ESTIMATING COSTS DURING THE INITIAL STAGE OF CONCEPTUAL PLANNING FOR PUBLIC ROAD PROJECTS: CASE-BASED REASONING APPROACH

Seokjin Choi[1], Donghoon Yeo[2], and Seung H. Han[3]

[1] Ph.D. Student, School of Civil and Environmental Engineering, Yonsei University, S. Korea
[2] Master's Course, School of Civil and Environmental Engineering, Yonsei University, S. Korea
[3] Associate Professor, School of Civil and Environmental Engineering, Yonsei University, S. Korea
Correspond to sjchoi@yonsei.ac.kr

ABSTRACT: Estimating project costs during the early stage of conceptual planning is very important when deciding whether to approve the project and allocate an appropriate budget. However, due to greater uncertainties involved in a project, it is challenging to estimate costs during this initial stage within a reasonable tolerance. This paper attempts to develop a cost-estimate model for public road projects under these circumstances and limitations. In the conceptual planning stage of a road project, there is only limited information for cost estimation, for example, such input data as total length of the route, origin and destination, number of lanes, general geographic characteristics of the route, and other basic attributes. This implies that the model should individuate suitable but restricted information without considering detailed features such as quantity of earthwork and a detailed route of a given condition. With these limited facts, this paper applies a case-based reasoning (CBR) method to solve a new problem by deriving similar past problems, which in turn is used to estimate the cost of a given project based on best-fitted previous cases. To develop a CBR cost-estimate model, the authors classified 8 representative variables, including project type, the number of lanes, total length, road design grades, etc. Then, we developed the CBR model, primarily by using 180 actual cases of public road projects, procured over the last decade. With the CBR model, it was found that the degree of error in estimation can be reasonably reduced, to below approximately 30% compared to the final costs estimated upon the completion of detailed design.

Keywords: Cost estimation; Public road project; Case-based reasoning

1. INTRODUCTION

Roads are a major infrastructure facility all over the world. In South Korea, the government invested 7,300 billion KRW in road construction and maintenance in 2007, and this accounted for almost 41.7% of the annual budget of the Ministry of Land, Transport, and Maritime Affairs (MLTM). In addition to Korea, many developing or developed countries are investing enormous amounts of money in road construction and maintenance projects. Accordingly, cost estimation for road project is considered an essential part in view of reasonable budget planning, and a number of methods have been proposed for this end.

These estimation methods are mainly based on quantity-take off and adopt various methodologies such as artificial neural networks and computerized calculation algorithms; thus, require experts' engineering and design capability. However, calculating the quantity of every item requires lots of efforts and enormous time. Furthermore, it is almost impossible to calculate the quantity of every item, especially in the early stage of planning phase, due to the limited information available and possibility of frequent route changes over the succeeding design phase.

In South Korea, the Ministry of Land, Transport, and Maritime Affairs (MLTM) formulates the simple model, which is widely used during the early stage of road construction project. Even though this method is easy to use by simply providing average unit cost per each length (km),, their average error rates are significantly high due to limited input information. Consequently, we developed a construction cost estimation model in the planning stage by using the case-based reasoning (CBR) method. The CBR method is the process of solving new problems based on the solutions of similar past problems. CBR is considered a powerful method for computer reasoning and analogous to the human problem-solving process.

2. PREVIOUS STUDIES

2.1 Existing Cost Estimation Models

In South Korea, several cost estimation models are utilized to forecast the road construction costs for several stages. For the conceptual planning phase, the aforementioned MLTM's unit cost model is utilized, which applies the average unit cost concept to estimate road construction cost in terms of road embankment and land acquisition cost related to the total length (table 1). One of advantage of this model is that it can be applied

even when the user is not informed with exact length about tunnels and bridges. The only required information is project type (widening or constructing new road), the total length of the road, the number of lanes, and the location of the road, which means whether the road is located in an urban or rural area.

Table 1. MLTM Cost Estimation Model [1]

	# of Lanes	Road construction Cost	Land Acquisition Cost	Total Cost
Widening project in Rural Area	2→4	126	15	141
New project in Urban Area	4	154	28	182

(100 million KRW/km)

As for a preliminary feasibility study stage, more computerized cost and quantity estimation programs have been widely adopted to calculate the quantity of earthwork, which accounts for the largest part of road construction costs. A preliminary feasibility study stands for the brief checking process conducted by the Ministry of Strategy and Finance (MOSF) to decide whether a specific project is worth promoting.

However, the MLTM model does not consider other important information which highly influences road construction cost such as geographic feature, regional uniqueness, and function of a road. Subsequently, it is found that this model's estimation shows about 60% average error rate, particularly for a more complicated project, due to a lack of information during the initial planning phase. Also, it only suggests the average unit cost of road widening in rural areas and the average unit cost of new road-building in urban areas. Therefore, other forms of road project such as 4-lane road-building in a rural area or a road-widening from 4 to 6 lanes in an urban area cannot be gauged. Thus, cost estimation result inevitably brings about a high error rate within limited forms of road projects.

In the United States, engineering firms utilize unit cost calculating programs such as Timberline or ME2. These two programs provide cost database so that the user can easily calculate the unit cost information with less effort. The California Department of Transportation (Caltrans) accumulates actual project results in a database in terms of the item code, unit cost, quantity of each item, district, average price per unit, and total amount. These data can be utilized for the future construction cost estimation. In addition to Caltrans, the Building Cost Information Service (BCIS) in United Kingdom estimates road construction cost based on actual project results. The BCIS retains the cost estimation data of 16,000 projects over 45 years, and this data is used by consultants, clients, and contractors to produce specific estimates for appraisals, early cost advice, cost planning, and benchmarking.

2.2 Previous Studies on Road Construction Cost

Christian and Newton [2] analyzed the historic levels of expenditures on road maintenance, rehabilitation, and new construction in the province of New Brunswick during the period 1965-1994. In addition, they developed three cost prediction models to determine an accurate cost for road maintenance. From the result of this model, they argued that maintenance funding needs to be increased by 25% to adequately meet the current and future needs of the existing road network. However, this research has a limitation of data collection because the data used to develop the model were only from 1992 to 1995.

Liu and Zhu [3] pointed out limitations of previous research, which regard specific estimation methods as generic techniques. Moreover, so far little attention has been paid to the unique requirements at each project stage. Accordingly, this research developed a theoretical framework that identifies the critical factors for effective cost estimation during each project phase of a conventional construction project using organizational control theory. However, they only suggested the framework, thus needs to develop more extended models and validity testing based on further empirical studies.

Wilmot and Cheng [4] developed a future highway construction costs estimation model of Louisiana. The model predicts overall highway construction costs in terms of a highway construction cost index, which includes the cost of construction labor, materials, and equipments. Application of this model showed that the model predicts past construction costs for the period 1984-1997, and predicts that highway construction costs in Louisiana will double between 1998 and 2015. This model estimates that highway construction costs in Louisiana are likely to increase more rapidly than would be anticipated if past trends were extrapolated or if the rate of general inflation were used as an estimation of the future increase in costs.

Besides the aforementioned research, Wilmot and Mei [5] improved a highway construction cost model using an artificial neural network method. In addition, this research developed a procedure that estimates the escalation of highway construction costs over time. A comprehensive set of factors that influence the cost of highway construction were included in the model formulations. These factors comprehensively reflect the construction costs of the facility (labor price, material price, and equipment price) and the characteristics of each contract (pay items, contract duration, and project location). This model demonstrates almost the same result with Wilmot and Cheng [4]; signifying that highway construction cost in Louisiana will double between 1998 and 2015.

These previous studies considered road construction cost in a various view. However, they do not focus on the early stage of road planning, and were confined to several specific forms of road project such as a highway project in Louisiana. Therefore, we develop the road construction cost-estimate model, specifically designed for the early planning stage where cost information is extremely limited.

3. CBR-based COST-ESTIMATE MODEL

3.1 Case-Based Reasoning

People usually tend to use previous cases to solve a new problem; they reuse information and knowledge of a past situation for a new one. Case-Based Reasoning (CBR) is a subfield of artificial intelligence that applies this human action of thinking [6, 7], and has been widely used in many areas of medicine, finance, and insurance since the early 1990s [8, 9].

CBR solves a problem through a cyclic process that consists of 4 major processes called the '4 REs': retrieve, reuse, revise, and retain. This method solves a new problem by retrieving the most similar cases, reusing the information and knowledge in these cases, revising the solution based on reusing previous cases, and retaining the new experiences by incorporating them into the existing case-base [6]. From this process, CBR provides a wide range of advantages [7]:

- A user can propose solutions quickly, avoiding the time necessary to derive those answers.
- CBR allows users to propose solutions in domains that aren't completely understood.
- CBR can be used as a means of evaluating solutions when no algorithmic method is available.
- Cases help users to focus on important parts of a problem by pointing out what features of a problem are the crucial ones.

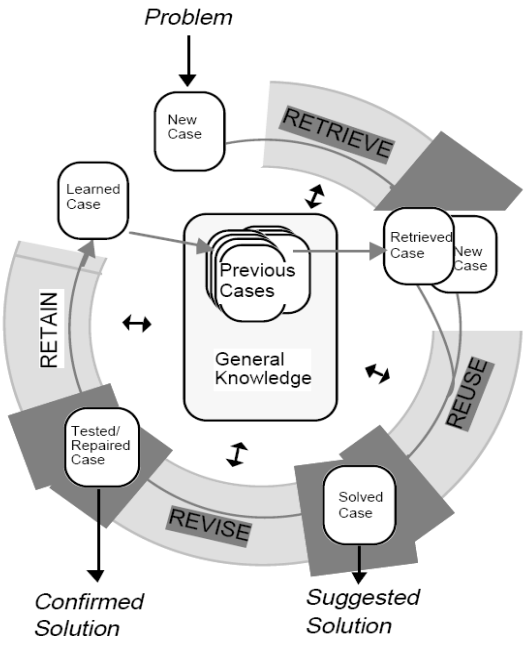

Figure 1. The CBR Cycle through the 4 REs [6]

3.2 Cost-Estimate Model

To build a cost-estimate model for a public road project at the initial stage of conceptual planning, CBR is used due to several advantages as below:

- Users can easily apply the CBR model by entering only a few input variables without understanding the model's algorithm.
- The detailed route of a road at the initial stage of conceptual planning can be easily changed due to economic limits, political issues, public grievances, etc. By including features of the road that aren't influenced by the changes as input variables, the result of the CBR model can be consistent in spite of the changes.
- By retaining and cumulating the new cases, the performance of the model can be continually improved as long as the model is used.

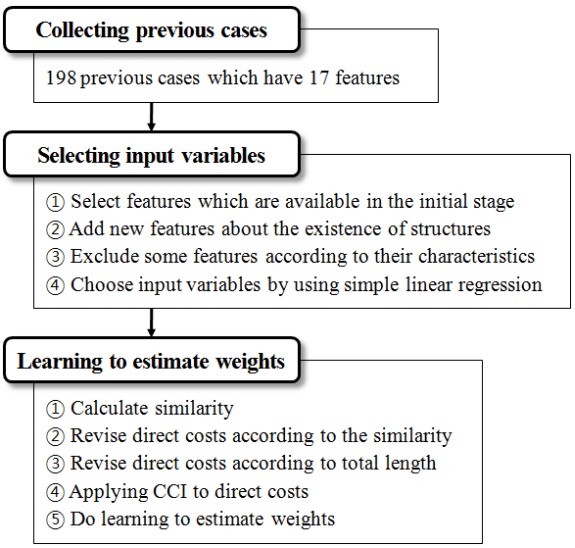

Figure 2. Procedure for Building the Cost-Estimate Model

3.2.1 Collecting Previous Cases

The cost-estimate model using CBR solves a new problem through deriving similar past problems, and thus it is highly influenced by the quality and quantity of previous cases [7]. To this end, this research collected 198 previous public road project cases constructed between 2000 and 2008 across the nation from 6 Regional Construction and Management Administrations (RCMA) in South Korea. Each case consists of 17 features: project owner, project type (new construction or expansion of existing road), provincial area, the number of lanes, the width of the road, the width of a single lane, total length, road grade (function), design speed, Minimum radius of curvature, maximum slope (the slope of the vertical section), the classification of the land acquisition, regional characteristic, geographic features, the quantity of earthwork, direct cost, and overhead cost.

3.2.2 Input Variables Selection

As the aforementioned 17 features were collected from completed road projects, some inputs are not available in the initial stage of conceptual planning. To sort out the features that can be grasped in the initial stage, therefore, several in-depth interviews with public owners who work on transportation policy or public budget—the primary users of the cost-estimate model—were performed. The interviewees selected 11 road project features as information that is available during the initial stage: project owner, project type, provincial area, the number

of lanes, and the width of the road, the width of a single lane, total length (planned), road grade (function), and design speed, regional characteristic and geographic features.

Then, a group discussion with interviewees was continued to uncover other information about road projects that is known during the initial stage. During the discussion, the large effect of the existence of possible underground water dissipation (weak ground) and inclusions of structures such as bridges or tunnels on the total construction cost was considered. Unfortunately, the existence of weak ground cannot be known during the initial stage because such a feature requires a geological survey. On the other hand, the existence of a long bridge or long tunnel can be detected even on a large-scale map with logical inference while the exact length and the type are still unknown. For example, if there is a wide river across the possible route connecting origin and destination or a mountainous range on the route, the users can intuitively infer that there will be a long bridge or a long tunnel. Since the conceptual route of a road is decided on a large-scale map such as 1:25,000 or 1:50,000 scale in the initial stage, 'the existence of a wide river (approximately more than 200 m)' and 'the existence of a mountain range' are added as new features, and entered into the collected data set.

Consequentially, a total of 13 features are available in the CBR model. However, some attributes are very similar to each other; thus, the multicollinearity problem can occur (e.g., 'the number of lanes' and 'the width of road' show a very similar distribution). Also, others can create new features by multiplying each other; for example, multiplying 'the width of the road' by the 'total length' produces the new 'road area measure'. The characteristics of these features are summarized as follows:

- 'The number of lanes' and 'the width of the road' show the same distributed pattern due to their similarity. For this reason, 'the width of the road' is substituted with 'the number of lanes,' which can be entered intuitively by the cost-estimate model user.
- In addition, 98.24% of the collected cases show the same value - 3.5 meters - in 'the width of a single lane' feature. It is therefore excluded from the model.
- The 'design speed' is determined by the 'road class.' Thus, the feature 'road class' can be substituted for 'design speed,' and only 'road class' is needed in the model accordingly.
- The feature 'regional characteristic' classifies the area of the road into city regions and rural regions while the 'geographic feature' classifies the surrounding terrain into three types: urban area, plains, and mountains. 'Geographic feature' is selected due to its detailed classification.
- The feature 'project owner' represents the RCMA that owns the project, and 'province' includes 9 provinces in South Korea. There are 6 RCMAs in South Korea, and each RCMA controls 1 or 2 provinces individually. Therefore, there is a high correlation between 'project owner' and 'province' features, and only 1 feature should be included in the model to avoid the multicollinearity problem.
- As above, multiplying 'the width of the road' by the 'total length' produces the 'road area measure,' and this new feature can substitute the 2 old ones.

Based on these characteristics, 4 groups of features were established (Table 2). After that, simple linear regression was applied to find the coefficient of determination (usually, indicated as R^2) of each group, and group A, which showed the highest R^2 value, was chosen as a set of input variables of the cost-estimate model.

Finally, 8 features were selected with further consideration of the availability and user's convenience as input variables for the cost-estimate model: provincial area, project type, the number of lanes, total length, road grade (function), geographic feature, the existence of a wide river, and the existence of a mountain range. Since those input variables are not easily affected by possible changes in the detailed route of a given road, we can steadily trust the model results, even though there are some changes in the road plan caused by economic budget limits, political issues, public resistance to the route, etc.

As a dependent variable in CBR model, direct cost is used; the overhead cost is inappropriate due to its wide variation according to included managerial costs, indemnities, etc. Also, indirect cost can be simply added as a percent of total direct cost at this stage. Direct cost can be applied as either a total cost or a unit cost. However, the unit cost calculated by dividing the total cost by the total length tend to decrease as the total length increases due to the rule of 'scale of economy' - the larger the project, the smaller the unit price. For this perspective, unit cost is inappropriate because it ignores the scale of a project.

3.2.3 Algorithm of the Model

Of 198 previous public road project cases, 180 were used to build the cost-estimate model, and 50 randomly selected cases were used for the learning process of the

Table 2. Selecting Input Variables According to R^2 Value

Group	Specific Variables	Common Variables	R^2
A	Province, the number of lanes, and total length	project type, geographic feature, road grade, the existence of a wide river, and the existence of a mountain range	0.610
B	Province, road area		0.596
C	Project owner, the number of lanes, and total length		0.602
D	Project owner, road area		0.586

CBR model when determining the weights of each attribute.

To build the CBR model, all variables of each case were scored first. Variables of previous case on nominal scale such as 'province,' 'project type,' 'geographic feature,' etc. were given 100 points when their value are exactly matched with that of the target case. On the other hand, scoring a variable that has a numerical value such as the 'total length' is very hierarchical; hence, variables were given 100 points when they take on the similar value of the target case by less than 10% difference, 80 points when the difference was less than 20%, 60 for less than 30%, and zero points in other cases. With this hierarchy, the model can retrieve candidate previous cases that have similar characteristics to the target case even when their total lengths are somewhat different.

The similarity score that represents the similarities between the target case and the previous case can be calculated by multiplying those variables' points by each variable's weight, which is estimated later. After that, top 5 cases that have the highest similarity scores are considered similar cases with the target project.

The direct costs of the 5 previous cases were then adjusted by their similarity score, total length, and construction cost index (CCI) successively. The similarity score was applied to determine each case's proportion to the estimated direct cost of the target case. Adjusting with the total length is a way to apply the unit cost concept to the model with considering the scale of economy. After that, converting the past cost to the current price was done by applying the CCI developed by the Korea Institute of Construction Technology (KICT). All of the cost data was converted to the value of December 2008. Finally, the average of the direct costs of the 5 previous cases was calculated to estimate the direct cost of the target case. This process of computation is summarized as follows:

$$COST_{est} = \frac{\sum_{n=1}^{5}\left(COST_{pre\,n} \times \frac{SS_n}{\sum_{i=1}^{5} SS_i} \times \frac{TL_{tar}}{TL_n} \times \frac{CCI_{200812}}{CCI_n}\right)}{5}$$

where, $COST_{est}$ = estimated direct cost of the target case
$COST_{pre\,n}$ = direct cost of previous case n
SS_n = similarity score of previous case n
TL_{tar} = total length of the target case
TL_n = total length of previous case n
CCI_{200812} = CCI in December 2008, 133.8
CCI_n = CCI of previous case n

3.2.4 Learning Process for the Weights

To estimate the weights of each input variable, a genetic algorithm (GA) was applied in the learning process by using Premium Solver for Excel v7.0. A genetic algorithm is the evolutionary process to find an optimum solution using evolutionary techniques such as inheritance, mutation, selection, and crossover [10]. To estimate the weights of each variable, the next fitness function was applied:

$$min \sum_{n=1}^{N} \left| COST_{act\,n} - COST_{est\,n} \right|$$

where, $COST_{act\,n}$ = actual direct cost of target case n
$COST_{est\,n}$ = estimated direct cost of target case n
N = total number of cases, 50 in this study

Table 3 shows the results of weights that optimize the projection of CBR model.

Table 3. Weights of Input Variables

Input Variable	Weight
Provincial area	0.0239
Project type	0.1525
The number of lanes	0.1599
Total length	0.1079
Road class	0.2121
Geographic feature	0.0624
The existence of a wide river	0.1682
The existence of a mountain range	0.1131
Total	1.0000

4. RESULTS OF THE MODEL

Eighteen cases (10% of the 180 cases used in model building) were randomly selected out of model samples and used to verify the cost-estimate model. Table 4 compares the cost-estimating results of the CBR model and the MTLM model. The CBR model showed an average error rate of 27.29% with a standard deviation of 17.87%, and this result is improved from that of the MTLM model (56.95%). The difference between the two models is statistically significant at a significant level, 0.05 (95%) from the t-test.

Table 4. Results Comparison

	Error Rate	Std. Dev.	Sig.
CBR model	27.29%	17.87%	0.021
MTLM model	56.95%	47.27%	

5. CONCLUSIONS

In this research, the cost-estimate model was suggested for the initial stage by adopting CBR methodology. We collected 198 previous cases that have 17 features such as total length, province, road area, etc., and refined the input variables according to their availability during the initial stage and the user's convenience. As a result, 8 input variables were derived from simple regression analysis. The suggested CBR model with a genetic algorithm showed better result to the existing model; an average error rate of 27.29% and a standard deviation of 17.87%. Also, users such as public servants and political decision-makers can make reasonable budget plans and effective policies by using this model.

For further improvement of this model, a systematic logical basis should be established and refined through a

numerical value scoring process. In addition, setting up a similarity criterion and similarity threshold value will give users of the cost-estimate model more confidence.

ACKNOWLEDGEMENT

This research was supported by a grant from the Construction and Transportation R&D Program (06CIT-A03) funded by the Ministry of Land, Transportation, and Maritime Affairs (MLTM) of the Korean government.

REFERENCES

[1] The Ministry of Land, Transport, and Maritime Affairs (MLTM), *The Road Construction Handbook*, The Ministry of Land, Transport, and Maritime Affairs, Gwacheon, South Korea, 2007.
[2] Christian, J., Newton, L., "Highway Construction and Maintenance Costs", *Canadian Journal of Civil Engineering*, Vol. 26, pp. 445-452, 1999.
[3] Liu, L., Zhu, K., "Improving Cost Estimates of Construction Project Using Phased Cost Factors", *Journal of Construction Engineering and Management*, Vol. 133(1), pp. 91-95, 2007.
[4] Wilmot, C. G., Cheng, G., "Estimating Future Highway Construction Costs", *Journal of Construction Engineering and Management*, Vol. 129(3), pp. 272-279, 2003.
[5] Wilmot, C., G., Mei, B., "Neural Network Modeling of Highway Construction Costs", *Journal of Construction Engineering and Management*, Vol. 131(7), pp. 765-771, 2003.
[6] Aamodt, A., Plaza, E., "Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches", *AI Communications*, IOS Press, Vol. 7(1), pp. 39-59, 1994.
[7] Leake, D. B. (editor), *Case-Based Reasoning: Experiences, Lessons & Future Directions*, MIT Press, Cambridge, Massachusetts, 1996.
[8] Riesbeck, C. K., Schank, R. C., *Inside Case-Based Reasoning*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1989.
[9] Schank, R. C., Kass, A., Riesbeck, C. K., *Inside Case-Based Explanation*, Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1994.
[10] Vose, M. D. *The Simple Genetic Algorithm: Foundations and Theory*, MIT Press, Cambridge, Massachusetts, 1999.