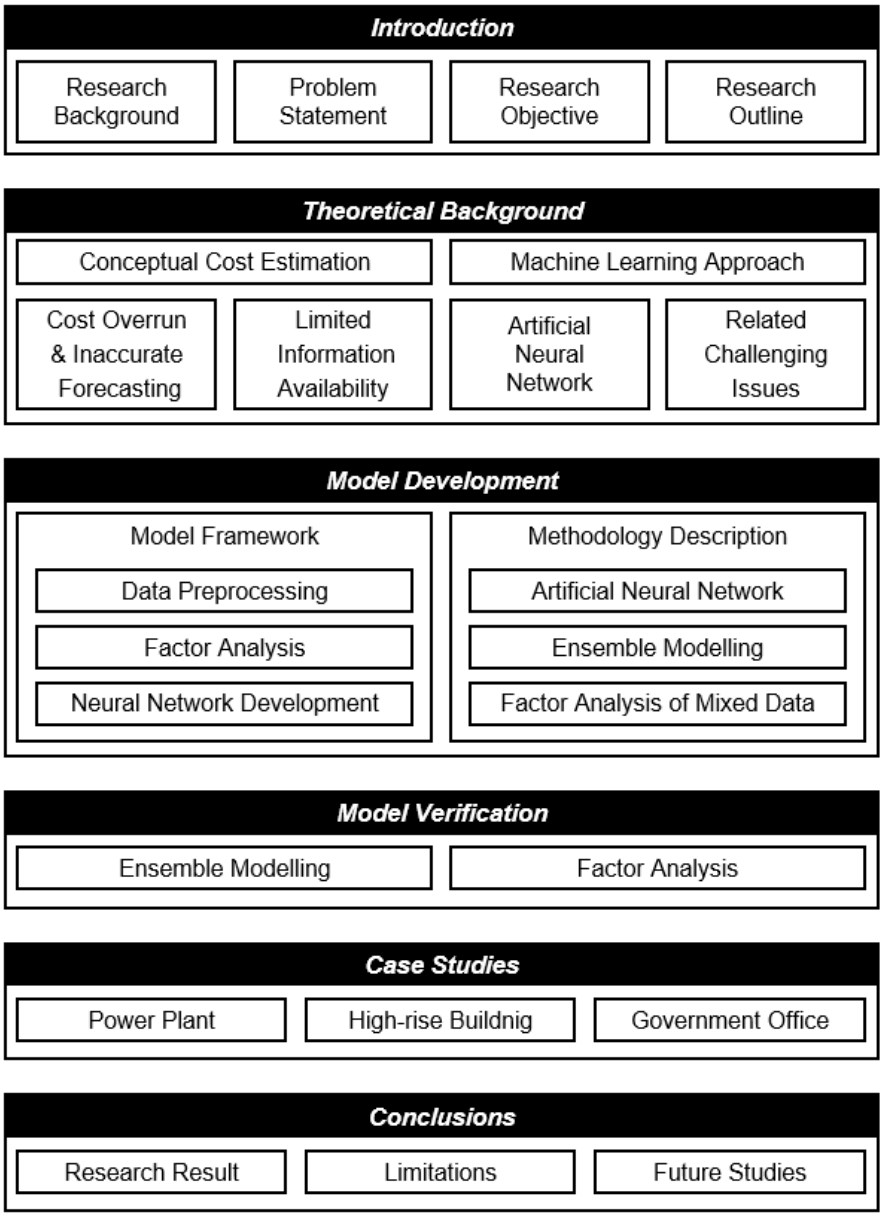


Figure 1-2. Outline of Dissertation

Chapter 2. Theoretical Background

This chapter introduces the essential basics of conceptual cost estimation and previous literature on developing conceptual cost estimation model using machine learning to form the theoretical foundation of this research. In the first section, a comprehensive overview of conceptual estimation is summarized with its concepts, roles, importance, and methodologies. Also, the problems and issues that are currently facing in during conceptual estimation at an early stage of project development are reviewed in the next section. In the third section, data-driven conceptual cost estimation research using a machine learning algorithm (especially artificial neural networks) is studied to identify the requirements to develop a conceptual cost estimation model for a construction project. In the last section, several challenging issues when developing data-driven conceptual cost estimation model are also described.

2.1 Conceptual Cost Estimation

The primary goals at the first stage of the project are identifying project requirement and establishing a plan for satisfying those requirements. Identifying requirements of project owner include deciding on the scope of the project that is suggested. Various options will be proposed at this phase and should be evaluated based on estimated cost, expected quality and benefit within the uncertain and risky environment of the construction project. This planning phase is particularly critical as decisions made during the early stages of the project development can cause more significant consequences than the decisions which can be made later in the project life-cycle. For most clients and contractors, completing the project within the budget expected in the planning phase is a common and important determinant for their decision making. Economic feasibility studies may be tried at this stage to analyze initial costs and project benefit including internal investment studies. This initial process may also affect the client's and contractor's decision on whether or not to go with the project. These initial appropriation and economic feasibility studies for a construction project are based on the cost estimates prepared at the early stages of the project. For this purpose, the client should estimate the budget for the project development and the contractor should estimate the price for the bid.

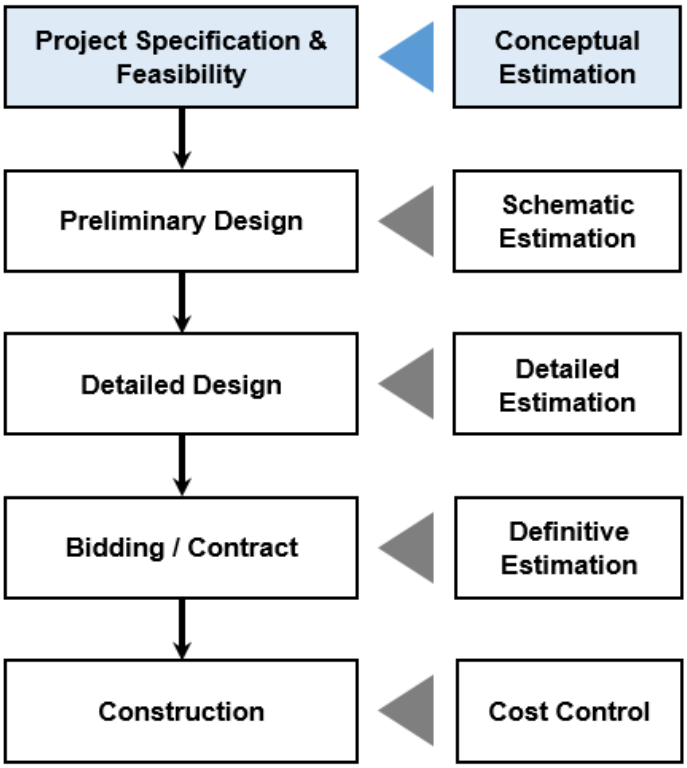


Figure 2-1. Project Development Cycle and Cost Estimation Classification

To this end, the effort of estimating a conceptual cost is one of the first actions that are made to ascertain if the proposed project is an economically viable enterprise or whether it has an opportunity for the business. Reliable conceptual cost estimation can be the first step in the successful execution of a project.

2.1.1 Overview of Conceptual Cost Estimation

Cost estimation is the process of forecasting project cost that has to be acquired to complete the project. Different types of cost estimation are applied in construction practice depending on their purpose, information availability, and the phases of project development currently. Several cost estimation methods for different stages of a project can be found in the literature and classified by its applicable phase of estimation from the preliminary phase to the final phase. In the preliminary phase of the project, conceptual cost estimation is used to support feasibility studies of project development for both client and contractors.

The conceptual estimates are established at the initial phase of project development for the client and the initial stage of the bidding phase for contractors. The project cost takes into consideration every expense involved in engineering, procurement, and construction efforts for the project development. Estimating efforts must be made in the early stages of a project to evaluate the future project, and it should be based on the current and anticipated scope of work statement and typically some specific key milestones that need to be achieved for the particular project.

Table 2-1. Generic Cost Estimate Classification Matrix

Estimate Class	Primary Characteristics	Secondary Characteristics		
	Level of Project Definition	End Usage	Expected Accuracy Range (a)	Preparation Effort (b)
Class 5	0% to 2%	Screening or Feasibility	L: -20% to -50% H: +30% to +100%	1
Class 4	1% to 15%	Concept Study or Feasibility	L: -15% to -30% H: +20% to +50%	2 to 4
Class 3	10% to 40%	Budget, Authorization, or Control	L: -10% to -20% H: +10% to +30%	3 to 10
Class 2	30% to 70%	Control or Bid/ Tender	L: -5% to -15% H: +5% to +20%	5 to 20
Class 1	50% to 100%	Check Estimate or Bid/Tender	L: -3% to -10% H: +3% to +15%	10 to 100

Note: (a) Typical variation in low and high ranges, (b) Typical degree of effort relative to least cost index of 1.

The cost index value of "1" represents 0.005% of project costs, then an index value of 100 represents 0.5%.

The conceptual estimate is an estimation of the cost of a proposed project based on rough project information and related conceptual engineering and design data. Conceptual cost estimates may be and are prepared and compiled at the start point in the projects lifecycle development from the initial early stages 0% to perhaps 5% of the design has been established. The estimate class designations defined by the Association for the Advancement of Cost Engineering (Christensen, P. and Dysert, L. R. 2003) are labeled Class 1, 2, 3, 4, and 5. Table 2.1 provides a summary of the characteristics of the five estimate classes. The maturity level of definition is the sole determining (i.e. primary) characteristic of Class. In Table 2.1, the maturity is roughly indicated by a percent of complete definition. Class 5 estimate is based upon the lowest level of project definition, and a Class 1 estimate is closest to full project definition and maturity. According to the description from AACE, the maturity level of project definition level is 0% to 2% of full project definition at class 5 estimates and 1% to 15% of full project definition at class 4 estimates. The level of project definition defines maturity or the extent and types of input information available to the estimating process.

Prepared based on very limited information, the estimated cost at this conceptual stage subsequently have a wide range of accuracy. Although

detailed estimation work can be started for accurate forecasting, preparing a detailed estimate is time-consuming and therefore more expensive. As the bidding phase progresses, more information will be gathered, and more detailed estimates will be produced, but it is not suitable for quick decisions early in the bidding phase. Likewise, due to the purpose and requirements of conceptual cost estimation, the estimates may be prepared within very limited amount time and efforts. Sometimes little more than proposed project type, location, and size or capacity are known at the time of conceptual estimation. Even though the specific details are not clearly revealed or specified at this early stage of the proposed project, an accurate estimate is required to determine the viability of the proposed project.

Conceptual estimation supports a number of strategic business planning, such as but not limited to market studies, project screening, assessment of initial profitability, evaluation of project alternatives, and long-term resource planning. The prepared conceptual cost can be used as base materials to determine the profitability of the project, the return on investment (ROI), and a cash flow plan can be developed from the project budget. Also, the initial project budget can be used as an initial baseline and control tool, to use as a staring tool to select the best path forward in considering technology, future business opportunities and goals. Later updates of this early estimate will provide an audit trail of the projects life cycle. Therefore, an accurate

estimation of the project cost is crucial to contract administration since the forecasted cost serves as a guideline for budgeting, planning, and monitoring.

Accurate and logically established complete and precise project cost estimation is the important key of the competitive enterprise. Early cost evaluations are the essential decision-making tool for whether to proceed with a proposed project or not (perhaps the project needs to be stopped because of a considerable potential overrun in the final forecasted cost), or to delay the decision to proceed to another location, until sometime in the future three years from now when the economy is in more of a growth mode, or look for alternate process / technical / business solutions. Again reliable comprehensive conceptual estimating of the cost of a project is fundamental and vital to any organization, and this skill set gives a “competitive advantage” both owners and contractors including EPC companies. Therefore an accurate and timely conceptual estimating is so important in today’s competitive global marketplace.

2.1.2 Current Practice of Conceptual Cost Estimation

Estimating methodologies fall into two broad categories: stochastic and deterministic. In stochastic methods, the independent variables used in the cost estimating algorithms are generally something other than a direct measure of the units of the item being estimated. The cost estimating relationships used in stochastic methods often are somewhat subject to conjecture. With deterministic methods, the independent variables are more or less a definitive measure of the item being estimated. A deterministic methodology is not subject to significant conjecture. As the level of project definition increases, the estimating methodology tends to progress from stochastic to deterministic methods. Obviously, in the early stages of the project, the value of the independent variables is not measured and uncertain, the stochastic approach is used when estimating the conceptual cost of the project. Shown in table 2.2, the stochastic method is used in class five and four levels. In addition to estimator's judgment based on their knowledge and experience, analogy-based methods, capacity factored estimation methods, or parametric methods are applied to class 5 estimates. And equipment factored estimation methods or parametric methods are used for class 4 estimates with more detailed information obtained.

Table 2-2. Estimating Methodology by Cost Estimate Classification

Estimate Class	**Typical Estimating Method**	**Typical Estimating Method for Process Industry**
Class 5	Stochastic or Judgment	Capacity factored, parametric models, judgment, or analogy
Class 4	Primarily Stochastic	Equipment factored or parametric models
Class 3	Mixed, but Primarily Stochastic	Semi-detailed unit costs with assembly level line items
Class 2	Primarily Deterministic	Detailed unit cost with forced detailed take-off
Class 1	Deterministic	Detailed unit cost with detailed take-off

When predicting costs in the conceptual planning stage, the analogous method and unit method estimating total cost by a cost per certain unit is widely used for practice across many projects and countries (Schexnayder et al. 2003). The client traditionally uses the unit price method in the planning phase of the project to forecast the total expense of the project. This unit price method calculates construction costs as the product of the unit cost and the total quantity of the unit based on the assumption that the total construction cost is proportional to the unit. Cost capacity method is also applied as the unit price method (Ahuja et al. 1994). The fundamental concept behind the cost-to-capacity method is that the costs of a similar project but with different sizes vary nonlinearly. These methods are an order-of-magnitude cost estimation method that uses past project cost and capacity to develop conceptual cost estimates for a new project (Ellsworth, Richard K. 2009). The advantages of the unit price method and cost capacity method is its relatively easy application to formulate conceptual cost estimates with only project capacity information quickly.

While applying the unit price cost method and cost capacity method based on these project sizes, various factors have been applied to the estimates. The raised scale factor, shown in equation one below, accounts for the nonlinear relationship and introduces the concept of economies of scale where, as a facility becomes larger, the incremental cost is reduced for each

additional unit of capacity. For example, a scale factor of less than 1 indicates that economies of scale exist and the incremental cost of the next added unit of capacity will be cheaper than the previous unit of capacity.

$$\frac{C_2}{C_1} = \left(\frac{Q_2}{Q_1}\right)^x \qquad \text{(Equation 1)}$$

$C_2 = \text{Estimated cost of project 2}$

$C_1 = \text{Known cost of project 1}$

$Q_2 = \text{Known capacity of project 2}$

$Q_1 = \text{Known capacity of project 1}$

$\text{x} = \text{Scale factor}$

In particular, in the process industry, various types of factors calculated from the relationship between the capacity of the facility and equipment and the total project cost were applied in the conceptual cost estimation. This is called equipment factored estimate, which is produced by taking the cost of individual types of process equipment. This method requires the type of individual equipment and specification information necessary for each type of plant project. The information at the initial stage of the project may be uncertain or absent, but equipment factored estimation method may be used to repeat the same type of plant project with the same equipment and facilities. Generally, Lang factors, Hand factors, Wroth factors. Chilton factors, Peters-

Timmerhaus factors, and Guthrie factors are utilized for equipment factored estimating methods to produce class 4 estimates (Christensen, P. and Dysert, L. R. 2003). Also, as an amount of information that can be collected at the initial stage of the project increases, many cases utilize the parametric method. Parametric estimating entails the analysis of cost, programmatic and technical data to identify cost drivers and develop cost models. The approach essentially correlates project cost and input information with parameters describing the item to be estimated. Parametric estimation method generally involves the use of regression analysis to determine the best algorithms for a model.

While the number of project factors to be considered increases, the traditional method overlooks various characteristics within the project, which usually yields inaccurate results. To obtain reasonable results, the essential project characteristics for which the cost is being estimated must be the same as, or very close to, that of the facility with a known historical cost. For example, not all projects experience economies of scale related to costs. The scale factor that is applied must appropriately reflect both the past projects and the new projects. The scale factor that is used should also be specifically applicable to the range of sizes for the specific type of project being estimated. Also, the parametric approach is based on the assumption that the influence of influencing cost elements on construction costs is still the same for the

new project, but this assumption is no longer valid due to rapid and enormous changes in the way of project development and applied technology.

In addition to the size of the project, the analysis must consider the project scope, its location, and any unique design and site characteristics. Differences in location would almost always require the application of a locational cost adjustment factor. As with regional variations, time differences within past project information should also be considered. The difference in time requires a revision of cost related information due to the inflation rate. Various indexes are used to revise the cost value reflecting the difference of the area or time of the project information. And currently, based on the accumulation of project data, the construction cost index reflecting the project location and the inflation of each period is measured and announced. The published construction cost index indicates the average cost movement of related goods and services in the construction industry. These indexes can be applied to reproduce the estimated cost by reflecting the difference in time and region. Many public organizations or private companies including Association for the Advancement of Cost Engineering, Global Data, IHS Markit, Engineering News Record (ENR), Turner Construction, Chemical Engineering Magazine, and Intratec are estimate cost indexes in different ways and provide cost indexes according to the type of project.

As the size of the project and the number of stakeholder involved increases exponentially, and the way in which they develop the project has become more diverse nowadays, the usability of traditional estimation methods becomes less. The methods and factors mostly introduced were mostly made decades ago, and derived factors based on the data at that time. Therefore, companies are developing new ways of estimating conceptual cost based on these general formulas by applying the project data they have conducted. This database has been produced to provide construction professionals with a conceptual estimating tool that will allow them the ability to create more accurate conceptual cost estimates.

2.2 Issues on Conceptual Cost Estimation

As presented in the practice and literature, the early-stage project cost estimates are made with limited information but should take into account various project influencing factors. The traditional methods can formulate conceptual estimates with less information easily and quickly, but the accuracy of forecasting results are relatively low due to its limited consideration of various project characteristics. When considering the weakness of the traditional method along with the difficulties of conceptual cost prediction, it is not surprising that many cost standards accept a low level of accuracy for the conceptual cost estimation (Association for the Advancement of Cost Engineering International, 2011).

In this chapter, cost overrun issues due to the low accuracy of the conceptual cost estimation is discussed, and the necessity and importance of improving the accuracy and reliability of cost forecasting models are examined. In addition, the inadequate information during conceptual estimation in practice is also described, in terms of understanding the challenging issues to be overcome when developing accurate cost forecasting models.

2.2.1 Cost overrun and Inaccurate Forecasting

Having reviewed the practices of conceptual cost estimation, it is evident that cost overrun is widely reported and chronic problem across international construction project. Many media say excessive cost overrun in some projects and only a small number of projects get delivered under the expected budget (McGraw Hill Construction; Mitigation of Risk in Construction, 2011). The definition of cost overrun is the difference between the actual and estimated costs in the percentage of estimated costs. Actual costs are defined as determined costs at the time of project completion, and estimated costs are set as budgeted, or forecasted, costs when deciding to develop a project. The consequence cost overrun is a critical problem across the construction industry (Hemanta Kumar Doloi, 2011). Since a massive amount of cost is required for the project development, the cost overrun can reduce the company's and industry's competitiveness (Kim, 2005, Kirkham 2014).

Even though contractors identified increasing estimating accuracy as a key strategic issue in their project development, the actual final costs of construction projects still show a significant difference from their originally forecasted costs. According to Flyvbjerg (2006) database on large-scale

infrastructure projects, cost forecast has remained significantly inaccurate for several decades. Even forecasting accuracy has not improved that much over the past decades, despite the improved cost forecasting methodology and quality of the accumulated previous project data (Flyvbjerg, 2006).

To investigate such cost overruns and inaccuracy of cost forecasting of the construction projects, Researchers have taken a multi-faceted approach including psychological and political explanations. First, optimism bias is identified as causal factors of inaccurate cost forecasting from a psychological point of view. The optimism bias has been developed by Kahneman and Tversky (1979) and Lovallo and Kahneman (2003). It is defined as a cognitive behavior identified with most people to view future events in a more positive way than is warranted by experience. Kahneman and Tversky (1979) argue that there is a tendency to underestimate or ignore the presence of past information is perhaps the major source of error in forecasting. And Lovallo and Kahneman (2003) define that prevalent behavior the "planning fallacy," and they argue that it comes from stakeholders having an "inside view." The inside view made forecasts by focusing on the project at present, rather than on the outcomes of similar actions that have already been completed. In construction cost estimation view in the planning phase of new projects. The thought of collecting similar project data and related statistics about the new projects enters a manager's is

not often when taking inside view (Flyvbjerg, 2006). Second, from the point of political explanation, strategic misrepresentation explains the inaccuracy of forecasting project cost. As the construction industry is highly competitive for contract approval and funding securement, estimators deliberately and strategically underestimate project costs and overestimate benefits from project completion in order to increase the probability of project approval. It is commonly shown that clients want less construction cost of the project for their benefit from the project development and contractors may try the reduction of project cost regardless of project scope for low bidding by strategic pressures.

These psychological and political descriptions for the inaccuracy of cost forecasting indicate that something other than technical issues explains the cost overrun across the construction project and inaccurate project cost forecasting. However, psychological and political factors are inevitable, and cost should be estimated for project development and decisions must be made based on the properly estimated costs while considering the optimism bias and strategic misrepresentation. Data based decision making is recommended in the literature which explain problems from the described psychological and political issues. The outside view takes advantage of the utilization of data from past projects or similar projects and the adjustment for the unique characteristics of the project. In this aspect, the American Planning

Association (APA, 2005) encourages planners to use reference class forecasting in addition to traditional methods as a way to improve the accuracy of cost estimation. Reference class forecasting is a method for systematically taking advantage from outside view from past project data. This outside view can prevent estimators from relying on estimating practice. Likewise, even if the technical aspect including quality of the project data and effectiveness of cost forecasting model cannot explain totally inaccuracy of cost forecasting, the technical issues should be clarified to ensure the accuracy of cost estimates to be communicated to decision makers.

2.2.2 Limited Information Availability

As shown in table 2-1 in the previous chapter, the accuracy range of an estimate is dependent upon a level of project definition, estimate input information and the estimating process. The amount and the level of detail in the input information as measured by percentage completion is an important criterion of accuracy. However, the project definition levels in class 5 estimates and class 4 estimates are very low, from 0% to 2% and 1% to 15%, respectively and the amount of information available is quite limited.

In certain extreme cases, estimators do not have any information for estimating. In this case, a past project with similar project characteristics will be a significant time saver for estimators. Even if it is not that extreme situation, before estimating project cost in detail, most estimators at conceptual estimating phases try to make comparisons with historical jobs. Referencing projects completed in the past is a frequent practice adopted by many estimators during every estimating phases due to the lack of related information on the current project. Using historical data of past projects to forecast cost contributes to removing any psychological elements such as optimism bias that may be inherent to the estimator by providing outside view. Reliable cost estimates are required with limited time and resources during

the feasibility stage, for which historical data serves as the backbone of reliable cost estimate as it provides credibility, accuracy, and justification.

Table 2-3. Available General Project Data in Class 5 & 4 Estimates

General Project Data	Class 5	Class 4
Project Scope Description	General	Preliminary
Project Size / Capacity	Assumed	Preliminary
Project Location	General	Approximate
Integrated Project Plan	None	Preliminary
Project Master Schedule	None	Preliminary
Escalation Strategy	None	Preliminary
Work Breakdown Structure	None	Preliminary
Project Code of Accounts	None	Preliminary
Contracting Strategy	Assumed	Assumed

Note: For information on class 3 to 1, see Appendix.

However, due to the inherently complex nature of construction projects, storing completed project's data and accessing detailed information are difficult tasks because of minimal scope definition during the early phase of the project (Hu, Xin et al. 2016). Collection and storage of data from the previous project require a significant amount of time and efforts of project planners, who has a limited resource. From different sources that are

collected, such as construction companies or government organization, they may differ in contents and type of formats; they may require quite a lot of preparation (Gunnar and Zane, 2010). Such collected data may be lacking a standard form across different companies and organizations (Mitchell 1998). Also, in the highly competitive environment of the construction industry (Liu and Ling 2005), construction firms are disinclined to share their proprietary corporate data by which competitors might gain an advantage. Since information is limited during this stage, an organization or individual may end up making an inaccurate cost for building construction projects hence transferring the problem to the construction process (Robinson et al. 2015).

Due to the limited availability of information during the early stages of a project, estimators can prepare rough forecast based on their knowledge over their long experience. Estimators or construction managers typically leverage their knowledge, experience, and estimators to estimate project costs, i.e. they usually rely on their intuition. Even if the experience can reflect the past project, it is evident that the experience based judgment has a significant limitation regarding objectivity and reliability. As a result, the estimated cost is also largely dependent on professional experience and judgment.

2.3 Conceptual Cost Estimation using Machine Learning

Since the 1980s, with the accumulation of past project data, to achieve the benefits of utilizing historical project information for forecasting, developing conceptual cost estimation model using data-driven approach has been a prominent topic in the field of construction economics research. Developing conceptual cost estimation model based on analysis of historical data and utilization of data processing techniques is an important topic in the areas of construction cost management research (Elfaki et al. 2014; Newton, 1991). Estimators and researchers have been working to develop data-driven cost forecasting model that maximize the practical value of historical information available to improve cost estimate accuracy and reliability.

Many approaches using mathematical or statistical techniques have been attempted for analyzing the patterns within project cost database. Statistical methods such as regression analysis were used to find the relationships between project variables and project cost (Lowe, D. J. 2006; Emsley, M. W. and Harding, A. 2006; Hwang, S. 2009; Mahamid, I. 2011). However, the regression-based methods have disadvantages of processing nonlinear and complex relationships among variables and lack of learning process.

With changing computer scientific advances, machine learning techniques including Artificial Intelligence (AI) has been introduced as a solution to analyze numerous and complicated data. With machine learning algorithm, processors interpret project data to learn insights from the complicated relationship between project cost and its features. Machine learning techniques are statistical techniques based on models that explore algorithms that can learn from data without being explicitly programmed (Arthur Samuel, 1959), and make consequent predictions on not yet seen data (Kohavi, 1998).

As machine learning, the application of computational methods underlying experience-based decision making has similar conceptual goals and requirements of cost estimation; Machine learning was widely applied during the last few decades to forecast cost of a new construction project by analyzing patterns within historical project cost data.

2.3.1 Machine Learning Approach

The recent advancement of computer technologies and the relative abundance of data from a growing number of construction projects bring new opportunities for applying machine learning to the construction industry. The machine learning can provide an alternative option for conceptual cost estimation, and many research has shown application and performance of machine learning algorithm including artificial neural networks, k-nearest neighbor, and support vector machine for the various type construction project. Some of those techniques are Case-Based Reasoning, Support Vector Machines, Decision Trees, K-Nearest Neighbors, and Artificial Neural Networks are commonly utilized to forecast the conceptual cost of construction projects in recent researches. There are numerous researches on the utility of the machine learning algorithm for taking the advantages of accuracy increase of the conceptual cost estimation. These researches obtained improvement in the accuracy of the estimation when compared to traditional methods. The table 2-4 and table 2-5 summarize the previous literature using the machine learning based methodology for project cost estimation.

Table 2-4. Summary of Utilized Project Data in Previous Literature

Author (First Author & Published Year)	Applied Methodology	Country	Project Type	Number of Project
Hegazy, T. 1998	ANN	Canada	highway project	18
Adeli H, 1998	ANN	USA	highway project	121
T.M.S. Elhag, 1998	ANN	UK	school	30
Moselhi, O. 1998	ANN	Canada	structural steel building	75
Al-Tabtabai, H. 1999	ANN	Kuwait	highway project	40
Günaydın, H. M. 2004	ANN	Turkey	residential building	30
Kim, G. H. 2004	ANN	South Korea	residential building	530
E.M. Elkassas, 2009	ANN	UK	pipeline project	35
			industrial project	115
			building project	65

Table 2-4. Summary of Utilized Project Data in Previous Literature (Continued)

Author (First Author & Published Year)	Applied Methodology	Country	Project Type	Number of Project
Cheng, M. Y. 2010	ANN	Taiwan	building project	28
Arafa, M. 2011	ANN	Israel	building project	71
Sonmez, R. 2011	ANN	Turkey	continuous care retirement community	20
Dominic D A. 2012	ANN	UK	water-related project	98
Petroutsatou, K. 2012	ANN	Greece	road tunnel	33
Ayedh A. 2013	ANN	UK	building project	20
Hyari, 2015	ANN	Jordan	building & civil project	224
Dursun, O. 2016	ANN	Germany	building project	657

Note: ANN (Artificial Neural Network)

Table 2-4. Summary of Utilized Project Data in Previous Literature (Continued)

Author (First Author & Published Year)	Applied Methodology	Country	Project Type	Number of Project
Petroutsatou, C. 2006	MRA	Greece	road tunnel	33
Lowe, D. J. 2006	MRA	UK	building project	286
Mahamid, I. 2011	MRA	Palestine	public road project	131
An, S. H. 2007	SVM	South Korea	building project	62
Gunduz, M. 2011	Parametric	Turkey	light rail transit and metro track works	16

Note: MRA (Multivariate Regression Analysis), SVM (Support Vector Machine)

Table 2-4. Summary of Utilized Project Data in Previous Literature (Continued)

Author (First Author & Published Year)	Applied Methodology	Country	Project Type	Number of Project
An, S. 2007	CBR	South Korea	residential building	580
Doğan, S. Z, 2008	CBR	Turkey	building project	29
Ji, S. 2010	CBR	South Korea	residential building	124
Koo, C. 2010	CBR	South Korea	residential building	101
Ji, S. 2011	CBR	South Korea	military barrack	129
	CBR	South Korea	residential building	164
Jin, R. 2012	CBR	South Korea	residential building	99
	CBR	South Korea	business facility	41
Choi, S. 2013	CBR	South Korea	public road project	207

Note: CBR (Case-Based Reasoning)

As shown in table 2-4, numerous researches on the utility of the machine learning algorithm for conceptual cost estimation have been presented. Comparative studies between alternative machine learning algorithms are also performed due to increased attention and broad applicability of machine learning algorithm. Emsley et al. (2002) and Sonmez (2004), and Wang and Gibson (2010) tested the accuracy of linear regression and artificial neural networks. And Kim et al. (2004) compared the accuracy of ANN, linear regression and case-based reasoning (CBR). The results of these studies depend on the data used. Some researchers conclude that there is no significant difference (Emsley et al. 2002; Sonmez, 2004) while the other researchers argue that specific algorithm can have better accuracy (Kim et al. 2004; Wang and Gibson, 2010). This review of the comparison research suggests that no consensus has been established with the selection of alternative algorithms. It means that it is necessary to analyze the characteristics of the input data to find a suitable algorithm alternative.

2.3.2 Artificial Neural Network

One of the main difficulties of cost modeling by regression analysis is a determination of a proper model representing the relations between the factors and cost components adequately (Sonmez, R. 2011). Since linear model assumptions and the pre-assumed a linear relationship between the variables implies limitations, linear methods do not guarantee adequate representation of the relationships within variables in complex project data. Neural networks have been used as an alternative to regression analysis due to high applicability when solving the non-linear and complex problem. An alternative approach, as applied in this research, is to use neural networks to establish a mapping function between the input variables and project costs.

Artificial neural networks have been applied to forecast the conceptual cost of a construction project by imitating the human brain. Like many of the researches in the table 2-4 in the previous chapter, neural networks have been used widely and successfully to forecast conceptual cost of various type of projects including civil project (highway project, road project, tunnel project, water-related project) and building project (residential building project, structural steel building project, school project).

Since the 1990s, several researchers have shown the performance and applicability of artificial neural networks for construction cost estimation. Moselhi et al. (1991) and Adeli (2001), respectively presented the overall application of artificial neural networks in the fields of construction management research. It has been widely applied to construction project management for estimating the cost of highway projects (Wilmot et al. 2005; Pewdum et al. 2009) predicting the cost of water and sewer installations (Alex et al. 2010) and building projects (Emsley et al. 2002); mark-up estimation (Li et al. 1999); risk quantification (McKim 1993); and tender price forecast (Boussabaine et al. 1999). Also, many researchers presented that artificial neural networks have a better forecasting performance than the parametric estimation model or linear regression in their research. (Garza and Rouhana, 1995; Creese and Li, 1995; Smith and Mason, 1997; Bode, 1998; Hegazy and Ayed, 1998; Sonmez, 2004; Wang and Gibson, 2010).

In their research, the practicality of cost estimation using artificial neural networks are explained. The artificial neural networks technique has no restrictions on the number of cost variables, because of a self-learning ability of neural networks. Therefore, artificial neural networks eliminate the need for finding a cost estimation relationship that mathematically describes the cost of a system as a function of the variables that have the most effect on the cost of that system.

2.4 Model Complexity Issues of Conceptual Cost Forecasting Model Development

As seen in the previous chapter, many estimators have attempted to estimate conceptual cost based on accumulated project data and researchers have developed a cost forecasting model for estimation conceptual project costs using a machine learning algorithm. In these researches that develop a cost forecasting model based on past project data, it is commonly said that the quality and reliability of the data used for the forecasting model development has a crucial influence on the performance of the prediction model. Poor quality data can cause a significant competitive disadvantage due to unreasonable decision making in the business environment (Redman 1998). Also, the utilization of unreliable information cannot validate the result of the research (De Veaux and Hand 2005). The amount of project data, the number of data attributes, the dimension of the data, and the relationship within data attributes are affecting the complexity of forecasting model. In general, model complexity can be defined as a function of the number of parameters in a given predictive model, as well as whether the chosen model is linear, nonlinear, and so on. The more parameters a model has, the more complex the model is. Complex models are generally known to be less easily interpreted, at greater risk of overfitting.

In particular, machine learning algorithms are often black box models, which cannot explain the process of producing forecasts. Therefore, the previous models are too often developed without due consideration given to the effect that the choice of input variables has on model complexity, learning difficulty, and performance of the subsequently trained forecasting model. In-depth analysis of input data is necessary as the performance of the forecasting model is profoundly affected by the input data. In this aspect, this chapter focuses on the complexity of input data used for related researches using machine learning, and discuss the issues that need to be addressed to improve the performance of the conceptual cost forecasting model.

2.4.1 Lack of Project Data

Previous literature related to data-driven conceptual cost estimation has commonly approached empirical research (ex past facto research) using historical project data (Fellows and Liu, 2009). In such studies, at first, the necessary data for the conceptual cost estimation were collected from past construction projects according to the purpose and scope of forecasting. And the project dataset should be created based on collected from past projects, which consists of historical project information project characteristics including project type, project size, etc. Then, the data set is used to develop a cost estimation model by using an appropriate method of analysis such as neural network, decision tree, support vector machine, and case-based reasoning. The developed models are validated using the test set of data.

As the conceptual cost forecasting model relies heavily on historical information of previous project information, it is assumed that there is enough data compared to the complexity of the forecasting model. The machine learning algorithm requires project data as much as possible to be collected to ensure the ability of that data to adequately reflect the project and the accuracy of forecasting. Enough amount of information related to the conceptual cost estimation needs to be collected and analyzed. However, as

shown in the following graph, as the level of detail or expected prediction accuracy of forecasting model increases, exponentially increasing amount of data is the required. On the other hand, as the level of detail of the data or expected prediction accuracy of forecasting model increases, the amount of data that can be collected becomes smaller. As mentioned in Chapter 2.2, the level of detail of the data that can be obtained in the early stages of the project is low. And in the case of more detail data, it is difficult to obtain data because of a lack of project experience or difficulty in sharing information within the competitive construction industry. Also, the quantity and quality of the collectible information may vary, depending on its source; organization, project characteristics. In this regard, limited and uncertain information makes it difficult to formulate a reliable conceptual cost estimation model with the required amount of data for an algorithm to execute in the conceptual cost forecasting model.

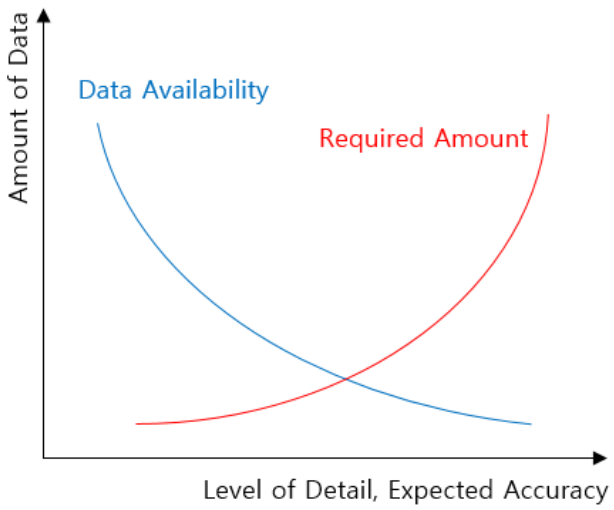


Figure 2-2. Required Data Amount and Availability of Data

As shown in table 2-5, especially in South Korea, there are many types of research to predict the construction cost by using case-based reasoning (CBR). CBR is a heuristic method based on the assumption that similar projects have similar characteristics and results and solves the problems by recognizing project similarity between a new project and past cases. CBR is used widely to deal with construction-related issues including conceptual cost estimation because it is suitable for situations where the number of project case is enough for developing case base, and the project characteristics are similar for different projects such as public residential building, apartment building.

However, because there is no previous project exactly same with a new project, the solutions applied to the past project may not work for the new project, particularly where there is not enough number of cases constructed in the past. In the case of a project such as EPC projects or skyscraper projects, the number of similar projects is limited, and even the same type of project has a large difference in project characteristics including project cost. If there are few similar projects and a similar project has few similar features, it is not appropriate to use CBR, and there are not many related examples. In the aspect of project data availability, few prior studies have attempted conceptual cost estimation for construction projects with a limited number of projects, such as power plant project, oil & gas project, and high-rise building project. It is difficult to find valid results through statistical methods or machine learning algorithms due to the lack of similar case data compared to the general construction project. Also, even though it is a project to create the same type of facilities, the project requirements and scope vary greatly according to the characteristics of the project, such as regions, clients, delivery methods, and other financial characteristics.

The table 2-4 and 2-5 also shows that most researchers tried to develop a prediction model using the projects collected from a single country. However, in case of the type of project with a limited number such as such as power plant project, oil & gas project, and high-rise building, collecting

project data from a single country can be challenging to produce a conceptual cost estimation model that yields meaningful forecast results.

Thus, the amount and detail of data required by the cost forecasting model differ from the data that can be obtained. Nevertheless, it is necessary to collect as much data as possible to improve the performance of the forecasting model. If the amount of detailed level data in the prediction model increases, the complexity of the prediction model may increase. In the next chapter, we will discuss issues that can arise as the complexity of the prediction model increases.

2.4.2 Attribute Selection and Characteristics

Performance of data-driven cost forecasting models depends on not only on the amount of data mentioned above but also on the properties of input data. (Rueda-Benavides and Gransberg, 2014). When analyzing the attributes of data, each attribute should be treated differently according to its importance. Previous researchers have attempted to select influential attributes primarily based on the authors' discretion, via interviews with experts, regression analysis, and other methods (Bell and Bozai, 1987; Setyawati et al. 2002; Gunaydin and Dogan, 2004). Some researchers also use principal component analysis (PCA) to identify principle component of collected data and to avoid problem regarding multicollinearity among variables. Also, Gardner et al. (2016) highlighted and tried to quantify the efforts of data collection by attributes and investigate the minimum amount of input data for reasonable and reliable estimates.

There are also some studies that develop cost forecasting models by estimating the weights according to the influence of the input variables. Genetic algorithms have been widely used to optimize the weights of the input variables and other parameters included in the forecasting model. In these studies, selection and optimization methods of the input variables are

investigated, and these methods show the capability of improving the performance of the forecasting model. Reducing the number of attributes in input data and assigning weights to each attribute can be the solution to reflect the relative importance of project information, but they would not be identical across projects and estimation methodology. Not only each property information has a different impact on project cost, but also the method of calculating weights is also tricky to use for other forecasting models using a different algorithm.

Likewise, the attribute selection approaches can be easily flawed because intuition based selection is difficult to reflect the influence and characteristics of each attribute. There are two general methods, forward and backward stepwise selection. In forward stepwise selection, predictors are added to the model one at a time starting at zero predictors, until all of the predictors are included. Backwards stepwise selection is the opposite, and involves starting with a model including all predictors, and then removing a single predictor at each step. The feature selection method of reducing the number of attributes (by choosing more informative attributes) can lessen the collecting efforts of unnecessary data and prevent redundant data from entering the prediction model; it does not guarantee both reliable level of accuracy of the model and adequate explanatory power of the data.

Gardner et al. (2016) investigate the correlation relationship between the accuracy of the prediction model and the number of input variables. The increase in the number of attributes in the conceptual cost forecasting model may result in better forecasting accuracy. When examining this pattern identified, it indicates that an appropriate number of attributes are necessary for the acceptable level of accuracy, and the explanatory power of many attributes has a positive effect on the performance of the cost forecasting model. To solve these problems mentioned, there is need to develop a more flexible conceptual cost forecasting, capable of being flexible enough to be applied in various type of projects regardless of the number and type of input data attributes.

Table 2-5. Summary of Input Variables in Previous Literature

Author (First Author & Published Year)	Project Type	Number of Quantitative Input Variables	Number of Qualitative Input Variables
Hegazy, T. 1998	highway project	2	7
Adeli H, 1998	highway project	2	0
T.M.S. Elhag, 1998	school	4	9
Günaydın, H. M. 2004	residential building	4	4
Kim, G. H. 2004	residential building	5	4
E.M. Elkassas, 2009	pipeline project	10	4
	industrial project	10	4
	building project	10	4
Cheng, M. Y. 2010	building project	6	4
Arafa, M. 2011	building project	6	1

Note: Quantitative variables do not include project cost, which is dependent variables

Table 2-5. Summary of Input Variables in Previous Literature (Continued)

Author (First Author & Published Year)	Project Type	Number of Quantitative Input Variables	Number of Qualitative Input Variables
Sonmez, R. 2011	continuous care retirement community	9	11
Dominic D A. 2012	water-related project	5	5
Petroutsatou, K. 2012	road tunnel	10	0
Dursun, O. 2016	building project	9	6
Petroutsatou, C. 2006	road tunnel	9	0 (※)
Lowe, D. J. 2006	building project	6	0 (※)
Mahamid, I. 2011	public road project	9	0 (※)
Gunduz, M. 2011	light rail transit and metro trackworks	18	0 (※)

Note: Quantitative variables do not include project cost, which is dependent variables

※: Qualitative variables are excluded due to methodological limitations of regression analysis and parametric method.

Table 2-5. Summary of Input Variables in Previous Literature (Continued)

Author (First Author & Published Year)	Project Type	Number of Quantitative Input Variables	Number of Qualitative Input Variables
An, S. 2007	residential building	4	5
Doğan, S. Z, 2008	building project	4	4
Ji, S. 2010	residential building	8	0 (※)
Koo, C. 2010	residential building	6	4
Ji, S. 2011	military barrack	9	8
Jin, R. 2012	residential building	10	0
	business facility	10	0
Choi, S. 2013	public road project	10	6

Note: Quantitative variables do not include project cost, which is dependent variables

※: Qualitative variables are excluded due to methodological limitations of regression analysis and parametric method.

In addition, depending on the type of the attribute, the attribute should be appropriately processed during the modeling process. Each of the variables needs to be analyzed to determine the best way of representing relationship within project data. The variables used for developing conceptual cost forecasting model can be roughly divided into quantitative and qualitative variables. The quantitative variables are numeric variables such as project cost, project duration, project size (area, height, length, and width), and other project related numbers. Where the range of these variables differed by the unit of variables, it is appropriate to standardize the value to ensure that the range of values was more evenly distributed. The remaining qualitative variables are categorical variables that represent one of a choice of project characteristics. The qualitative variables are generally treated as a series of binary variables by applying binary input coding.

Categorical variables such as project region, project type, delivery method are necessary information for conceptual cost modeling. The qualitative variables are essential at an early phase of project development (RunZhi Jin et al. 2014) when the quantitative variables are not sure at this stage. The quantitative variables may be determined at the initial phase of the project, but there may be uncertainties that can be changed as the project progresses due to changes in the requirements of the clients, value engineering, and design changes.

As found in Table 2-6 that summarizes the number and type of input variables used in the previous data-driven cost estimation research, categorical variables were not used in regression-based forecasting models and parametric methods. Also, there is a tendency to use more qualitative variables than quantitative variables. To increase the application of the qualitative variables, there have been studies to develop revision method of the predicted values by reflecting the characteristics of qualitative variables in CBR (RunZhi Jin et al. 2014), but this method is not applicable to other prediction methodologies.

Looking at the data attributes used in the research, there are many variables related to the size of the project such as area, height, the number of floor and capacity of facilities. In the case of the civil project including road and bridges, the variables related to the length and width of road or bridges are usually collected. And in the case of a residential building or public apartment, the project database consisted of variables such as the number of households, number of rooms in single households, building area of each floor, the number of the parking lot, and so on.

It is also reasonable to suspect that there is a high probability that a multicollinearity problem will occur when developing a prediction model based on the collected data. Nevertheless, most of the studies did not focus

on the multicollinearity problem comes from the linear relationship between the properties. The multicollinearity issue can be ignored according to the purpose of the prediction, but it is a problem that should not be overlooked when considering the effects of multicollinearity problems on predictive models. The multicollinearity issue will be discussed in more detail in the next chapter.

2.4.3 Overfitting

In the field of cost estimation research, researchers strive to collect as much data as possible for increasing the amount of learning of cost forecasting model. They tried to increase both the number of projects and the number of data attributes that included in the learning data. On the other hand, increasing the number of data attributes leads to an increase in data complexity of learning data. As model complexity increase (by adding more detailed attributes), the model will have a lower error by fitting the training data, this is a fundamental property of statistical models. But it possibly will increase the prediction error on new data at the same time. As described in figure 2-3, model complexity contributes to training error and test error in different ways. Typically, the error decreases with increasing model complexity, but since the model is more likely to be over-fit for a fixed training set. This is a problem with overfitting, which can be a key threat to validate the model. Overfitting causes training optimism about a performance of the model, and it can result in inaccurate forecasting of new data sets. The training optimism is a measure of how much worse our model does on new data compared to the training data.

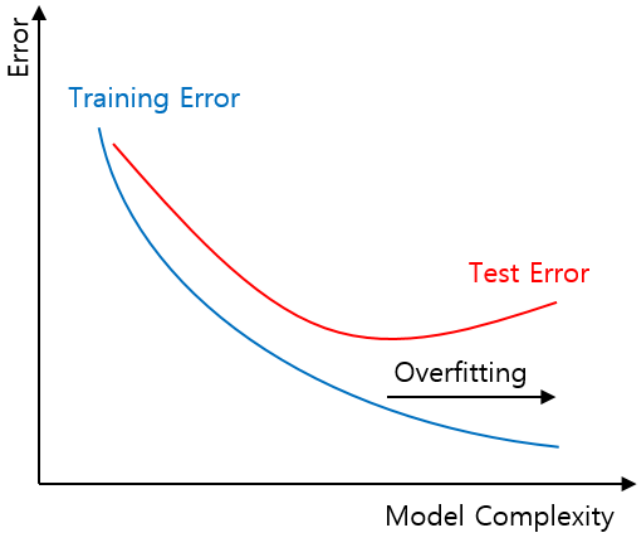


Figure 2-3. Model Complexity Contributing to Error

In previous chapters, the limitations of the reduction the data attributes, as the number of data attributes increases as the accuracy of the prediction model increases and the model has more explanatory power. It is a logical conflict with the overfitting problem described in the previous paragraph. This problem cannot be said to be right in one way or another, and the complexity of the model, the explanatory power of the model, and the accuracy of the model must all be considered.

To address this problem, understanding how different sources of error lead to bias and variance is necessary to examine the behavior of the forecasting model. The error due to bias is taken as the difference between

the expected (or average) prediction of our model and the correct value which we are trying to predict. And the error due to variance is taken as the variability of a model prediction for a given data point. Addressing bias and variance is about investigating the over-fitting problem. Bias is reduced, and variance is increased about model complexity. As more and more parameters are added to a model, the complexity of the model rises, and variance becomes our primary concern while bias steadily falls. Therefore, the performance of the forecasting model should consider both the accuracy (bias) and stability (variance).

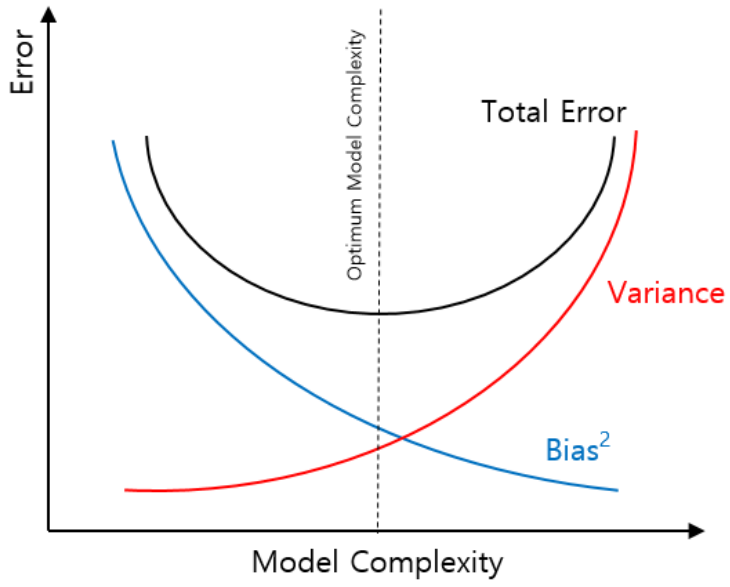


Figure 2-4. Bias and Variance Contributing to Total Error.

As shown in the graph above, there will be an optimal level of model complexity that minimizes errors due to bias and variances. If the forecasting model complexity exceeds this optimal level, a possibility of overfitting the prediction model exists; while if the model complexity falls short of the optimal complexity level, the model is under-fitting. In practice, there is not a standardized way to identify the level of optimal model complexity. Instead, the prediction error should be measured accurately, and various levels of model complexity should be investigated to identify the level of model complexity that minimizes the overall error of cost forecasting model.

2.4.4 Multicollinearity

As mentioned briefly in the previous chapter, attributes of the project are often highly correlated and such correlation leads to multicollinearity of the input data and it may significantly distort the forecasting results. Multicollinearity refers to the presence of correlation relationships between the predictors in a dataset. It indicates that changes in one variable are associated with shifts in another variable. The stronger the correlation, the more difficult it is to change one variable without changing another. It becomes difficult for the model to estimate the relationship between each independent variable and the dependent variable independently. The presence of multicollinearity may be due to the scarcity of data samples or is inherent in the data collected.

Multicollinearity reduces the precision of the estimate coefficients, which weakens the predictive power of the forecasting model. As shown in the following graph, high multicollinearity will inflate the standard error of estimates of the predictors, and thus decrease the reliability of the forecasting model.

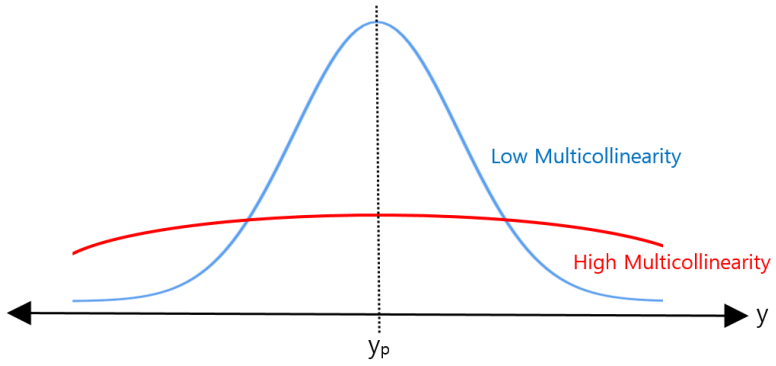


Figure 2-5. Distribution of Predicted Values according to Multicollinearity

The pre-processing statistical data techniques such as principal component analysis (PCA) or factor analysis (FA) are used to prevent this problem (Manly, 2005). The literature indicates that the multicollinearity problem is described in the field of statistics, but there is little attention in the areas of machine learning. Multicollinearity problems have been addressed less in the fields of machine learning studies. This is because the machine learning algorithm itself cannot automate the selection of relevant predictors, and the purpose of developing a forecasting model is to focus on the accuracy of the forecasting. However, the machine learning algorithms are not robust against the multicollinearity problem. This implies that the multicollinearity problem should be appropriately solved when using the machine learning algorithm.

Likewise, the complexity of the cost forecasting model is caused by various factors including the amount of project data, the number of data attributes, the characteristics of attributes, and multicollinearity within data and affects the performance of the prediction model in various ways. However, little attention has been paid to the effects of model complexity related to when developing the conceptual cost forecasting model. The model complexity tends to be ignored in the process of seeking only the accuracy of the forecasting model. In order to make conceptual cost forecasting model more reliable and robust, it is important to address the issues of data characteristics, because the dimension and complexity of the data can affect the accuracy and robustness of the prediction model. As mentioned before, model complexity issues cannot be said to be right in one way or another, and it should be treated as much an opportunity as it is a problem to improve the performance of the conceptual cost estimation model.

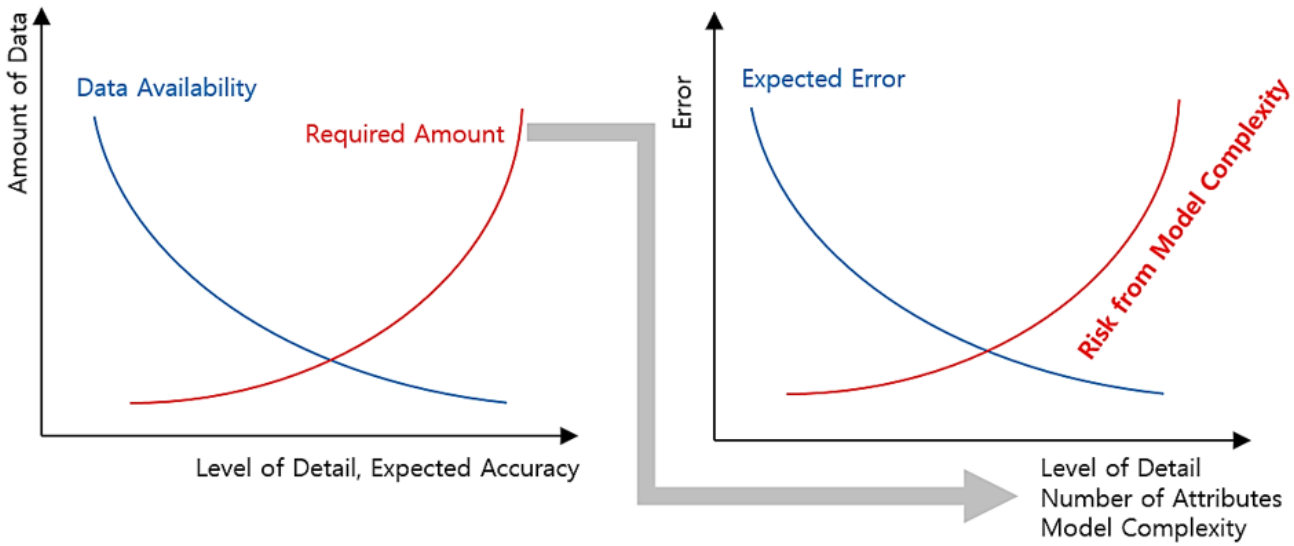


Figure 2-6. Risk of Model Complexity

2.5 Summary

In the first section, the role and importance of conceptual cost estimation are addressed to explain the significance of this research area. As decisions made during this stages of the project development can result in more significant consequences and conceptual cost estimation is a fundamental element for the initial appropriation and economic feasibility studies of a project planning, developing conceptual cost estimation model is an important topic in construction cost management research. And the traditional estimating methods for the conceptual estimation including the unit price method, cost capacity method, cost index method and parametric estimation method are introduced. Also, the application and limitation of the traditional estimation methods are also reviewed.

After the investigation of the existing practice of conceptual cost estimation, issues faced during the conceptual estimation process is also described in the next section. Current conceptual cost estimation practice still experiences a prevalent case of significant difference between estimated cost and actual cost. In the previous literature, the psychological factor 'optimism bias' and political factor 'strategic misrepresentation' are explain the cost overrun and inaccurate budgeting phenomenon. Even though, the

performance issues of cost forecasting model cannot explain totally inaccuracy of cost forecasting, the technical performance of conceptual cost estimation should be studied to ensure the reliability of cost related decision making. From this review of explanation of inaccurate forecasting of conceptual cost, the necessity and importance of improving the accuracy and reliability of cost forecasting models are established. In addition, the issues with the inadequacy of information during conceptual estimation in practice is also discussed, in terms of understanding the challenges to be overcome when developing accurate cost forecasting models.

Next section, related previous research on conceptual cost estimation are investigated. To achieve the benefits of historical project data utilization and advancement of data analysis techniques, researchers have developed a conceptual cost estimation model using machine learning algorithm. Case-based reasoning, k-nearest neighbor, support vector machine, decision tree, and artificial neural networks are applied to forecast the construction cost of the various type of projects. Among them, the application of artificial neural networks to predict the conceptual cost of the construction project are reviewed in detail. Due to the advantage of being able to effectively grasp the nonlinear relationship between variables without regard to the number of input variables, artificial neural networks were applied to cost forecasting model at conceptual stages of project development.

Having reviewed previous literature on data-driven conceptual cost estimation research, several issues with model complexity are identified. Although many researchers have gathered past project data to create a database to generate a cost forecasting model, there is still insufficient data to be learned to estimate the cost of the project. In particular, it is difficult to find a developed case of data-driven cost forecasting models for projects that lack past examples. In addition, there are many types of research related to the selection of influencing input variables. Considering the explanatory power of the input data and the forecasting model, reducing the number of input variables does not guarantee the performance of the forecasting model. Contrary to this, when the amount of data and the number of attributes increase, overfitting and multicollinearity problems arise. Therefore, the accuracy of the forecasting model should be analyzed in relation to the complexity of the developed model. In order to make conceptual cost forecasting model more reliable and robust, it is important to appropriately address the issues of model complexity can affect the accuracy and robustness of the prediction model.

Chapter 3. Model Development

Based on the review of existing research presented in the previous chapter, this chapter provides an explanation of the conceptual cost estimation model development. First, the conceptual framework describes the comprehensive of the purpose of the developed model along with a list of required functions for the model development. Then, a detailed description of the model development process is provided in the next section. The model development process including data preprocessing, artificial neural networks modeling, ensemble method selection, and factor analysis. The detail explanation of methodologies used for developing the conceptual cost estimation model is provided in another section. Artificial neural networks, ensemble modeling including bootstrap aggregating and adaptive boosting are explained as learning methodology for analyzing construction project data. And factor analysis method including principal component analysis, multiple correspondence analysis, and the combined algorithm is described as a tool for investigating model dimension and complexity.

3.1 Model Development Framework

When developing a conceptual cost forecasting model to forecast the cost of the construction project accurately, this research attempts to investigate the issues as mentioned earlier of model complexity including project data scarcity, attributes selection, overfitting and multicollinearity. To address these problems, this research proposes to develop a conceptual cost forecasting model incorporating artificial neural network, ensemble modelling, and factor analysis. The artificial neural network learns the nonlinear relationship of data. Ensemble modeling can help to increase the amount of learning and to solve the problem of overfitting. Factor analysis can find the optimal level of model complexity avoiding overfitting and solve multicollinearity problems. The combination of these three methodologies will serve to solve each of the problems mentioned above. The following figure shows the overall methodological approach of the proposed model corresponding to the model complexity issues as mentioned earlier.

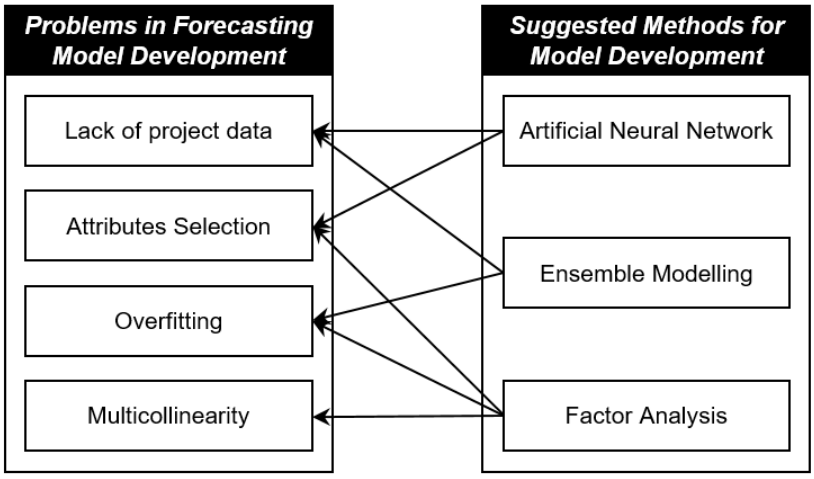


Figure 3-1. Methodological Approach of Model Development

3.1.1 Purpose and Process Overview

The purpose of the conceptual cost forecasting model developed in this research is forecast the cost of the construction project, with rough preliminary project information, even when the level of details from the information provided is meager. In order to function this purpose, the proposed model is developed to have two main functional components; data preprocessing including factor analysis, forecasting model development including neural networks and ensemble modeling. The proposed model is developed to learn the input project data by using an ensemble method based on artificial neural network and to obtain the predicted value for the test data set. The following figure 3-2 summarizes an overall process and components of the proposed model.

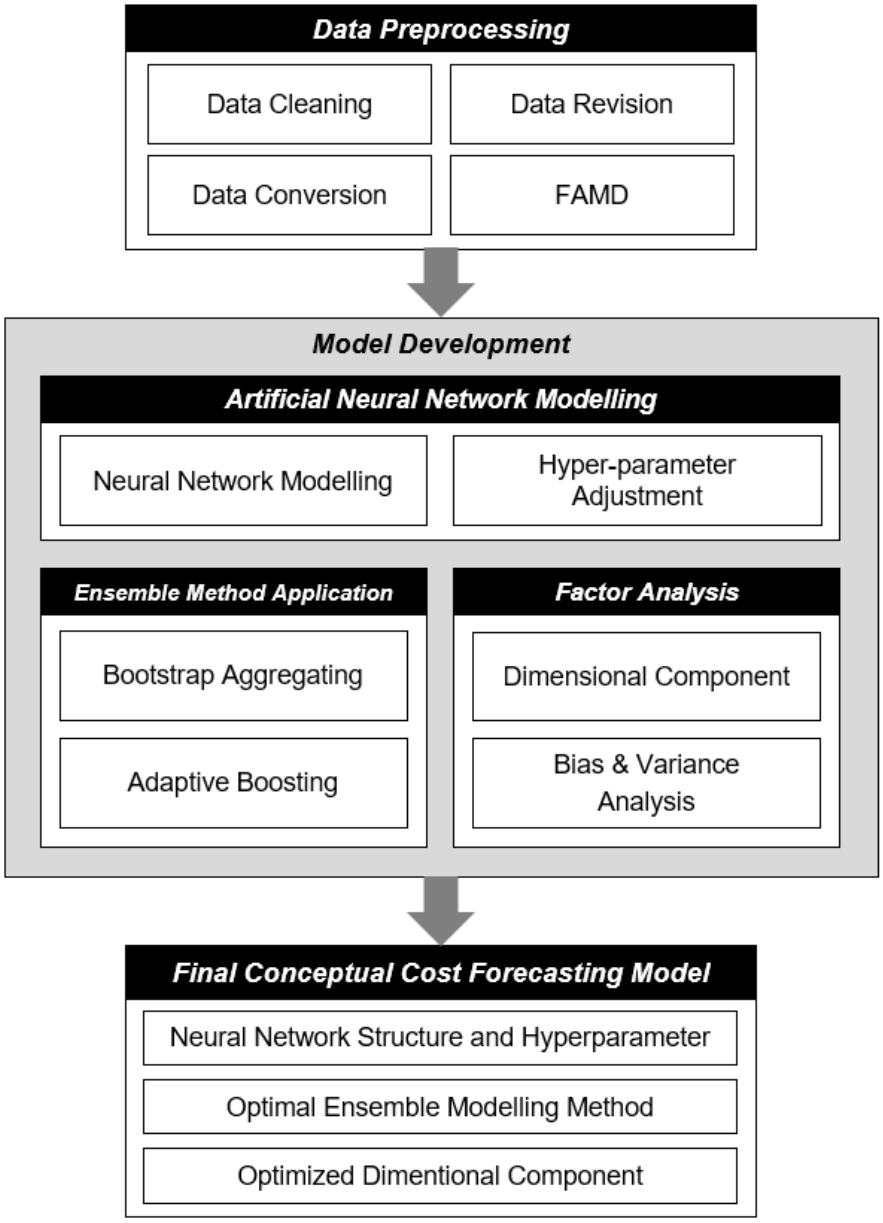


Figure 3-2. Model Development Process

In order to function the above purpose, the following data processing and analysis procedure needed to be executed. At first, the input data is preprocessed so that it can be processed first in the prediction model. This work includes data cleaning, data revision, and data conversion. To address issues of the model complexity and to derive the optimal forecasting model, we perform a factor analysis of input data. For analyzing dimension of input data, factor analysis of mixed data (FAMD) algorithm is applied to generate data projected to a new orthogonalized dimension. Then the orthogonalized data are learned by ensemble method based on artificial neural network. During this process, the hyper parameters of the artificial neural network in the forecasting model are tuned according to the input data for improving accuracy. The accuracy of the proposed model is measured by calculating the difference between the predicted value and the actual value for the test set. Finally, the accuracy of the prediction model is analyzed according to the number of components of the analyzed dimension, and a model is having the best accuracy and explanatory power is derived. The following figure shows the overall data processing process of the proposed model.

3.1.2 Data Preprocessing

As mentioned a lot in previous chapters, the quality of the input data determines the quality of the machine learning based forecasting model. Therefore, data preprocessing has been addressed by many researchers as a first key stage of data preparation for applying machine learning methods. Data quality can be defined as accuracy, relevance, and completeness of data and the data preparation involves enhancing and enriching the quality of the data concerning their intended purpose (Rajagopalan and Isken, 2001). There is a preliminary work of data preprocessing procedures including data cleaning, conversion, integration, reduction, and discretization (Yu, 2007). Due to its importance, researchers spend much time and efforts on data preparation and data preprocessing (Soibelman and Kim, 2002; Zhang et al. 2003). They highlighted its fundamental importance by estimating that data preparation and data preprocessing takes approximately over 60% of the total data analysis work.

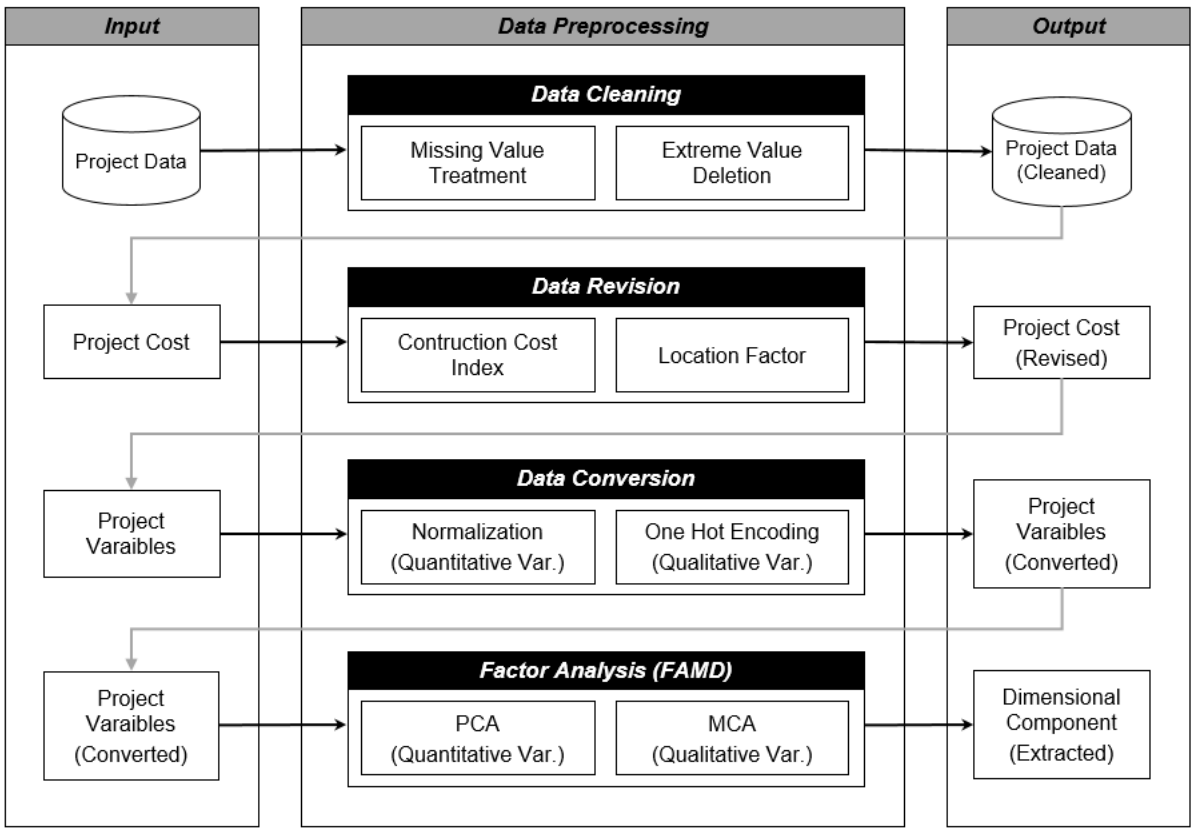


Figure 3-3. Data Preprocessing Procedure

Data Cleaning

For the first stage of data preprocessing, data cleaning is performed by removing data that is considered an outlier. Data cleaning process also involves removing the missing values in the database. Listwise deletion method is applied when removing missing value. This method has the disadvantage of reducing the quantity of the data. But as the developed conceptual cost forecasting model requires project data with the same number of attributes, the listwise deletion method is applied.

And outliers and extreme values are detected based on interquartile ranges. The outliers and extreme values in project data identified by applying the following equation are removed.

$$Q_3 + OF \times IQR < x \leq Q_3 + EVF \times IQR$$

$$\text{or } Q_1 - EVF \times IQR \leq x < Q_1 - OF \times IQR$$

$$Q_1 = 25\% \ quartile$$

$$Q_3 = 75\% \ quartile$$

$$OF = outlier \ \ factor \ \ (= 3.0)$$

$$EVF = extreme \ \ value \ \ factor \ \ (= 6.0)$$

Data Revision

As the projects were distributed in the timeline, the project cost should be standardized reflecting the cost escalations. As explained in chapter 2.1.2 cost index method is applied to revise the project cost information. The revising process is illustrated in the formula below.

$$C_b = C_a \times \left(\frac{I_b}{I_a}\right)$$

$$C_b = \ project\ cost\ at\ time\ b(current)$$

$$C_a = \ project\ cost\ at\ time\ a(past)$$

$$I_b = cost\ index\ at\ time\ b(current)$$

$$I_a = cost\ index\ at\ time\ a(past)$$

Appropriate construction cost index should be used to reflect the project type and the inflation of each period. In this research, Cost index data are available from a variety of sources are used. In this research, the Chemical Engineering Plant Cost Index, which is announced by Chemical Engineering magazine, is used as a tool for adjusting plant construction costs. And for skyscraper project data, local construction cost indexes were used to convert the construction costs of the buildings projects to the specific base of each country. All the projects' costs were converted to costs on the base year of 2010. Each country had a particular base city for which the costs were converted to New York City for the USA, Beijing for China, Dubai for UAE,

Sydney for Australia and Singapore as Singapore. Lastly, for apartment building project data and office building project data in South Korea, Korean construction cost index is used, which is published by the Korea Institute of Construction Technology. Using this index provided, the project cost information adjusted from one period to another based on the equation above.

Data Conversion

Input data must be converted to be appropriately processed in a cost forecasting model. As project data consists of a mixed type of data, the data need to be treated in a different way according to the type and characteristics of each attribute. Standardization and one hot encoding are applied to quantitative variables and qualitative variables, respectively.

For quantifiable variables, normalization or standardization is necessary to convert the collected data into standardized values since all attributes should be analyzed under identical standards because of different measurement scales of each attributes (Koo et al. 2010). Variables such as project cost, project size or project capacity must be adjusted from values measured on different scales to a common scale. Multi-layer perceptron in the artificial neural network is also sensitive to feature scaling, so it is highly recommended to scale the input data. For standardization of quantifiable variables, the standard score is calculated as the following formula. The

standard score is the signed number of standard deviations by which the value of a data point is above the mean value of what is being observed or measured.

$$z = \frac{x - \mu}{\sigma}$$

z = standard scores, μ = mean, $\sigma =$ standard deviation

Computing a z-score requires knowing the mean and standard deviation of the entire population to which a data point belongs as shown in the formula introduced. The calculated standard score is a dimensionless value obtained by subtracting the mean of the population from an individual raw score and then dividing the difference by standard deviation of the population. The result of standardization (or Z-score normalization) is that the features will be rescaled that they'll have the properties of a standard normal distribution that the mean is 0, and the standard deviation is 1. These standardized values allow the comparison of corresponding standardized values for different attributes and datasets in a way that eliminates the effects of certain gross influences.

For qualitative variables including nominal, ordinal and categorical variables, it is necessary to be converted into a numerical format that could be provided to machine learning algorithms. The machine learning algorithms require all input variables and output variables to be numeric. And

the proposed forecasting model needs to reflect information available in conceptual stages, which are categorical information such as project type, project region, project funding status, project delivery methods or structural types of project. Therefore, these categorical variables must be transformed into a binary value by applying one hot encoding. The one hot encoding is a process by which categorical variables are converted into a binary form that could be provided to machine learning algorithms. The one hot encoding scheme expresses each item of the categorical variable as 0 and 1 value. An example of this approach can be seen in the following figure.

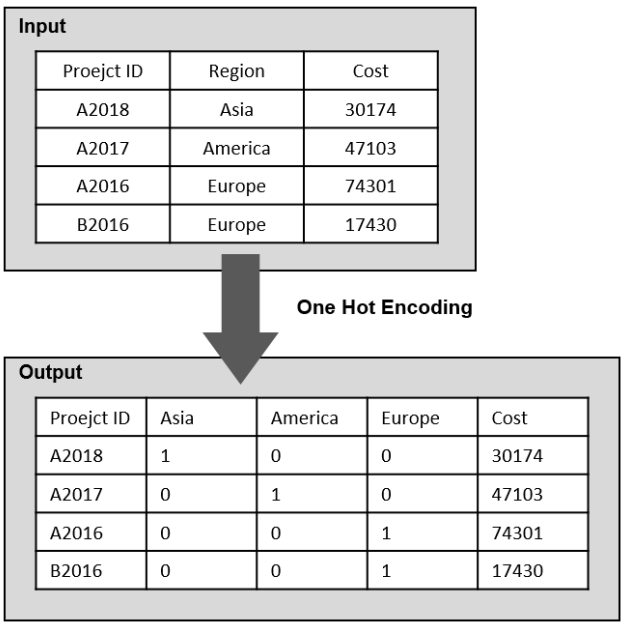


Figure 3-4. Explanation of One Hot Encoding

Factor Analysis

In order to analyze the performance of the forecasting model according to the complexity of the input data, this research attempts to apply the factor analysis. Factor analysis in data preprocessing stage refers to changing the data to be input to the prediction model into a new coordinate system using the factor analysis of mixed data (FAMD) algorithm. For analyzing dimensionality of data, orthogonalization should be performed to make transformed input variables independent. For the orthogonalization, the input data is processed by FAMD algorithm. More detailed information of FAMD is in the next chapter 3.2.

The dimension component variables are linear combinations of the original variables and independent of each other. So the use of the dimension component variables will not generate a multicollinearity problem while having a capability of data explanation. The orthogonalized dimension component variables have disadvantages of difficult interpretation. It should be acknowledged as a problem and a limitation when developing the black box model such as artificial neural networks. Since the artificial neural networks are used for proposed cost forecasting model in this research, the explanation of the variables is a limitation, and we will exclude the interpretation of individual variables in the remaining dissertation.

3.1.3 Forecasting Model Development

In order to develop a cost forecasting model based on the preprocessed data described in the previous chapter, the input data is learned based on an artificial neural network algorithm. The forecasting model is developed through the procedure of artificial neural network modeling, hyperparameter adjustment, ensemble modeling method application, and prediction model accuracy analysis by dimensions of input data. The following figure shows the overall process of the proposed forecasting model development. The modeling process involved investigating the performance of different ensemble methods and effects of model complexity on cost forecasting model.

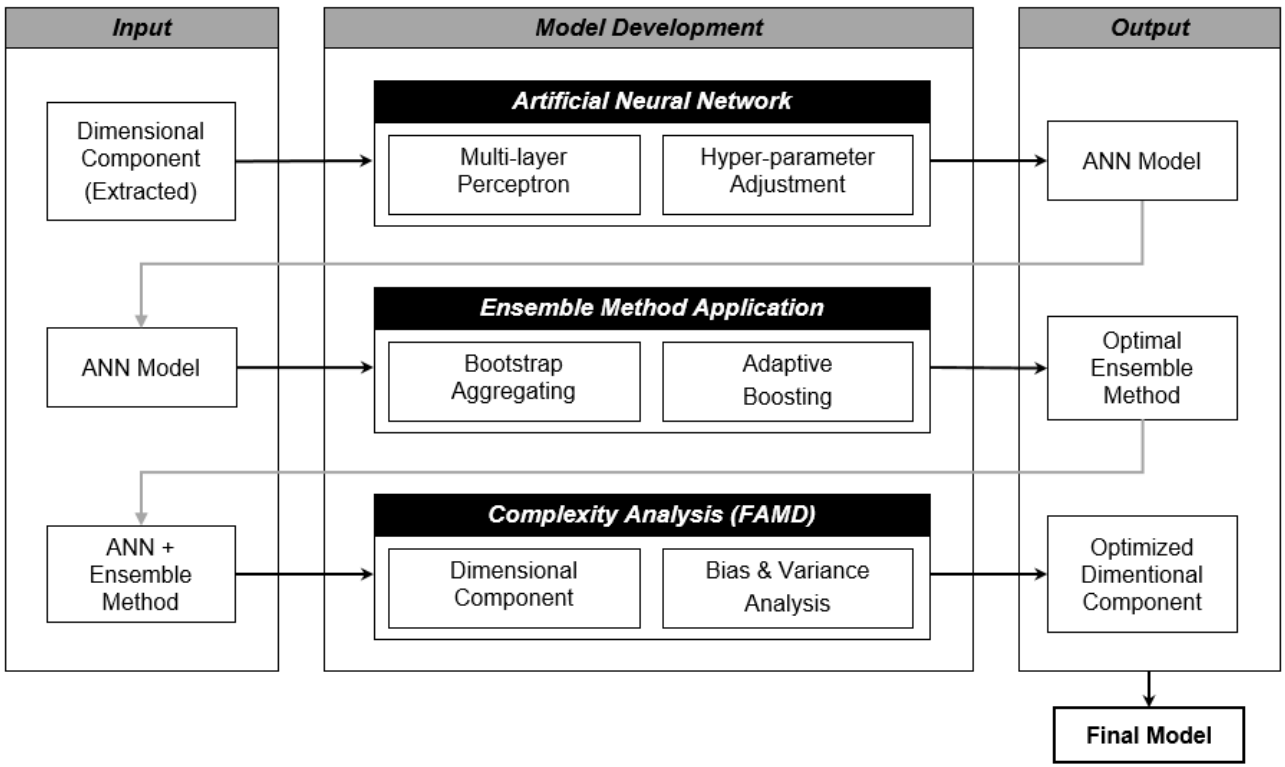


Figure 3-5. Forecasting Model Development Procedure

Artificial Neural Network Modeling

The proposed conceptual cost forecasting model has been designed to include an input layer corresponding to the project attributes of input data and an output layer of one processing element as the project cost to predict. Multi-layer perceptron is implemented to train using back-propagation with no activation function in the output layer, which can also be seen as using the identity function as the activation function. Therefore, it uses the square error as the loss function, and the output is a set of continuous values. Multi-layer perceptron is a supervised learning algorithm that learns a function by training on a given dataset.

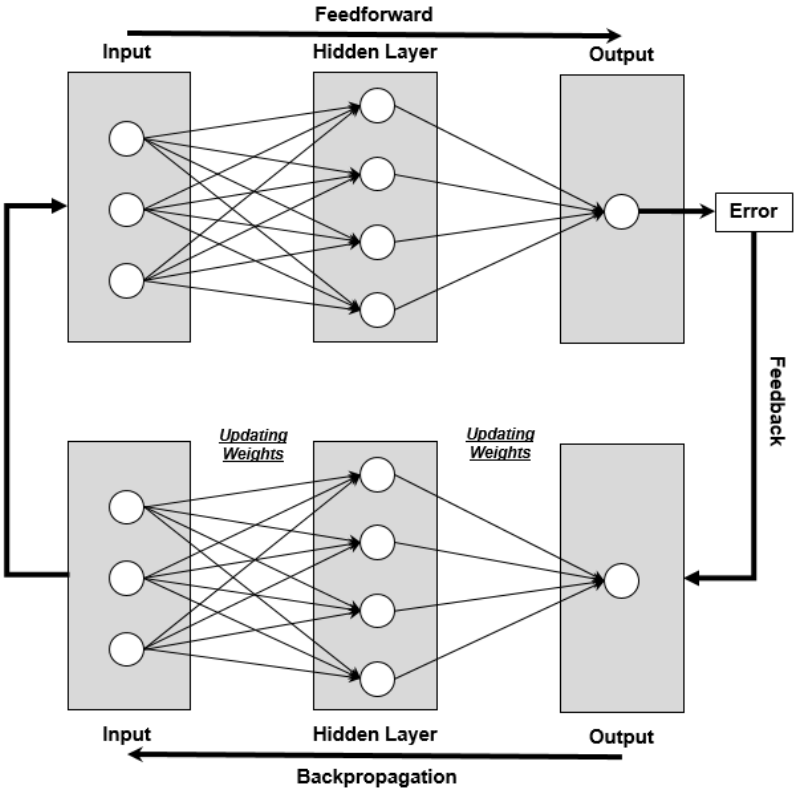


Figure 3-6. Back-propagation Algorithm

Hyperparameter Adjustment

Most machine learning algorithms involve hyperparameters which are variables set before actually optimizing the model's parameters. Setting the values of hyperparameters can be seen as model selection, choosing which model to use from the hypothesized set of possible models. Neural networks can have many hyperparameters, including those which specify the structure of the network itself and those which determine how the network is trained.

Multi-layer perceptron also requires tuning a number of hyperparameters such as the number of hidden neurons, layers, and iterations. Among them, the hidden layer extracts and remembers the important features from the input project data to predict the target values of the output layer (Rafiq MY et al. 2001). Therefore, a valid number of the hidden layer should be determined.

Many research has explained the problems with practicality associated with defining the optimal network size and set parameters of the artificial neural networks (Rafiq MY et al. 2001; Setyawati BR et al. 2002) so that they require an amount of time with trial and error. Hyperparameters are often set by hand, or determined by some search algorithm such as coordinate descent, grid search, random search and model-based methods since there is no rule to determine the optimal hyperparameters (Shtub A et al. 1999; Setyawati BR et al. 2002). This research determines hyperparameters of cost forecasting model using grid search. But due to reducing the computing efforts, the range of parameters for grid search is determined by trial and error process based on the measurement of model performance.

Ensemble Method Application

Due to limited information available at the early phase of project and difficulty of collecting a number of project data, ensemble modeling techniques are applied to the proposed conceptual cost forecasting model to increase the amount of training for fewer project data and alleviate the complexity of the model. As the ensemble methods designed to improve the stability and accuracy of machine learning algorithms to minimize the errors, the performance of the forecasting model can be improved. Bootstrap aggregating and adaptive boosting is applied for artificial neural network model, created by learning input project data. The two ensemble modeling methods have different sampling methods and have different effects depending on the characteristics of the input data. In the proposed model, both the ensemble modeling methods are tested for the artificial neural network model made from the input data and select a model with better performance.

Complexity Analysis

The proposed model attempts to apply the factor analysis. In order to analyze the performance of the forecasting model according to the complexity of the model. In the proposed model development process, the data converted by the FAMD algorithm is used as input data to the cost forecasting model. As a result of applying the FAMD algorithm to the project data, dimensional components are derived in the order that best describes the project data. The dimension component consists of the weights of the individual input variables. Each component has an inertia value, which means that the larger the value, the better the description of the input data.

The performance of the cost forecasting model by varying the dimension components from two to ten (depending on the number of attributes the input data, the number of dimensional components can be bigger or smaller). The model performance is evaluated at each dimensional component. The test error is used to determine the best model. This involves the development of a forecasting algorithm that simultaneously finds the minimum model error and optimal level of model complexity.

3.2 Methodology Description

3.2.1 Artificial Neural Network

Artificial neural networks approach is an analogy-based, non-parametric computational information processing system that has an information processing procedure similar to a biological human brain's neural networks (Anderson et al. 1992). They have characteristics of the biological neural networks; the ability to learn from the experience, to generalize based on the given information, and to extrapolate results for new datasets (Haykin 1994). Artificial neural networks provide the capability to solve problems without the benefits of an expert and the need for a selection of data. This approach is especially appropriate for complex, hard-to-learn problems where no formal underlying theories or classical mathematical and traditional procedures exist (Adeli, 2001) by identifying patterns in data that are not obvious (Anderson and McNeill, 1992). Artificial neural networks are fundamentally different from statistical methods like a regression in one way- they learn inductively by examples, and they are able to generalize solutions (Flood et al. 1994). Other data-driven approach mentioned in chapter 2, including regression analysis, case-based reasoning and fuzzy logic analysis, find it difficult addressing problems such as imprecision, incomplete and uncertainty of data and other variables affecting costs and implicit

combinatorial effects and inter-relationships of cost variables (Flood and Kartam 1994), areas where artificial neural networks is often at its best. On the positive side, neural networks demonstrate many features that are coherent with the nature of cost estimation. Major strengths include their ability to self-organize knowledge from training data, their ability to generate results from incomplete information and their ability to cope with complex relations. These resemble the way human estimators develop expertize through experience.

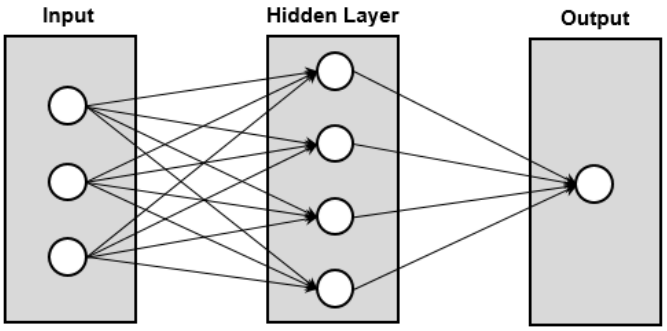


Figure 3-7. Structure of Artificial Neural Networks

The most common form of learning is the back propagation method, which is a supervised learning method (Setyawati, Creese and Sahirman 2003). Describing the learning process in order, the input data set is presented to the network in its input layer. These are then transferred to the hidden layer by some form of activation function, usually a linear activation function.

Weights are assigned randomly to the input values in the hidden layer and then their cumulative weighted values transferred to the output layer. If the training algorithm adopted is a supervised one, the result of the training, called the output, is compared to the expected real value at the output layer and the error, measured by the difference between the output and the predicted value, is calculated. This feedback is sent to the network, and an error function is used to minimize the amount of the error in the next training cycle.

3.2.2 Ensemble Modeling

As described in chapter 2.4.3, the main causes of error in machine learning are due to bias and variance. Ensemble methods designed to improve the stability and accuracy of machine learning algorithms to minimize the errors. The combinations of multiple models decrease variance, especially in the case of unstable models, and may produce a more reliable classification than a single classifier. In this research, the ensemble modeling techniques are expected to increase the amount of training for fewer project data and alleviate the complexity of the model.

In the case of small amount of project data to be learned, this research attempts to apply the ensemble method for developing cost forecasting model. The ensemble is a machine learning idea in which the concept is to learn with multiple data sets and models using the same machine learning algorithm (Breiman, 1996). The ensemble method by sampling techniques are roughly divided into two; bootstrap is aggregating and boosting. Random forest and gradient boosting decision tree are the ensemble methods can be applied to decision tree method. Decision trees are not used in this study, so these two ensemble methods will not be introduced. Returning to the description of the ensemble method to be used in this study, both bootstrap aggregating and

boosting start with creating N new training data sets by random sampling with replacement from the original training set. When sampling with replacement, every sample is returned to the data set after sampling. So a particular data point from the observed data set could appear zero times, or more in a given bootstrap sample (Sonmez, 2011). Then, N models are obtained by learning from the N training set. This process is described in the following figure. As a result of bootstrap aggregating, the aggregation of the models is made by voting in case of categorical attributes, and by averaging the predictions in case of numerical attributes (Wang et al. 2012).

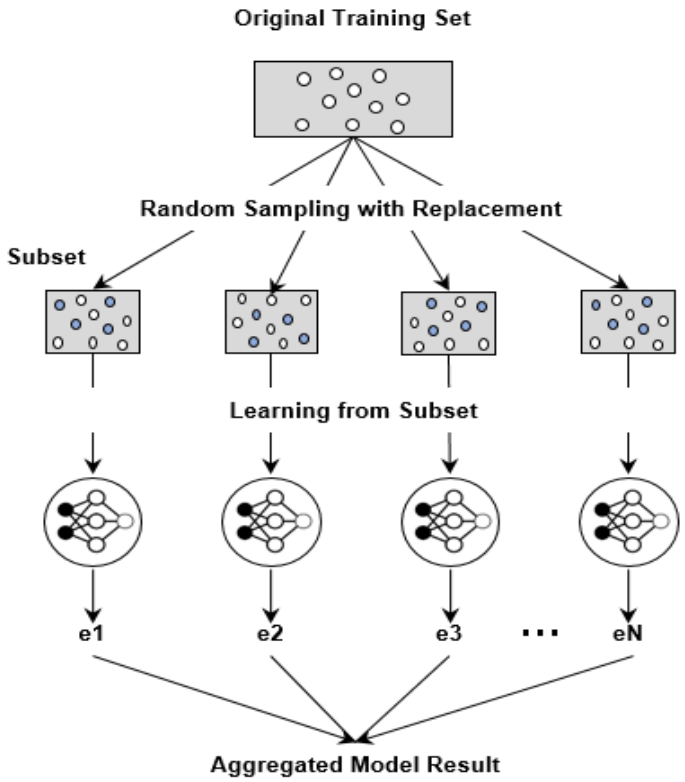


Figure 3-8. Process of Bootstrap Aggregating

The purpose of bootstrap aggregating is to mimic the process of sampling observations from the population by resampling data from the observed sample (Efron & Tibshirani, 1993). Bootstrap sampling is commonly used to establish a level of uncertainty for the estimated parameters. The bootstrap method can also be utilized to improve the prediction performance of neural networks when sparse data is available for

training (Tsai & Li, 2008). It is expected that the ensemble method which increases the amount of learning with a small number of data can improve the performance of conceptual forecasting model positively. This can be expected to have a better effect on the conceptual cost estimation phase with insufficient data as mentioned in Chapter 2.

In the case of bootstrap aggregating, any element has the same probability to appear in a new data set by sampling with replacement. This means the training stage is parallel and each model is built independently for bootstrap aggregating. However, for boosting the observations are weighted, and therefore some of them will take part in the new sets more often. Boosting algorithm sequentially builds the new model. In boosting algorithm, each classifier is trained on data subset, taking into account the previous classifiers' success. After each training step, the weights are redistributed. Misclassified data increases its weights to emphasize the most difficult cases. In this way, subsequent learners will focus on them during their training. In the training stage with a boosting algorithm, the algorithm allocates weights to each resulting model. A model with good a classification result on the training data will be assigned a higher weight than a poor one. So when evaluating a new learner, boosting algorithm needs to keep track of learners' errors, too.

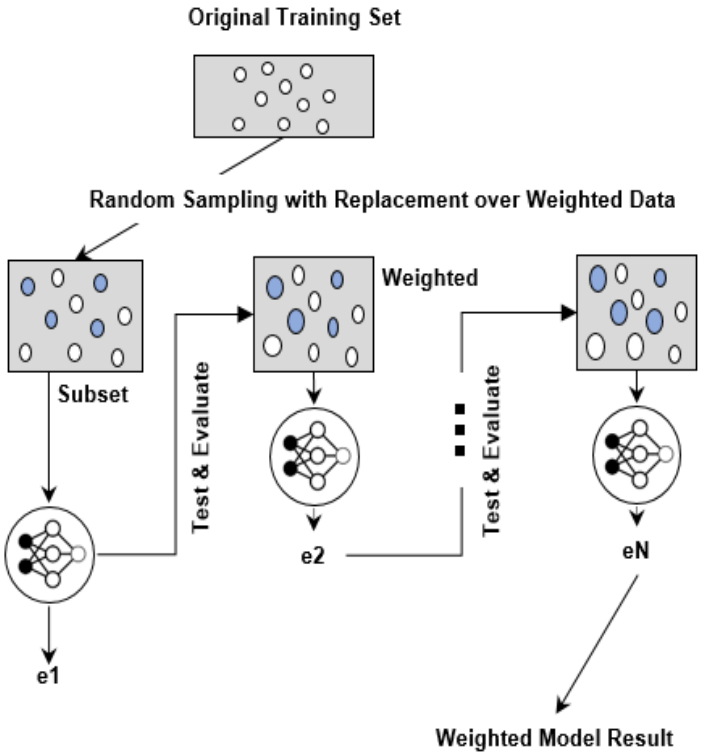


Figure 3-9. Process of Adaptive Boosting

Some of the boosting techniques include its extra-condition to keep or discard a single model when testing the result of the single model. For example, in adaptive boosting, the most renowned and planned to be tested in this research, an error less than 50% is required to maintain the model. Otherwise, the iteration is repeated until achieving a learner better than a random guess. Adaptive boosting can be used in conjunction with many other types of learning algorithms to improve performance. The output of the

different learning algorithms ('weak learners') is combined into a weighted sum that represents the final output of the boosted classifier. Adaptive boosting is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers (Freund, Y. and Schapire, R. E. 1996).

Both bootstraps aggregating and boosting method decrease the variance of your single estimate as they combine several views from different models. The individual models can be weak, but as long as the performance of each one is slightly better by learning from the randomly sampled subset, the final model can be proven to converge to a strong learner. Therefore, the result of both methods may be a model with higher accuracy and stability. If the problem is that the single model gets a very low performance, bootstrap aggregating will rarely get a better bias due to its averaging effect. However, boosting could generate a combined model with lower errors as it optimizes the advantages and reduces pitfalls of the single model. By contrast, if the difficulty of the individual model is over-fitting, then bootstrap aggregating is the better option. Boosting for its part doesn't help to avoid over-fitting; in fact, this technique is faced with this problem itself. Boosting is sensitive to noisy data and outliers. In some issues, it can be less susceptible to the over-fitting problem than other learning algorithms. For this reason, bootstrap aggregating is useful more often than boosting. To this end, when considering

overfitting and model dimension issues described in the chapter 2.4.3. The bootstrap aggregating and boosting method should be applied according to the characteristics of the data and the purpose of the situation applied. In this study, the performance of each method will be compared and analyzed by the complexity of the input data.

3.2.3 Factor Analysis of Mixed Data

In order to analyze the performance of the forecasting model according to the complexity of the input data, this research attempts to apply the factor analysis of the data learned by the forecasting model. For analyzing dimensionality of data, orthogonalization should be performed to make transformed input variables independent. For the orthogonalization, the input data is processed by factor analysis of mixed data (FAMD) algorithm.

FAMD is the factorial method dedicated to data in which a group of individuals is described both by quantitative and qualitative variables. FAMD works as a principal component analysis (PCA) for quantitative variables and as multiple correspondence analysis (MCA) for qualitative variables. Mathematically FAMD looks for the function on me the more related or most correlated to all K quantitative and Q qualitative variables. The input data include K quantitative variables k=1 to K, and Q qualitative variables q=1 to Q. The quantitative variables and qualitative variables are standardized during the analysis in order to balance the influence of each set of variables. And z is a quantitative variable follow the process described in the following equation.

$$\gamma(z,k) = \ the\ correlation\ coefficient\ between\ var.k\ and\ z$$

$$\eta^2(z,\ q) = \ the\ squared\ correlation\ ratio\ between\ var.z\ and\ q$$

In the principal component analysis of K, we look for the function on I (a function on I assigns a value to each). It is the case for initial variables and principal components the most correlated to all K variables in the following formula.

$$\sum_k \gamma^2(z,k)$$

In multiple correspondence analysis of Q, we look for the function on I more related to all Q variables in the following formula.

$$\sum_k \eta^2(z,q)$$

In FAMD {K,Q}, we look for the function on me the more related to all K+Q variables in the following formula.

$$\sum_k \gamma^2(z,k) + \sum_k \eta^2(z,\ q)$$

As a result of applying FAMD, both types of variables play the same role. Existing individual variables in original data contribute to the components of the new dimension with different weights. The components that can represent existing data are linear combinations of individual variables in existing data.

3.3 Summary

In this chapter, this study developed a conceptual cost forecasting model incorporating artificial neural network, ensemble modeling, and factor analysis. The purpose of the conceptual cost forecasting model developed in this research is to forecast the cost of the construction project accurately. Also, this research attempts to investigate the issues of the model complexity. In order to function this two purpose, the proposed model is developed to have functional components; data are preprocessing including factor analysis, neural networks development, and ensemble modeling.

First, the data processing is essential to the process of developing data-driven cost forecasting model. The four steps of data preprocessing are presented including data cleaning, data revision, data conversion, and factor analysis. Factor analysis in data preprocessing stage refers to changing the data to be input to the prediction model into a new coordinate system using the FAMD algorithm. And the proposed cost forecasting model consists of several functional components, including artificial neural networks modeling, ensemble method selection, and factor analysis. The combined use of the artificial neural network, ensemble method, and factor analysis is tested to obtain the optimal structure of neural networks, optimized hyper parameter

setting, optimal ensemble method, and an optimal number of dimensional components for improving the performance of the developed model.

The detail explanation of methodologies used for developing the conceptual cost estimation model is also provided in the next section. Artificial neural networks, ensemble modeling including bootstrap aggregating and adaptive boosting are explained as learning methodology for analyzing construction project data. And factor analysis method including principal component analysis, multiple correspondence analysis, and combined algorithm (FAMD) are explained as a tool for investigating model dimension and complexity.

Chapter 4. Model Verification

The verification and validation procedure of a cost forecasting model usually limited within the set of project data used for the test. This chapter conducts comparative experiments with several types of project data to verify and validate the proposed concept used for developing conceptual cost estimation model. The data set include combined cycle power plant project (C1) and high-rise building project (C2).

The first experiment tests the effects of utilization of the ensemble methods. Each accuracy of the developed forecasting model is measured according to whether or not to use the ensemble method including bootstrap aggregating and adaptive boosting. And the next experiment tests the performance of the conceptual cost forecasting model according to the dimensions of the data. The dimension component of data is formulated by a factor analysis method for the mixed type of project data. As shown in the figure below, both experiments were conducted in the order of case database construction, data preprocessing, factor analysis, forecasting model development, and results analysis.

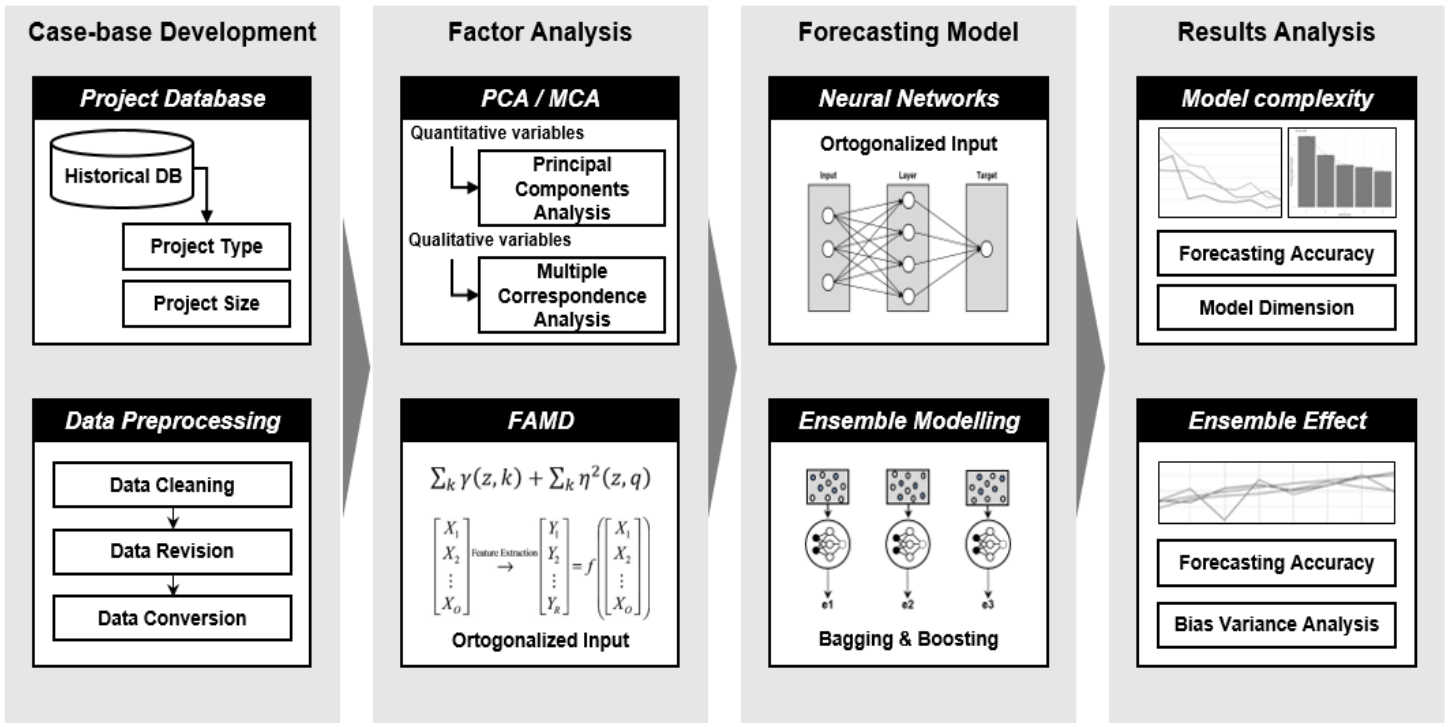


Figure 4-1. Model Verification Process

4.1 Ensemble Modeling Method

In the previous chapter, ensemble methods are explained to provide the capability of improving the stability and accuracy of machine learning algorithms. The combinations of multiple models from sampling method decrease variance, especially in the case of unstable models, and may produce a more reliable performance than a single model. The ensemble method can also be utilized to improve the forecasting performance of neural networks when data availability is limited for training. It is expected that the ensemble method can enhance the performance of the forecasting model by increasing the amount of learning with a small number of data.

4.1.1 Experimental Design

To carry out an experiment for investigating the effects of ensemble modeling methods applied to artificial neural networks, the following hypothesis are established, and the validity of the hypothesis is verified through experiments. The hypothesis is the accuracy of cost forecasting model using artificial neural networks can be improved by applying ensemble modeling methods. To verify the hypothesis, this paper conducts a comparative experiment on cost forecasting model using artificial neural

networks based on two different ensemble modeling method; bootstrap aggregating and adaptive boosting. Hypothesis testing will proceed through the T-test or the ANOVA test. Due to different random weight initializations that can lead to different validation accuracy, the iterative analysis is needed to obtain valid results. The cost forecasting model is developed using combined cycle power plant projects, high-rise building project, and government office building project. The number of bootstrap aggregating sampling is 500. And the error rate is calculated through a random subsampling method.

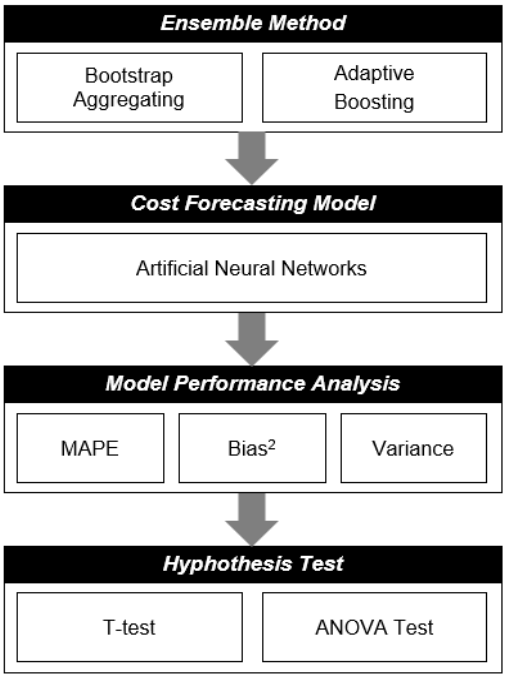


Figure 4-2. Experimental Design for Ensemble Method

Mean absolute percentage error (MAPE) is used to measure the accuracy of the forecasting model. Mean absolute percentage error (MAPE) is calculated to estimate error between the predicted value and the actual value for the test set. MAPE is a popular measure of prediction accuracy of a forecasting method in statistics to express accuracy as a percentage. The following formula defines the MAPE; the difference between actual value and forecast value is divided by the actual value again, and the absolute value of this calculation is summed for every fitted or forecasted point in time and divided again by the number of fitted points n.

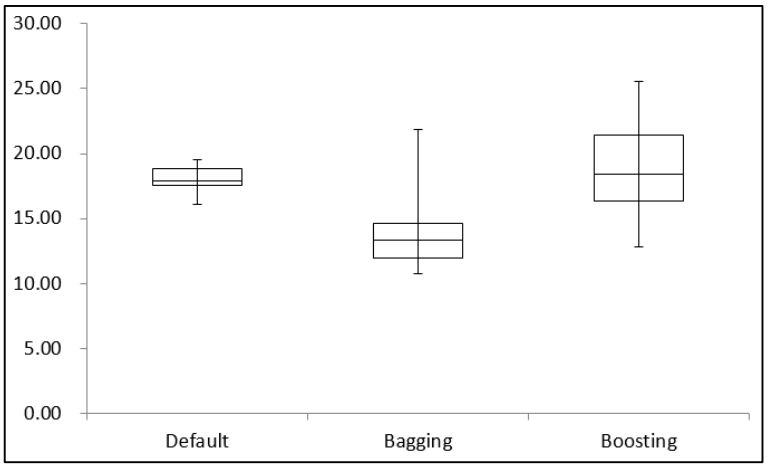
$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_T}\right| \times 100$$

$$A_T = \text{actual value}$$

$$F_t = \text{forecast value}$$

4.1.2 Experiment Results

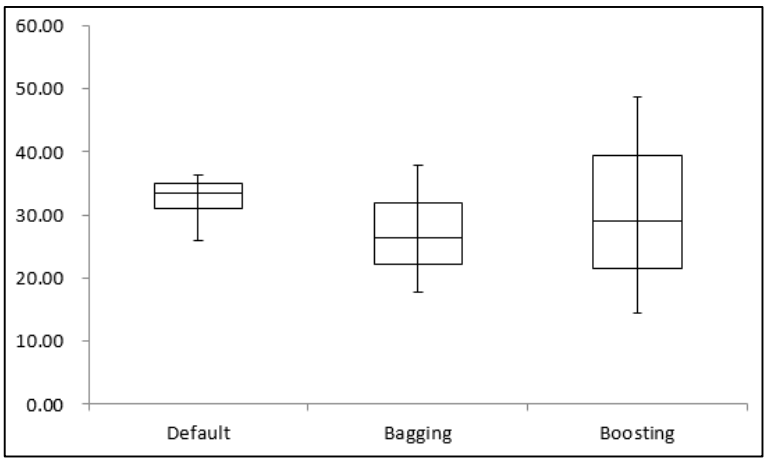
When examining the result from cost forecasting model for the combined cycle power plant, as shown in following figures (4-2) and table (4-1), mean absolute percentage error (MAPE) of forecasting models demonstrates noticeable features by applying ensemble modeling methods. In addition to MAPE, mean squared error (MSE), bias, and variance shows distinct value from the forecasting model not applying ensemble modeling. First, examining the hypothesis that established for this experiment, the application of ensemble modeling improves the performance of the artificial neural network based prediction model. There was a statistically significant difference in the accuracy of the predictive model when comparing and not applying the system.



Note (a): Default refers to not using ensemble modeling

Note (b): Bagging refers to bootstrap aggregating

Figure 4-3. Boxplot of MAPE by Ensemble Modeling Application (C1)



Note: Bagging refers to bootstrap aggregating

Figure 4-4. Boxplot of MAPE by Ensemble Modeling Application (C2)

Table 4-1. T-test Result of Bootstrap Aggregating Application (C1)

Category	Only ANN	ANN + Bootstrap Aggregating
Mean of MAPE (%)	17.86	14.11
Variance of MAPE	1.71	11.22
t - score	3.29	
p - value	0.0065	

Note: threshold chosen for statistical significance = 0.05

Table 4-2. T-test Result of Bootstrap Aggregating Application (C2)

Category	Only ANN	ANN + Bootstrap Aggregating
Mean of MAPE (%)	32.64	26.75
Variance of MAPE	10.68	47.78
t - score	2.32	
p - value	0.02	

Note: threshold chosen for statistical significance = 0.05

As described in figure 4-3, high-rise building project has relatively larger values of MAPE due to its data characteristics (large variance on cost and limited attributes). Nevertheless, the effect of the bootstrap aggregating application was shown in this cost forecasting model for the high-rise building project at conceptual stages.

However, unlike the cases of applying the method of the bootstrap aggregating, the adaptive boosting method did not show a valid difference, in both combined cycle power plant project and high-rise building project (figure 4-2 and 4-3). Also, a result of T-test states that the effect of the adaptive boosting method is not significant in this model in both cases (table 4-3 and 4-5). Even the MAPE value of model with adaptive boosting application displays bigger value than original model in combined cycle power plant case. When examining the large variance values, the boosting method seems to have lower robustness of forecasting.

Table 4-3. T-test Result of Adaptive Boosting Application (C1)

Category	Only ANN	ANN + Adaptive Boosting
Mean of MAPE (%)	17.86	18.75
Variance of MAPE	1.71	14.75
t - score	-0.69	
p - value	0.5033	

Note: threshold chosen for statistical significance = 0.05

Table 4-4. T-test Result of Adaptive Boosting Application (C2)

Category	Only ANN	ANN + Adaptive Boosting
Mean of MAPE (%)	32.64	31.00
Variance of MAPE	10.68	149.71
t - score	0.39	
p - value	0.7074	

Note: threshold chosen for statistical significance = 0.05

4.1.3 Discussions

Results of application of bootstrap aggregating confirm the ideas of ensemble methods being able to improve the accuracy of estimates. In both type of project (power plant and skyscraper) area, the not common building construction project that the accumulated project case itself is relatively rare and the characteristics of each case and the deviation of project cost are significant. As mentioned earlier in chapter 2 and 3, the ensemble modeling was expected to be useful in forecasting the cost of these type of project, and the application of bootstrap aggregating satisfied this expectation.

On the other hand, adaptive boosting cannot demonstrate even statistically significant. To investigate the low performance of adaptive boosting, other error-related metrics such as MSE, bias, and variance are compared with the default model and model with bootstrap aggregating. For both two types of project, bias is reduced as ensemble methods applied. Unlike a theory, however, bootstrap aggregating and adaptive boosting shows different performance on reducing variance in both datasets while biases are reduced in both data sets. As mentioned before, adaptive boosting increases the variance in both datasets. Therefore, it is necessary to look at bias and variance by model complexity (model dimension). As seen in the next

verification experiment, the effects of ensemble modeling methods vary through the number of dimensional components of data.

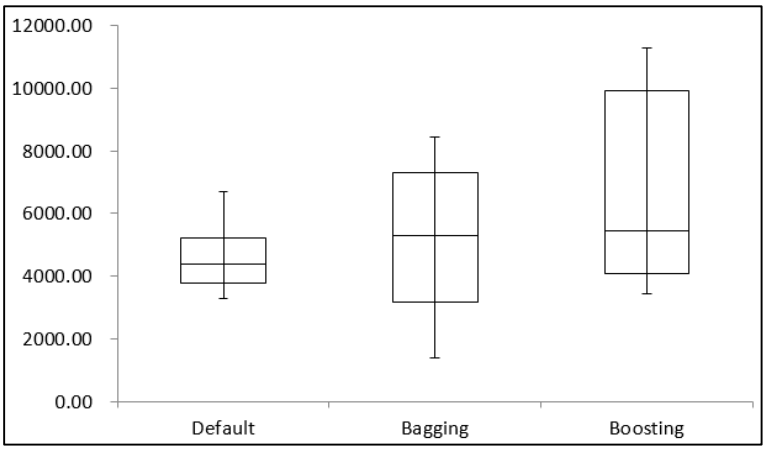


Figure 4-5. Comparison of Variance (C1)

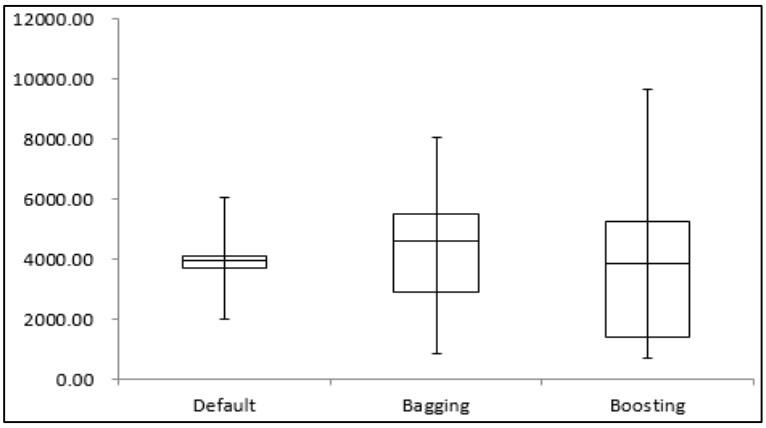


Figure 4-6. Comparison of Variance (C2)

4.2 Factor Analysis

In this research, the effects of model complexity and multicollinearity are investigated when developing conceptual cost forecasting model. These issues have not been addressed enough in the previous study focused on the accuracy of the forecasting model. Under limited availability of project data at conceptual estimation stages, estimators and researchers tried to collect as much data as possible for increasing the amount of learning of cost forecasting model. However, increasing the number of data attributes leads to an increase in data complexity, which results in an over-fitting problem. On the other hand, a reduction the data attributes can have a negative impact on the accuracy of the forecasting model. In order to suggest a solution to this contradictory problem, this research attempts to apply factor analysis method to developing a forecasting model. In this verification section, the performance of cost forecasting models is investigated according to the number of dimensional components extracted by factor analysis of input data.

4.2.1 Experimental Design

To carry out an experiment for investigating the effects of model complexity on cost forecasting model, the performance of developed cost forecasting models is analyzed according to the number of dimensional component of input data. The dimensional component of data is created by applying FAMD (factor analysis of mixed data) to input data. In order to verify the effect of model complexity on the cost forecasting model, the data of combined cycle power plant projects are utilized, and two previously introduced ensemble methods are applied in this experiment. The number of bootstrap aggregating sampling is 500. And the error rate is calculated through a random subsampling method. Mean absolute percentage error (MAPE) is used to measure the accuracy of the forecasting model,

4.2.2 Experiment Results

As can be seen in the following figures (4-4), the performance of the forecasting model is measured when the model dimension is changed from 2 to 10. The x-axis refers to the number of dimensional components, and the dotted line is the cumulative sum of the inertia values of each dimensional component derived from the factor analysis.

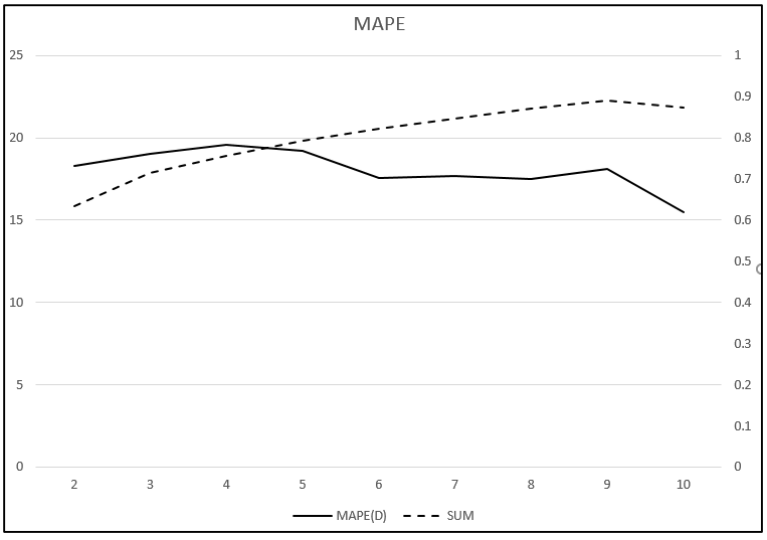


Figure 4-7. MAPE by Dimensional Component (C1)

This value can reflect the explanatory power of the forecasting model to input, as the number of dimensional components increases, this value also increases. This is similar to increasing the complexity of the model as the

number of attributes of the input data increases. In the transformed data by applying FAMD algorithm, as the number of components increases, the features of the original data contributing to each component will be extracted more. Therefore, the accuracy of the forecasting model will be improved by reflecting more characteristics of the original data. Although these results do not show any significant reduction of error, it can be inferred that the error of the prediction model decreases.

4.2.3 Discussions

As shown in this experiment, not only the accuracy of the artificial neural network model, but also the effect of bootstrap aggregating and adaptive boosting methods varies corresponding to the number of dimensional components involved. It is necessary to investigate and understand the bias and variance regarding model complexity. First, the performance of the forecasting model according to the ensemble method differ by the number of dimensional components involved.

As shown in figure 4-5 below, the error rate of bootstrap aggregating (red line) tends to decrease overall as the number of dimensional components increases. The black line is an error rate of the artificial neural network prediction model without ensemble methods. Unlike this tendency of bootstrap aggregating (blue line), however, the error rate of adaptive boosting application decreases at the beginning but then increases again. In the previous experiment, the variance values when the adaptive boosting method was seen larger than the variance values of default condition and bootstrap aggregating application, and it can be concluded that the adaptive boosting method does not reduce variance, unlike the theory.

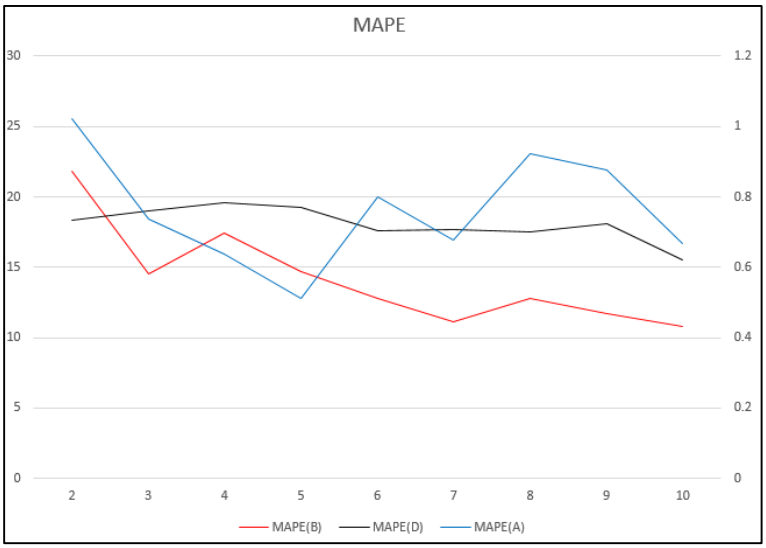


Figure 4-8. MAPE of Ensemble Modeling Application by Dimensional Component (C1)

In analyzing the results of forecasting by dividing the error into bias and variance, in this experiment, adaptive boosting played an expected role to reduce dispersion (figure 4-6), but it was found that the error rate of the model was increased by increasing deviation as model complexity increases (figure 4-7). It can be inferred that the sequential sampling method in adaptive boosting was done in the direction of expanding the deviation. Overall, this implies that the application of the ensemble methods does not guarantee the improvement of accuracy for every time. Therefore, the effects of ensemble methods should be tested according to the complexity of the model.

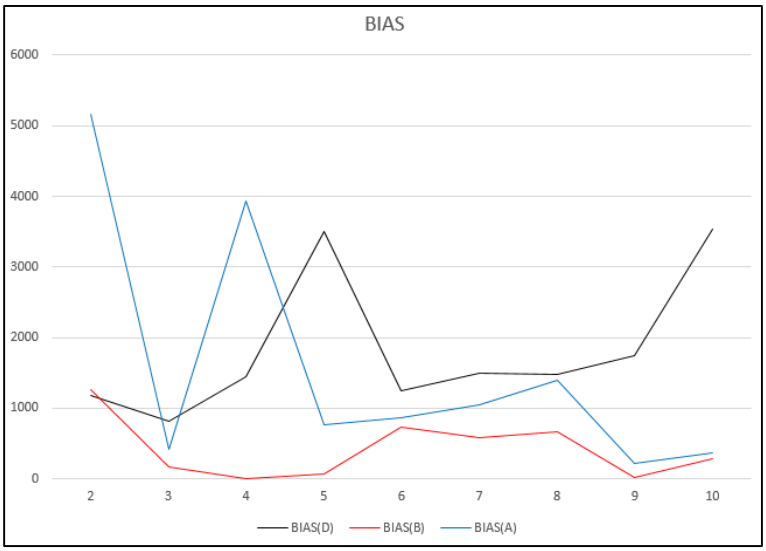


Figure 4-9. The Bias of Ensemble Modeling Application by Dimensional Component (C1)

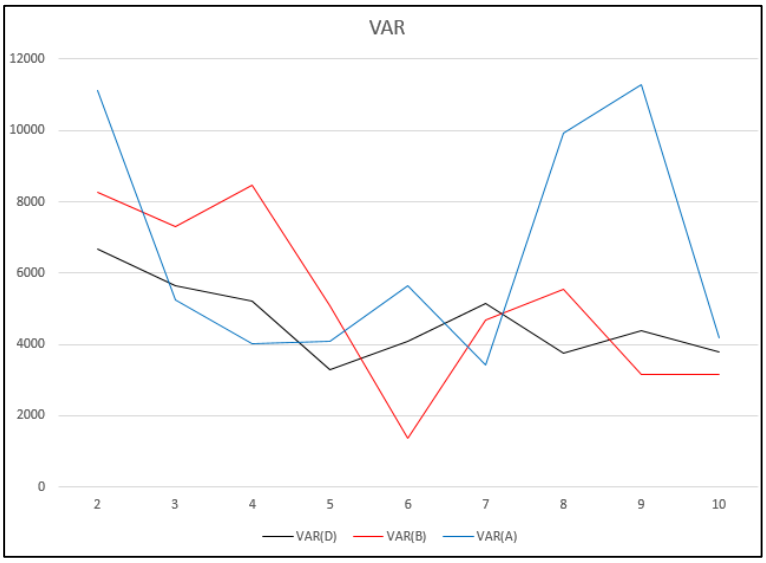


Figure 4-10. The Variance of Ensemble Modeling Application by Dimensional Component (C1)

Variance means how much of the dispersion is gathered and scattered. This value is a measure of the stability of the model. From this point of view, we should look at the accuracy of the forecast together with the variance. Comparing each error and variance, it is found that the MAPE value is small when the variance is small. When the value of the MAPE was the smallest, the variance had a small value at least within the second of the variance values for the whole dimensional components. The results show that the optimal forecasting model derived from the lowest MAPE value is also reliable in terms of stability.

4.3 Summary

In this chapter, to verify the proposed conceptual cost forecasting model's functions to produce accurate cost estimates, ensemble modeling application and factor analysis are tested as functional components of the proposed model.

The first experiment was designed to verify the positive effect of the application of ensemble modeling techniques on improving the accuracy of the forecasting model. Both bootstrap aggregating and the adaptive boosting, these two modeling techniques are applied to develop cost forecasting model based on combined cycle power plant and high-rise building project data, respectively. While the bootstrap aggregating application is showing the positive effects on the performance of cost forecasting model, the use of adaptive boosting cannot demonstrate the improvement of model performance, and even the results were poor as statistically insignificant.

Next, for testing the effects of model complexity on cost forecasting model, the second experiment was conducted with a database of combined cycle power plant projects. The model behavior of ensemble application is also analyzed by the dimensional components involved. As a result, not only

the accuracy of the artificial neural network model but also the effect of bootstrap aggregating and adaptive boosting methods varies corresponding to the number of dimensional components involved.

Two experiments are conducted to verify the proposed methodology functions appropriate corresponding to its purpose. From this process, the necessity of investigation of model performance by model complexity is derived. Also the possibility that the proposed model can be a tool to find better prediction results are shown. With this in mind, case studies are conducted to see if the proposed method can contribute to improving the performance and applicability of cost forecasting models.

Chapter 5. Case Studies

This chapter describes case studies using the proposed conceptual cost forecasting model. The case studies are designed to demonstrate and validate the effectiveness and applicability of the developed model. Total three types of project data including combined cycle power plant, high-rise building, and government office building are prepared. Power plant project and high-rise building project database is established based on projects carried out worldwide. And government office project database is based on the project completed in South Korea. Table 6-1 summarizes the explanation of case project data.

Table 5-1. Summary of Case Project Data Description

Case Project Type	Region	Number of Project	Number of Variables	Data Source
Combined Cycle Power Plant	Worldwide	86	11	Construction Intelligence Center
High-rise Building	Worldwide	44	7	Council on Tall Buildings and Urban Habitats (CTBUH).
Government Office Building	South Korea	104	12	Public Procurement Service (South Korea)

Note: the number of variables excludes the cost variables (output)

Each case studies are conducted to examine the accuracy of cost estimation results after developing the prediction model by learning the corresponding project cases with neural networks. When analyzing the accuracy of the forecasting model, random subsampling is applied as a validation method. This method is based on randomly splitting the data into subsets, for each test, 25% of the input data is supposed to be selected randomly for the creation of a test set. And mean absolute percentage error (MAPE) is used to measure the accuracy of the forecasting model as the verification experiment.

5.1 Combined Cycle Power Plant (C1)

Power generation project developments are essential in energy planning and policy. Energy infrastructure projects have risen sharply according to the increase in worldwide demand, financed by national governments and private capital development banks. A thermal power plant is a facility in which heat energy is converted to electric power. In most of the places in the world, the turbine is steam-driven combined cycle power plant. The most representative combination is gas turbine cycle and steam turbine cycle.

5.1.1 Case Base Description

The project data of combined cycle power plant is collected from Construction Intelligence Center (CIC). Although there is a larger and smaller capacity of power generation projects, we have collected worldwide project data from 500MW to 1000MW. The project database consists of total 86 projects data that are completed in 29 countries; the projects are located in Asia and Pacific (14), Europe (11), Middle East and Africa (16), North America (27), and South and Central America (20).

The collected information consists of three quantifiable variables and eight categorical variables. Quantitative variables include project capacity, project duration, and construction duration. And qualitative variables include project stage, project operation type, region, country, primary fuel, funding status, financing structure, and funding mode. The detailed information of categorical variables is described in the appendix section. The collected variables contain necessary information defined by AACE class 5 estimates. (class 4 estimates are covered partially.)

As the projects were distributed in the timeline from 1999 to 2018, the project cost should be normalized reflecting the cost escalations. In this research, Chemical Engineering Plant Cost Index (CEPCI) is used as a tool for adjusting plant construction costs from one period to another. The CEPCI consists of composite index assembled from a set of four sub-indexes: Equipment (Heat Exchangers/Tanks, Pipe/valves/fittings, Process Machinery, Pumps and Compressors; Construction Labor; Buildings; and Engineering & Supervision. Most of these components correspond to Producer Price Indexes (PPIs), updated and published monthly by the U.S. Department of Labor's Bureau of Labor Statistics.

5.1.2 Results and Discussions

Table 5-2 summarizes the results of the proposed conceptual cost forecasting model application. The table shows the comparison of implementation and combination of proposed methodologies used. The combined use of bootstrap aggregating and FAMD shows the best result; 10.76% of MAPE. Given the high accuracy of class 5 and class 4 estimates covers -20% to + 30% (class 5) and -15% to + 20%, This result can be said to have sufficient value for the development model.

Table 5-2. MAPE for Ensemble Methods and FAMD Application (C1)

Category	**Default (only ANN)**	**ANN + Bootstrap Aggregating**	**ANN + Adaptive Boosting**
Default (only ANN)	16.06	13.35	16.27
ANN + FAMD	15.49	10.76	12.82

In terms of the stability of the model, when the ANN and the bootstrap aggregating were applied, the variance of final model with the lowest error was the second smallest among all models according to all dimensional components. The model with the smallest variance was at the input of six-

dimensional component values and had a MAPE value of 12.7% when bootstrap aggregating was applied. And when adaptive boosting was applied, the model with the smallest variance among all model has the lowest MAPE value. Five-dimensional components values were input for the model with adaptive boosting. The collected information is very limited, especially the number of quantifiable variables are only three. Among them, only generation capacity has correlated a relationship with project cost. The result of correlation analysis is presented in the following tables. It seems that the neural network algorithm of this model has learned the nonlinear relationship between variables well.

Table 5-3. Correlation Analysis Result of Quantifiable Variables (C1)

Attributes	Project Duration	Construction Duration	Generation Capacity	Project Cost
Project Duration	-	0.237	0.094	0.152
Construction Duration	-	-	-0.114	-0.008
Generation Capacity	-	-	-	0.624*
Project Cost	-	-	-	-

Note*: correlation exists if the correlation coefficient is bigger than 0.5

5.2 High-rise Building (C2)

In recent years, there has been a considerable increase in the number of development of high-rise building projects. As the high-rise buildings projects require a tremendous amount of investment, careful and deliberate cost planning is essential. In the case of a skyscraper project, there are not many similar cases, and there is not enough data to refer to the cost estimate, so the cost forecast of skyscraper project has considerable difficulties and risks compared to general building projects.

5.2.1 Case Base Description

The project data of high-rise building project is collected from Council on Tall Buildings and Urban Habitats (CTBUH). Due to the lack of a similar past case, their construction data can be scarce. In this research, only 44 high-rise building project data is collected from 5 countries; United States (19), China (10), Singapore (2), United Arab Emirates (8) and Australia (5).

The collected information consists of four quantifiable variables and three categorical variables. Quantitative variables include the height of the building, the gross floor area of tower building, gross domestic product (GDP)

of the country where the project is developed, and the number of the parking lot. And qualitative variables include country, finishing types and structural types of tower building. The detailed information of categorical variables is described in the appendix section.

For data revision, local construction cost indexes were used to convert the construction costs of the buildings projects to the specific base of each country. All the projects' costs were converted to costs on the base year of 2010. Each country had a particular base city for which the costs were converted to New York City for the USA, Beijing for China, Dubai for UAE, Sydney for Australia and Singapore as Singapore. The location factors are used to convert investment estimates from the US to another country of interest. They are calculated based on high volumes of local data of different locations, relating to productivity, labor costs, steel and energy prices, equipment import needs, freight, taxes and duties on imported and domestic materials and regional business environment, among others.

5.2.2 Results and Discussions

Due to the limited number of project (44 projects only), it is required to increase the training amount of artificial neural network model by applying ensemble modeling. Moreover, it was challenging to obtain better-predicted results due to the large scale of the project and the enormous construction cost. For high-rise building project data, adaptive boosting has a lower value of MAPE, which means it has good forecasting performance. In the course of several trial test, the adaptive boosting technique sometimes yielded an exaggerated value like highly accurate or too far from the target. For this dataset of a high-rise building, the overall performance of adaptive boosting application was high enough but among the dimensional values.

Table 5-4. MAPE for Ensemble Methods and FAMD Application (C2)

Category	**Default (only ANN)**	**ANN + Bootstrap Aggregating**	**ANN + Adaptive Boosting**
Default (only ANN)	35.77	22.26	21.50
ANN + FAMD	29.86	17.78	14.41

5.3 Government Office Building (C3)

The project information is collected from materials “Analysis of Construction Expenses Classified by Public Facilities,” annually published by Public Procurement Service in South Korea. This material is provided to be used as a basic guideline for making decisions such as reviewing the feasibility of the projects and budgeting which reflects the proper construction cost when ordering a building from a public institution. Also, this resource can be used for contractors and designers as a reference for estimating the construction cost with design information.

5.3.1 Case Base Description

The project information is collected from materials “Analysis of Construction Expenses Classified by Public Facilities,” annually published by Public Procurement Service in South Korea. Total 104 public projects are collected including government office for various professional, research facilities, medical service facilities, and library.

The collected information consists of nine quantifiable variables and three qualitative variables. Quantitative variables include height, floor,

project duration, total area, land area, construction area, coverage ratio, floor area ratio, and the number of the parking lot. And categorical variables include region, project types and structural types. Besides, this database has three kinds of cost related information; total cost, contract cost, and survey amount. In the survey amount information, detailed cost elements such as material cost, labor cost, and expenses are documented. The government materials can be also inferred from the total construction cost and the survey amount information.

The cost information in the database is also processed to reflect inflation by time using a Korean construction cost index published by the Korea Institute of Civil Engineering and Building Technology (KICT).

5.3.2 Results and Discussions

Because this database has a different type of cost attributes, the cost forecasting model is applied to estimate the contract cost and survey amount, respectively. While the combination of bootstrap aggregating and FAMD shows 12.29% of MAPE for forecasting contract value, 8.27 % of MAPE is shown for forecasting survey amount. Forecasting contract cost, the final models with the lowest MAPE value are developed with three-dimensional components when bootstrap aggregating is applied, and six-dimensional components are required when adaptive boosting is applied. In the default neural networks and the bootstrap aggregating application, the final model had the lowest variance. When boosting was applied, the final model had the second lowest variance.

Table 5-5. MAPE for Ensemble Methods and FAMD Application (C3)

Category	**Default (only ANN)**	**ANN + Bootstrap Aggregating**	**ANN + Adaptive Boosting**
Default (only ANN)	29.32	26.49	33.61
ANN + FAMD	15.87	12.29	13.20

Table 5-6. Forecasting Result of Contract Cost and Survey Amount (C3)

Category	**Default (only ANN)**	**ANN + Bootstrap Aggregating**	**ANN + Adaptive Boosting**
Survey Amount	13.54	8.27	11.62
Contract Cost	15.87	12.29	13.20

Note: The values are the best predicted MAPE (%).

Table 5-7. Correlation Analysis Result of Cost Variables (C3)

Attributes	Survey Amount	Contract Cost	Project Cost
Survey Amount	-	0.106	0.899*
Contract Cost	-	-	0.134
Project Cost	-	-	-

Note*: correlation exists if the correlation coefficient is bigger than 0.5

The table 5-6 compares the error rate when forecasting different output and table 5-7 shows the relationship among survey amount, contract cost and project cost. The correlation analysis between cost variables as shown in Table 7 is conducted because the researchers assumed that there would be a correlation between cost variables. However, there was only a correlation between the total construction cost and the survey amount. The total cost is

the sum of the survey amount and the government material cost. Contract costs, on the other hand, were not correlated with different cost values. This means that the cost forecasting model that learned the Public Procurement Service data could be used independently for each role.

Originally, the survey amount is estimated with quantity take off and calculating the unit price based on the design information during the detailed design phase. This case studies show the possibilities of accumulated data from bottom-up based estimation approach can be applied to a top-down estimation approach. The Public Procurement Service will be able to estimate construction costs at a level of 10% accuracy using the project information that is expected when the initial plan is set up before the start of the design for the new project. In the case of contractors, the estimated contract amount can be predicted by using the proposed model during the bidding phase. And the result can be utilized as a necessary reference for analyzing the adequacy of the contract amount calculated by themselves.

5.4 Summary

In this chapter, the description of conducting the case studies is presented to demonstrate and validate the effectiveness and applicability of the developed model.

Total three types of project data including combined cycle power plant (86), high-rise building (44) and government office building (104) are utilized. Under limited numbers of case project especially power plant project and high-rise building project, the proposed conceptual cost forecasting model shows the reliable level of accuracy. The results of two case studies confirm the ideas of ensemble methods application being able to improve the accuracy of estimates, especially in neural networks with sparse data.

The last case study used the data of government office building projects from public procurement service. This data is generated through estimation with quantity take off and calculating the unit price based on the design information during the detailed design phase. By applying this data to the proposed forecasting model, the broader possible application of the proposed estimation approach is investigated.

Chapter 6. Conclusions

This chapter summarizes the research results and this study's contribution to the technical and academic points of view is described. This chapter finally provides its limitations and required future works for enabling the research results to be applied to the conceptual estimation practice for projects development in the future.

6.1 Research Results

Conceptual cost estimation at the planning stage of project development is critical for successful planning and execution of the construction project as the estimated cost is important information for decision making for all stakeholders. These initial appropriation and economic feasibility studies for a construction project are based on the conceptual cost estimates. For this purpose, the client should estimate the reliable budget and the contractor should estimate the accurate price for the project execution. Despite the importance of conceptual cost estimation, accurate estimation of cost budgets is a difficult task due to the increasing complexity of the project and limited information availability in the early phase of the project. Even though contractors improved cost estimation methodology for increasing estimating

accuracy, the actual costs of construction projects still show a significant difference from their originally forecasted costs.

Having reviewed previous literature on data-driven conceptual cost estimation research, several challenging issues are identified. Although many researchers have gathered past project data to create a database to generate a cost forecasting model, there is still insufficient data to be learned to estimate the cost of the project. In particular, it is difficult to find a developed case of data-driven cost forecasting models for projects that lack past examples. In addition, there are many types of research related to the selection of influencing input variables. Considering the explanatory power of the input data and the forecasting model, reducing the number of input variables does not guarantee the performance of the forecasting model. Contrary to this, when the amount of data and the number of attributes increase, multicollinearity and model complexity-related problems arise. Therefore, the accuracy of the forecasting model should be analyzed in relation to the complexity of the data. And the multicollinearity problem should be appropriately solved when developing conceptual cost forecasting model.

To address these problems, this research suggested and developed the conceptual cost forecasting model to forecast the cost of the construction project accurately with limited information. Neural networks ensemble

model incorporating factor analysis is designed to improve the performance of conceptual cost estimation and analyze the performance of the forecasting model according to the complexity of the input data.

The proposed conceptual cost forecasting model is verified and validated by several experiments and case studies. Three types of construction project data including combined cycle power plant, high-rise building, and government office building are utilized for the case studies. Under limited numbers of case project especially power plant project and high-rise building project, the proposed conceptual cost forecasting model compensate for the limitations of historical project data utilization by showing the reliable level of accuracy. The results of two case studies confirm the ideas of ensemble methods application being able to improve the accuracy of estimates, especially in neural networks with sparse project data. And the third case study used the data of government office building projects investigates more broad possible application of the proposed cost forecasting model.

Also, the results of this study have the following meaning. The experiments and case studies in chapter 4 and 5 show not only the accuracy of cost forecasting model but also the effect of bootstrap aggregating and adaptive boosting methods application varies corresponding to the number of

dimensional components involved. The combined use of artificial neural networks, ensemble modeling, and factor analysis revealed that the suggested machine learning algorithm (artificial neural networks and ensemble modeling) does not guarantee the improvement of forecasting performance. The result of the research argues that it is necessary to investigate the performance of cost forecasting model by dimensional components of data as the effects of data-driven modeling methods varies through the complexity of the model. From this point of view, this research emphasizes the necessity of investigation of model performance by model complexity and provide the possibility of obtaining better performance of cost forecasting model.

6.2 Research Contributions

This research possibly has contributions to the existing body of knowledge on conceptual cost estimation practices and studies on developing data-driven cost forecasting models. The findings and results of this research have the following main contributions to a scientific and industrial aspect, respectively.

From a scientific point of view, the findings and model development approach presented in this research can guide the development and validation of a reliable and accurate cost forecasting model. This research re-examines industry-specific knowledge regarding conceptual cost estimation research, and awaken the necessity to improve the performance of cost forecasting models by addressing identified the problems and issues faced while developing conceptual cost estimation model. In more detail, by proposing more flexible methods incorporating artificial neural network, ensemble methods, and factor analysis, this research shows an opportunity to improve the accuracy and stability of conceptual cost estimation.

And from the industrial point of view, this research contributes to making a shift from traditional uses of project data to more enhanced

resources for performance prediction support by exploring the usefulness of project data. In the past, the project management professional and researchers see the past project as sources of qualitative information such as risk and benchmarking data to improve the project management practices. In the perspective of the present research, the above uses of past projects are termed as underutilization of the information, and the content of this thesis have highlighted the usefulness of previous project data, beyond limited use of past project data. This research provides a new and productive role of the past project data or statistics by providing a methodology for utilizing past project information effectively for conceptual cost estimates. If the information created in a new way from the previous project can be validated, it will add further value to the already existed use of project information and encourage data-driven conceptual cost estimation, which offers better ways to get insight from the past project data.

With the proposed conceptual cost forecasting model, the more accurate cost in the conceptual stage can be estimated. This can support the decision-making of an organization where accurate conceptual cost estimation is required for project development. By establishing stable budgeting, this can ultimately prevent cost overrun in advance. With limited information in the planning phase of the project, the suggested model can satisfy clients with conceptual cost estimation results with reliable level accuracy.

6.3 Limitation and Future Research

Although it noted that this research performed based on limited research scope, additional research and examinations have to perform to further validate the proposed conceptual cost forecasting model. Regarding this, three research suggests the future research area as follows.

Since cost data of construction projects are confidential and thus very difficult to obtain more detailed data from public institutions or construction companies, the main limitation is inherent due to the very limited data set used in this research, Also, the proposed machine learning approach is more effective especially when the amount of data is collected as much as possible. Therefore, this limited data availability limits generalization of the findings of this research.

To have more concrete conclusions and better performance of cost forecasting model, a more extensive dataset consisting of more projects with similar nature should be used. Although this research tested the proposed methodology for three types of projects, more extensive experiments on various project types and data sets need to be considered to generalize the findings of this research over the construction industry.

Next, regarding the methodology used in the study, the dimensional components derived from the FAMD algorithm are used as the input parameter of the neural network-based model. It is not easy to identify the relationship between individual variables and each dimension. Moreover, since the artificial neural network model is also a black box model, the cost forecasting model proposed in this research is hard to understand the effect of individual variables on project cost. In this respect, the results of this research cannot answer questions about which variables should be used or how many variables should be used to improve prediction performance. Therefore, experiments to compare and analyze the amount and characteristics of input variables is required. Lastly, because the proposed methodologies in this research are flexible to data types and forecasting algorithm, more machine learning methods can be involved in the experiment, for example, Support vector machine, k-Nearest Neighbor for verifying the applicability of proposed concepts and improve the performance of conceptual cost forecasting model.

Bibliography

Adeli, H. & Wu, M. (1998). Regularization neural network for construction cost estimation. Journal of construction engineering and management, 124(1), 18-24.

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

Ahn. (2016). Front-End Cost Estimation by Selective Case-Based Reasoning for Building Construction Projects (Doctoral dissertation, Seoul National University).

Al-Tabtabai, H. Alex, A. P. & Tantash, M. (1999). Preliminary cost estimation of highway construction using neural networks. Cost Engineering, 41(3), 19.

An, S. H., Park, U. Y., Kang, K. I., Cho, M. Y., & Cho, H. H. (2007). Application of support vector machines in assessing conceptual cost estimates. Journal of Computing in Civil Engineering, 21(4), 259-264.

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

Awojobi, O., & Jenkins, G. P. (2016). Managing the cost overrun risks of hydroelectric dams: an application of reference class forecasting techniques. Renewable and Sustainable Energy Reviews, 63, 19-32.

Bakhshi, P. & Touran, A. (2014). An overview of budget contingency calculation methods in construction industry. Procedia Engineering, 85, 52-60.

Batselier, J., & Vanhoucke, M. (2016). Practical application and empirical evaluation of reference class forecasting for project management. Project Management Journal, 47(5), 36-51.

Batselier, J., & Vanhoucke, M. (2017). Improving project forecast accuracy by integrating earned value management with exponential smoothing and reference class forecasting. International journal of project management, 35(1), 28-43.

Bayram, S. & Al-Jibouri, S. (2016). Efficacy of estimation methods in forecasting building projects' costs. Journal of construction engineering and management, 142(11), 05016012.

Bayram, S., & Al-Jibouri, S. (2016). Application of reference class forecasting in Turkish public construction projects: contractor perspective. Journal of management in engineering, 32(3), 05016002.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In Neural networks: Tricks of the trade (pp. 437-478). Springer, Berlin, Heidelberg.

Black, J. (2013). Quality Guidelines for Energy Systems Studies: Capital Cost Scaling Methodology. DOE/NETL-341/013113.

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

Bode, J. (2000). Neural networks for cost estimation: simulations and pilot application. International Journal of Production Research, 38(6), 1231-1254.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Callegari, C., Szklo, A., & Schaeffer, R. (2018). Cost overruns and delays in energy megaprojects: How big is big enough?, Energy Policy, 114, 211-220.

Cheng, M. Y. Hoang, N. D. Roy, A. F. & Wu, Y. W. (2012). A novel time-depended evolutionary fuzzy SVM inference model for estimating construction project at completion. Engineering Applications of Artificial Intelligence, 25(4), 744-752.

Cheng, M. Y. Tsai, H. C. & Sudjono, E. (2010). Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry. Expert Systems with Applications, 37(6), 4224-4231.

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), 04013026.

Chou, J. S. (2011). Cost simulation in an item-based project involving construction engineering and management. International Journal of Project Management, 29(6), 706-717.

Christensen, P., & Dysert, L. R. (2003). AACE international recommended practice no. 17R-97 cost estimate classification system. AACE International, USA.

Christensen, P., & Dysert, L. R. (2005). AACE International Recommended Practice No. 18R-97 Cost Estimate Classification System–As Applied in Engineering, Procurement, and Construction for the Process Industries (TCM Framework: 7.3–Cost Estimating and Budgeting). AACE.

Dao, B., Kermanshachi, S., Shane, J., Anderson, S., & Hare, E. (2016). Identifying and measuring project complexity. Procedia Engineering, 145, 476-482.

Dao, B., Anderson, S., & Esmaeili, B. Developing a Satisfactory Input for Project Complexity Model Using Principal Component Analysis (PCA). In Computing in Civil Engineering 2017 (pp. 125-131).

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. Journal of Construction Engineering and Management, 134(2), 146-152.

Doloi, H. K. (2011). Understanding stakeholders' perspective of cost estimation in project management. International journal of project management, 29(5), 622-636.

Drucker, H. Burges, C. J. Kaufman, L. Smola, A. J. & Vapnik, V. (1997). Support vector regression machines. In Advances in neural information processing systems (pp. 155-161).

Dursun, O. & Stoy, C. (2016). Conceptual estimation of construction costs using the multistep ahead approach. Journal of Construction Engineering and Management, 142(9), 04016038.

Dysert, L. R. (2003). Sharpen your cost estimating skills. Cost Engineering, 45(6), 22.

Elhag, T. M. S. & Boussabaine, A. H. (1998, September). An artificial neural system for cost estimation of construction projects. In Proceedings of the 14th ARCOM annual conference.

Ellsworth, R. K. (2007). Cost to capacity factor development for facility projects. Cost engineering, 49(9), 26-29.

Emsley, M. W. Lowe, D. J. Duff, A. R. Harding, A. & Hickson, A. (2002). Data modeling and the application of a neural network approach to the prediction of total construction costs. Construction Management & Economics, 20(6), 465-472.

Firoozabadi, K. J. Rouhani, S. & Bagheri, N. (2013). Review of EPC projects cost estimation and minimum error technique introduction. International Journal of Science and Engineering Investigations, 2.

Flyvbjerg, B. (2006). From Nobel Prize to project management: Getting risks right. Project Management Journal, 37(3), 5–15.

Flyvbjerg, B. (2007). Eliminating bias in early project development through reference class forecasting and good governance. In K. J. Sunnevåg (Ed.), Decisions based on weak information: Approaches and challenges in the early phase of projects (pp. 90–110). Trondheim, Norway: Concept Program, The Norwegian University of Science and Technology.

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

Freund, Y. & Schapire, R. E. (1996, July). Experiments with a new boosting algorithm. In Icml (Vol. 96, pp. 148-156).

Gardner, B. J. (2015). Applying artificial neural networks to top-down construction cost estimating of highway projects at the conceptual stage.

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.

Garg, A., & Tai, K. (2013). Comparison of statistical and machine learning methods in modelling of data with multicollinearity. International Journal of Modelling, Identification and Control, 18(4), 295-312.

Gransberg, D. D. & Riemer, C. (2009). Impact of inaccurate engineer's estimated quantities on unit price contracts. Journal of Construction Engineering and Management, 135(11), 1138-1145.

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.

Gunduz, M. Ugur, L. O. & Ozturk, E. (2011). Parametric cost estimation system for light rail transit and metro trackworks. Expert Systems with Applications, 38(3), 2873-2877.

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.

Hu, X. Xia, B. Skitmore, M. & Chen, Q. (2016). The application of case-based reasoning in construction management research: An overview. Automation in Construction, 72, 65-74.

Huang, J. Li, Y. F. & Xie, M. (2015). An empirical analysis of data preprocessing for machine learning-based software cost estimation. Information and software Technology, 67, 108-127.

Hwang, S. (2009). Dynamic regression models for prediction of construction costs. Journal of Construction Engineering and Management, 135(5), 360-367.

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.

Islam, M. S., & Nepal, M. (2016). A Fuzzy-bayesian Model for Risk Assessment in Power Plant Projects.

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

Ji, S. H. Park, M. & Lee, H. S. (2011). Case adaptation method of case-based reasoning for construction cost estimation in Korea. Journal of Construction Engineering and Management, 138(1), 43-52.

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.

Juszczyk, M. (2017). The challenges of nonparametric cost estimation of construction works with the use of artificial intelligence tools. Procedia Engineering, 196, 415-422.

Kahneman, D. (1994). New challenges to the rationality assumption. Journal of Institutional and Theoretical Economics, 150(1), 18–36.

Kaushik, N. (2013). Prediction of project performance; Development of a conceptual model for predicting future performance of an OG&C project in an EPC environment.

Kassambara, A. (2017). Practical Guide To Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra (Vol. 2). STHDA.

Kim, G. H. An, S. H. & Kang, K. I. (2004). Comparison of construction cost estimating models based on regression analysis, neural networks, and case-based reasoning. Building and environment, 39(10), 1235-1242.

Kim, G. H. Yoon, J. E. An, S. H. Cho, H. H. & Kang, K. I. (2004). Neural network model incorporating a genetic algorithm in estimating construction costs. Building and Environment, 39(11), 1333-1340.

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.

Kirkham, R. (2014). Ferry and brandon's cost planning of buildings. John Wiley & Sons.

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.

Li, H. (1995). Neural networks for construction cost estimation: Building Research and Information, 23(5), 279-284.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. ACM Computing Surveys (CSUR), 50(6), 94.

Liu, M. & Ling, Y. Y. (2005). Modeling a contractor’s markup estimation. Journal of Construction Engineering and Management, 131(4), 391-399.

Liu Peng, L. L. (2005). A review of missing data treatment methods. Int. Journal of Intel. Inf. Manag. Syst. and Tech, 1(3).

Lovallo, D., & Kahneman, D. (2003, July). Delusions of success: How optimism undermines executives' decisions. Harvard Business Review, 56–63.

Lowe, D. J. Emsley, M. W. & Harding, A. (2006). Predicting construction cost using multiple regression techniques. Journal of construction engineering and management, 132(7), 750-758.

Lucko, G. & Mitchell Jr, Z. W. (2009). Quantitative research: Preparation of incongruous economic data sets for archival data analysis. Journal of Construction Engineering and Management, 136(1), 49-57.

Mahamid, I. (2011). Early cost estimating for road construction projects using multiple regression techniques. Construction Economics and Building, 11(4), 87-101.

McKim, R. A. (1993). Neural network applications to cost engineering. Cost Engineering, 35(7), 31.

Mitchell Jr, Z. W. (1998). A statistical analysis of construction equipment repair costs using field data & the cumulative cost model (Doctoral dissertation, Virginia Tech).

Morlini, I. (2002). Facing multicollinearity in data mining. In XLI Convegno della Società Italiana di Statistica (pp. 55-58). Cleup.

Moselhi, O. & Siqueira, I. (1998). Neural networks for cost estimating of structural steel buildings. AACE International Transactions, IT22.

Petroutsatou, C. Lambropoulos, S. & Pantouvakis, J. P. (2006). Road tunnel early cost estimates using multiple regression analysis. Operational Research, 6(3), 311-322.

Orme, G. J., & Venturini, M. (2011). Property risk assessment for power plants: Methodology, validation and application. Energy, 36(5), 3189-3203.

Petroutsatou, K. Georgopoulos, E. Lambropoulos, S. & Pantouvakis, J. P. (2011). Early cost estimating of road tunnel construction using neural networks. Journal of construction engineering and management, 138(6), 679-687.

Project Management Institute. (PMI). (2008). A guide to the project management body of knowledge (PMBOK® guide) – Third Edition. Newtown Square, PA: Author.

Rafiq, M. Y. Bugmann, G. & Easterbrook, D. J. (2001). Neural network design for engineering applications. Computers & Structures, 79(17), 1541-1552.

Rajagopalan, B. & Isken, M. W. (2001). Exploiting data preparation to enhance mining and knowledge discovery. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 31(4), 460-467.

Setyawati, B. R. Creese, R. C. & Sahirman, S. (2003). Neural networks for cost estimation (Part 2). AACE International Transactions, ES141.

Shtub, A. & Versano, R. (1999). Estimating the cost of steel pipe bending, a comparison between neural networks and regression analysis. International Journal of Production Economics, 62(3), 201-207.

Son, J., & Rojas, E. M. (2010). Impact of optimism bias regarding organizational dynamics on project planning and control. Journal of construction engineering and management, 137(2), 147-157.

Sonmez, R. (2011). Range estimation of construction costs using neural networks with bootstrap prediction intervals. Expert systems with applications, 38(8), 9913-9917.

Tibshirani, R. J. & Efron, B. (1993). An introduction to the bootstrap. Monographs on statistics and applied probability, 57, 1-436.

Touran, A. (2003). Probabilistic model for cost contingency. Journal of construction engineering and management, 129(3), 280-284.

Tsai, T. I. & Li, D. C. (2008). Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems. Expert Systems with Applications, 35(3), 1293-1300.

Wang, Y. R. & Gibson Jr, G. E. (2010). A study of preproject planning and project success using ANNs and regression models. Automation in Construction, 19(3), 341-346.

Yu, W. D. (2007). Hybrid soft computing approach for mining of complex construction databases. Journal of Computing in Civil Engineering, 21(5), 343-352.

Zima, K. (2015). The Case-Based Reasoning model of cost estimation at the preliminary stage of a construction project. Procedia Engineering, 122, 57-64.

APPENDIX

Appendix A. Estimate Input Information and Maturity Matrix

The appendix A is supplementary to the required project information for each level of cost estimation process described in Chapter 2. The table A-1 explains the extent and maturity of estimate input information against the five estimate classification levels. This is a checklist of basic deliverables found in common practice in the process industries. The maturity level is an approximation of the completion status of the deliverable. The completion is indicated by the following letters.

- None: development of the deliverable has not begun.
- Started: work on the deliverable has begun. Development is typically limited to sketches, rough outlines, or similar levels of early completion.
- Preliminary: work on the deliverable is advanced. Interim, cross-functional reviews have usually been conducted. Development may be near completion except for final reviews and approvals.
- Complete: the deliverable has been reviewed and approved as appropriate.

Table A-1. Estimate Input Checklist and Maturity Matrix (Primary Classification Determinate)

	Estimation Classification				
General Project Data	**Class 5**	**Class 4**	**Class 3**	**Class 2**	**Class 1**
Project Scope Description	General	Preliminary	Defined	Defined	Defined
Plant Production/Facility Capacity	Assumed	Preliminary	Defined	Defined	Defined
Plant Location	General	Approximate	Specific	Specific	Specific
Soils & Hydrology	None	Preliminary	Defined	Defined	Defined
Integrated Project Plan	None	Preliminary	Defined	Defined	Defined
Project Master Schedule	None	Preliminary	Defined	Defined	Defined
Escalation Strategy	None	Preliminary	Defined	Defined	Defined
Work Breakdown Structure	None	Preliminary	Defined	Defined	Defined
Project Code of Accounts	None	Preliminary	Defined	Defined	Defined
Contracting Strategy	Assumed	Assumed	Preliminary	Defined	Defined

Appendix B. Project Data Attributes (C1)

The project data of combined cycle power plant is collected from Construction Intelligence Center (CIC). The table below (Table A-2) describes the details of data attributes from the combined cycle power plant project used for model verification experiments (chapter 4) and case studies (chapter 5.1).

Table A-2. Descriptions of Project Data Attributes (C1)

Attribute	Description
Project Capacity	Project capacity refers to the generating capacity of the power generation project, in units of MW.
Project Duration	Project duration refers to the length of the period which the project was developed, and this duration starts from the time in which the project was announced officially by the owner.
Construction Duration	Construction duration refers to the length of the period which the construction was conducted, and this duration starts from the time in which the main construction works commenced on the project.
Project Stage	Project stages describe the current and exact status of the project. The CIC has classified the project stages as
Project Operation Type	Project operation type defines whether the project is developed as a single phase/entity or as multiple phase/entity. The CIC has classified the project types as Parent and Sub project.

Table A-2. Descriptions of Project Data Attributes (C1) Continued.

Attribute	Description
Funding Status	Funding status is an attribute that describes how much funding the project is done. The CIC has classified the funding status of the project as a fully funded project and partially funded project.
Financing Structure	The funding structure is an attribute that describes how the funding of the project is made up of debt and equity
Funding Mode	Funding mode is an attribute that distinguishes whether the project is delivered from public or private, or whether it is a public-private joint venture project.
Project Value	The project value is the total sum value/capital cost for the project, to be converted into US$ if sourced in other currencies.
Region	This attribute defines exactly where the project development and construction is undertaken.
Country	This attribute defines exactly where the project development and construction is undertaken.

Appendix C. Project Data Attributes (C2)

The project data of high-rise building project is collected from Council on Tall Buildings and Urban Habitats (CTBUH). The table below (Table A-3) describes the details of data attributes from the high-rise building project used for model verification experiments (chapter 4) and case studies (chapter 5.2). The project data of combined cycle power plant is collected from Construction Intelligence Center (CIC).

Table A-3. Descriptions of Project Data Attributes (C2)

Attribute	Description
Height	The height of the building is measured from the level of the lowest level of the architectural building to the top of the building, including spires.
Gross Floor Area	The total gross floor area includes the area within the tower footprint, not including adjoining podiums, connected buildings or other towers within the development.
Structural Types	The structure types are classified by the structure materials used for the building construction. The attributes are classified into the concrete structure, steel structure, and composite structure.

Table A-3. Descriptions of Project Data Attributes (C2) Continued.

Attribute	Description
Finishing Types	The finishing types describe the quality of the finishing materials of the building. The quality of finishing materials is classified as normal quality and high quality
Parking	This attribute means the number of car parking spaces contained within the building
Country	This attribute defines exactly where the project development and construction is undertaken.
GDP	The gross domestic product of the country where the building is constructed, in the year of completion

Appendix D. Project Data Attributes (C3)

The project information of government office building is collected from materials "Analysis of Construction Expenses Classified by Public Facilities," annually published by Public Procurement Service in South Korea. The table below (Table A-4) describes the details of data attributes from the government office building project used for model verification experiments (chapter 4) and case studies (chapter 5.3).

Table A-4. Descriptions of Project Data Attributes (C3)

Attribute	Description
Height	The height of the building is measured from the level of the lowest level of the architectural building to the top of the building.
Floor	This attribute means the number of floors within the building, not including underground floors.
Structural Types	The structure types are classified by the structure materials used for the building construction. The attributes are classified into the concrete structure, steel structure, and composite structure.
Project Duration	The project duration refers to the length of the period which the construction was conducted.

Table A-4. Descriptions of Project Data Attributes (C3) Continued.

Attribute	Description
The number of parking lot	This attribute means the number of car parking spaces contained within the building
Total Area	The total area is the sum of the floor area of a single building in the ground. The total area includes the ground floor as well as the underground floor and parking lot facilities.
Land Area	The land area refers to the horizontal projection area in accordance with the building law.
Construction Area	The construction area refers to the horizontal projection area (horizontal projection area) of the part enclosed by the outer wall of the building or the center line of the column, usually the floor area of the first floor.
Coverage Ratio	The coverage ratio refers to the ratio of the building area to the land area.
Floor Area Ratio	The floor ratio refers to the ratio of the building floor area to the land area.
Project Types	The project type is classified according to the use purpose of the building.
Region	This attribute defines exactly where the project development and construction is undertaken.

國 文 抄 錄

人工神經網 方法을 活用한 建設事業

槪念見積 費用豫測 方法 開發

건설프로젝트의 성공적인 수행을 위해서는 단계별로 체계적인 공사비 견적이 이루어져야 하며 특히, 프로젝트 계획단계에서는 프로젝트의 성공여부를 결정하는 중요한 의사결정이 이루어지기 때문에 이를 뒷받침하는 정확한 공사비 예측이 필요하다. 하지만 프로젝트 초기단계에서는 불확실한 프로젝트 범위, 부족한 관련 정보, 갈수록 복잡해지는 프로젝트의 특성으로 인해 정확한 공사비 예측에 어려움을 겪고 있다. 이에 효과적인 공사비 예측을 위해서 실무자와 연구자들은 과거 프로젝트 정보를 활용하여 새 프로젝트의 비용을 예측하는 다양한 방법을 개발하고 발전시켜 왔다. 점차 기계 학습을 포함한 데이터 분석 기술이 발전하고, 축적된 프로젝트 수행 실적이 증가함에 따라, 많은 연구자들이 데이터를 기반으로 새 프로젝트의 비용을 예측하는 방법론을 개발하고 정확도를 높이기 위한 연구를 진행하고 있다.

데이터 기반의 비용 예측 모델은 프로젝트의 특성을 반영하고 예측의 정확성을 보장하기 위해서 높은 수준의 데이터의 양과 품질을 요구하지만 건설 프로젝트 초기 계획단계의 특성 상, 확보할 수 있는 데이터의 양과 상세 수준에 한계가 있으며, 활용할 수 있는 데이터의 특성에 맞는 비용예측모델을 개발해야한다. 하지만 기존의 연구는 활용하는 데이터의 양, 변수의 수, 유형 등의 데이터 특성에 의해 좌우되는 예측모델의 복잡도가 성능에 미치는 영향을 고려하는데 한계점을 가지고 있으며, 본

연구에서는 데이터의 특성에 따른 다양한 문제에 대하여 모델 복잡도의 관점에서 통합적으로 접근하여 분석하였다.

이에 본 연구는 전술한 한계점을 극복하기 위해, 건설프로젝트 초기단계에서 활용되는 비용 예측 방법론 및 데이터 특성을 분석하고, 분석결과를 반영한 건설프로젝트 비용예측모델을 제시하였다. 건설프로젝트 초기단계에 확보할 수 있는 데이터의 특성에 맞추어 인공신경망, 앙상블기법, 요인분석을 결합한 비용 예측 방법론을 활용하여 모델 복잡도를 분석결과를 반영한 예측모델을 개발하였으며, 실제 프로젝트 데이터를 활용한 실험을 통해 제안한 방법론의 효과와 타당성을 검증하였다. 복합화력발전소 프로젝트, 초고층 프로젝트, 정부청사 프로젝트 등 총 3가지 유형의 건설프로젝트 데이터를 수집하고 사례 연구를 수행하여 제안한 비용예측모델의 유연성 및 적용가능성을 검토하였다.

본 연구에서 제안한 인공신경망, 앙상블 기법, 요인 분석을 결합한 건설프로젝트 비용 예측모델을 활용하여 건설프로젝트 계획단계 비용 예측의 정확도 및 신뢰성을 향상시킬 수 있을 것으로 기대된다. 또한 제안한 방법론을 활용하여 다양한 특성을 가진 프로젝트 데이터를 바탕으로 개발한 예측모델의 정확도와 설명력을 검증하였으며, 이를 통해 다양한 특성을 가진 건설프로젝트 데이터에 보다 유연하게 대응할 수 있는 비용예측모델을 개발 할 수 있을 것으로 기대된다.

주요어: 공사비용 견적, 개념 견적, 인공신경망, 앙상블 기법, 요인 분석, 모델 복잡도

학 번: 2013-30174