Knowledge-Based Model for Forecasting Percentage Progress Costs

Kim, Sangyong*
School of Construction Management and Engineering, University of Reading, Berkshire, Reading, RG6 6AW, UK

Abstract
This study uses a hybrid estimation tool for effective cost data management of building projects, and develops a realistic cost estimation model. The method makes use of newly available information as the project progresses, and project cost and percentage progress are analyzed and used as inputs for the developed system. For model development, case-based reasoning (CBR) is proposed, as it enables complex nonlinear mapping. This study also investigates analytic hierarchy process (AHP) for weight generation and applies them to a real project case. Real case studies are used to demonstrate and validate the benefits of the proposed approach. By using this method, an evaluation of actual project performance can be developed that appropriately considers the natural variability of construction costs.

Keywords: analytic hierarchy process, case-based reasoning, cash flow, progress costs, S-curve

1. Introduction

Successful management within the limitations of a budget is an important concern in any construction project. When there is a lack of information and reliable methods that support the estimation process, it is difficult to initiate estimating report during the project planning stage[1]. To control the cost within an acceptable level, the appropriate and accurate measurement of various project–related determinants and an understanding of the magnitude of their effects are required. As such, the importance of early estimation cannot be over-emphasized.

In addition, the vast majority of project progress cost models are based on the use of S-curves to forecast project expenditure cost flow. Barraza et al.[2] introduced a stochastic S-curve as an alternative to the deterministic S-curve that generated most likely budget and duration values. Blyth and Kaka[3] produced an individual S-curve for a particular project using a multiple linear regression model. Suhani[4] addressed the use of the banana envelope to show the integration of project resources with the project schedule stated as early and late dates. In truth, the variables that determine the shape of S-curves are very difficult to quantify and may not lend themselves to curve fitting. In addition, the thousands of sub–projects involved make it impossible to individually examine and construct an S-curve for each sub–project. Finally, S-curves constructed from actual cumulative progress are not smooth, and are often highly uneven.

There has been a large amount of work in the field of mathematical and statistical modelling of cost of progress forecasts. Although multiple regression analysis (MRA) has been used many times to forecast case flow based on statistics, it is not appropriate when describing non-linear
relationships, which are multidimensional, consisting of a multiple input and output problem[5]. To address this limitation, parametric models were developed using artificial neural networks (ANNs), and these models were found to be very useful at the early stages of a project life cycle [6–8]. ANN models have also been proposed for forecasting project progress and its cash flow. Chao and Chien[9] proposed ANN models for estimating S–curves using a cubic polynomial for fitting S–curves. The results provided better accuracy than the regression and average curve methods. Boussabaine and Kaka[10] and Boussabaine et al.[11] proposed a method for forecasting project progress and cash flow at a set interval of every tenth of project time without using an S–curve formula. However, ANNs can lose their effectiveness when the patterns are very complicated or noisy, knowledge representation and problem structuring are ill–defined, and training is trapped in local minima[12]. Thus, this study offers a case–based reasoning (CBR) method that overcomes these limitations by separating the mathematical component.

CBR is a relatively recent problem solving method that has been attracting increasing attention because it seems to more closely resemble the psychological process humans follow when trying to apply their knowledge to the solution of problems. CBR is a problem–solving technique that reuses past cases and experiences to find solutions to the problems. While other major artificial intelligence (AI) techniques rely on making associations along generalized relationships between problem descriptors and conclusions, CBR is able to benefit from utilizing specific knowledge of previous experiences, and evaluate the proposed solution and update the system by learning from this experience[13–15].

The aim of this study was to develop a CBR model to assist contractors in forecasting, planning and controlling the cost of progress. However, the scope of this study is limited to forecast expenditure cost, which is part of cash flow management. The accuracy of estimates from the progress matching method is unstable in a project’s beginning stage because estimations are based entirely on the fitted historical cases. These are similar in progress to the project, but just for that short period of time. Their progress may differ at later stages of the project. To counter this instability and improve accuracy, we propose incorporating the previously mentioned CBR model into the estimation process by integrating analytic hierarchy process (AHP) into the progress matching method. Thus, incorporating prediction uncertainty into deterministic forecasts can improve the reliability and credibility of the model outputs.

2. Methodology

Data that incorporate all the input variables, which were selected by interviewing fifteen experts with over five years of field experience, are required. When potential input variables were identified, the weight of each variable was calculated by AHP, and strategies for data collection established. In addition, an appropriate CBR model was developed and validated, and preliminary testing of a developed model was carried out, using a relatively small number of data sets. In order to use AHP to generate weights, one of the cases in the input case–based library was removed. For the purpose of developing a CBR model, a new excel visual basic application system was selected.

2.1. Case-Based Reasoning

CBR is a kind of computerized tool that imitates
the analogical reasoning of the human brain in problem solving[16]. The principle of CBR is based on the assumption that similar problems have similar solutions, According to Riesbeck and Schank[17], CBR solves problems by capturing previous experiences and matching the important features of new problems to those of the old cases that have been successfully solved. The main source of knowledge in CBR is the case that can be reused, even if it only partially matches the problem in hand[18]. In particular, CBR can deal efficiently with both numerical and nominal data, and can effectively handle cases that have incomplete data or variable data structures[19]. Furthermore, CBR has powerful learning capabilities that do not require time-consuming training and testing operations[18].

![Figure 1. Case-based reasoning cycle[20]](image)

In the six—Re processes, changes initiated from outside of the CBR can be modeled easily as shown in Fig. 1: (1) Retrieve the most similar cases from stored previous cases; (2) Reuse the retrieved cases to attempt to solve the problem; (3) Revise the proposed solution if necessary; (4) Retain the new solution as a part of a new case; (5) Review the results from applying the solution; and (6) Restore the case into the case base library.

### 2.2. Analytic Hierarchy Process

AHP is a multi-factor decision-making method that uses hierarchical structures to represent a decision problem and then delivers priorities for the decision-maker throughout judgments[21–24]. Many previous studies[25–30] consider the AHP methodology to be well suited for decision-making due to its role as a synthesizing mechanism in decisions. For example, An et al.[28] compared three different weighting methods and concluded that the AHP was more accurate, reliable, and explanatory than descent gradient methods to determine the relative weights of importance for making preliminary estimates of new construction costs. Once the hierarchy is built, the decision-maker systematically evaluates its components, which represent considered factors, by comparing their importance in a pair-wise manner. This study applies the AHP to calculate the weights of the aspects and the attributes within each aspect. Pair-wise comparisons of importance of the factors at each level of an AHP are made in terms of importance when comparing factors with respect to their relative importance[29–32] (see Table 1).

### Table 1. Scale of relative importance[28]

<table>
<thead>
<tr>
<th>Intensity of Relative Importance</th>
<th>Definition</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Equal importance</td>
<td>Two activities contribute equally to the objective</td>
</tr>
<tr>
<td>3</td>
<td>Moderate importance of one over another</td>
<td>Experience and judgment slightly favour one activity over another</td>
</tr>
<tr>
<td>5</td>
<td>Essential or strong</td>
<td>Experience and judgment strongly favour one activity over another</td>
</tr>
<tr>
<td>7</td>
<td>Very strong importance</td>
<td>An activity is strongly favoured and its dominance is demonstrated in practice</td>
</tr>
<tr>
<td>9</td>
<td>Extreme importance</td>
<td>The evidence favouring one activity over another (is of the highest) possible order of affirmation</td>
</tr>
<tr>
<td>2, 4, 6, 8</td>
<td>Intermediate values between the two adjacent judgments</td>
<td>When compromise is needed</td>
</tr>
</tbody>
</table>

Reciprocals of above nonzero number

If activity i has one of the above nonzero numbers assigned to it when compared with activity j, then j has the reciprocal value when compared to i.
The last step is devoted to measuring the overall consistency of the provided AHP judgments by means of the consistency ratio (CR) proposed by Saaty. The CR provides a way of measuring errors introduced during the eliciting of expert opinions. The value of the consistency index is applied as in Eq. (1)\[33\],

\[
CI = \frac{(\lambda_{\text{max}} - n)}{n-1} \quad (1)
\]

where \( n \) is the number of compared factors, and \( \lambda_{\text{max}} \) is the maximum eigenvalue of a judgment matrix corresponding to the group of compared factors. The CR value is given by dividing the consistency index (CI) value by the random consistency index (RI) value. The RI value depends on the number of compared factors. RI values for different numbers of factors are presented in Table 2.

<table>
<thead>
<tr>
<th>n</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>RI</td>
<td>0</td>
<td>0</td>
<td>0.58</td>
<td>0.9</td>
<td>1.12</td>
<td>1.24</td>
<td>1.32</td>
<td>1.41</td>
<td>1.45</td>
<td>1.49</td>
</tr>
</tbody>
</table>

Appropriate CR value justifies extracting expert knowledge that can guide effective retrievals of useful weights. If the CR is below \( 0.1 \)’, then the errors are fairly small and thus, the final estimating can be accepted. If it is more than \( 0.1 \)’, the judgments may be somewhat random and should perhaps be revised. The weight values expressing the importance of each CF are presented in Fig. 2. They will be assigned to the considered attributes for case-based retrieval of the most similar process plans due to the effective similarity function in the proposed application area.

3. Description of Data

The model must be affected by using collected information of data. The first stage in producing the CBR–AHP model required the gathering of data from sample cases of matching progress. These data were required to be of similar project type and duration. Typically, in such models, previous case projects are collected and retrieved according to some attributes. The proper formula parameters for case–based library are used to produce a standard matching progress as the basis of prediction for a new project classified in the same category. To illustrate the development of a system for estimating matching progress, data on the nature and actual progress of 57 building projects are shown in the case representation in Table 3.

3.1 Time standardization

A cost index represents the relative scale of cost for a fixed quantity of goods or services between different periods, and provides a good means for forecasting future construction costs that change over time in response to changing demand, economic conditions, and prices\[34\]. The data collected to develop a CBR–AHP model have diverse characteristics and differences, such as when and where the projects were constructed. Such differences may cause incorrect prediction results\[35,36\]. A cost index ought to be a reliable
tool for estimating future costs of construction activities, where construction activities are conducted months or years after costs were estimated [37].

<table>
<thead>
<tr>
<th>Table 3. Case Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
</tr>
<tr>
<td>Duration</td>
</tr>
<tr>
<td>No. of floor</td>
</tr>
<tr>
<td>No. of parking</td>
</tr>
<tr>
<td>Year</td>
</tr>
<tr>
<td>Floor area ratio</td>
</tr>
<tr>
<td>Building coverage ratio</td>
</tr>
<tr>
<td>Total area</td>
</tr>
<tr>
<td>Type of building</td>
</tr>
<tr>
<td>Location</td>
</tr>
<tr>
<td>Form system</td>
</tr>
<tr>
<td>Roof</td>
</tr>
<tr>
<td>Foundation</td>
</tr>
<tr>
<td>Superstructure</td>
</tr>
<tr>
<td>Substructure</td>
</tr>
<tr>
<td>Retaining wall</td>
</tr>
<tr>
<td>External wall</td>
</tr>
<tr>
<td>Internal wall</td>
</tr>
<tr>
<td>Ceiling</td>
</tr>
<tr>
<td>Floor</td>
</tr>
<tr>
<td>Output</td>
</tr>
</tbody>
</table>

First, the data used to establish the CBR–AHP model were collected from projects completed in 2000 and 2010. The data had to be converted to the identical time reference point defined by the Korea Institute of Construction Technology (KICT). The cost data of all the reference cases were converted to February 2005 cost level provided by the KICT. For example, February 2008 data were converted into February 2005 data by multiplying February 2008 cost data by the value (100.0/127.1=0.78678) calculated by dividing 100.0, the index value for February 2005, by 127.1, i.e., the index value for February 2008. The building cost index applied to the conversion was official statistical data prepared to estimate the price fluctuation of input resources by 100.0 time scale as the price of a direct building construction cost input in a project at a certain point in time.

Second, as is common in valuation of work done, progress measurement for all the projects took place at every 1% increment of construction progress in the range 0%~100%. All the fitted cases in the case-based library will also be used in matching progress. For each project, the data were first standardized by cumulative expenditure cost and progress time range. For each data set, the ratio of present cumulative expenditure cost against final cumulative cost was plotted against every percentage completion of the project, thus producing the sample expenditure cost progress. For example, the cumulative expenditure cost of progress at any time point $t_i$ between $t_j$ and $t_0$ in the case-based library from its fitted point is defined, and the average cost value is calculated at each of one hundred time points by Lagrange interpolation as shown in Eq. (2),

$$L(c_i) = \sum_{j=1}^{100} t_j \prod_{i=1(i\neq j)}^{100} \left( \frac{c_j - c_i}{c_j - c_i} \right)$$

---

522
4. Case Study

Expert knowledge can be applied to assess weights of importance. The expert is expected to have the required knowledge and experience to decide which model or system makes good predictions. CBR applications can be created using the hybrid CBR–AHP application development tool. The CBR model searches for matched cases contained in the case-based library, and summarizes them into a set of acceptable solutions. Decision-makers then select one of the recommended solutions. The system’s interface is organized by following the basic process used to construct the CBR–AHP application. The reasoning structure of the proposed system is presented in Fig. 3. The following six steps are involved in CBR application:

1) **Case definition**: The first step is used to define the initial components of the system. The names and value types for cost factors (CFs) are defined. The selected CFs should provide the best description of the relevant construction cost influencing attributes that result from prior experience. Table 3 presents an illustrative example of a case-based library contents.

2) **Similarity definition**: This step deals with the manner in which the similarity between a new problem description and the case-based library items is assessed. The methodology and various metrics for determining similarity during case-based retrieval are defined. **Similarity index** (SI) is assessed both at the case level (comparing cases against each CF) as well as at the CF level (comparing the value of each CF value to the new entered CF values). Weighted case similarities between the new problem and cases included in the case-based library are estimated according to the following formula:

\[
SI = \frac{\sum_{i=1}^{n} (W_i \times SS_i)}{\sum_{i=1}^{n} W_i} \times 100 \quad - (3)
\]

In Eq. (3), SI is a calculated numerical value that demonstrates the degree of similarity between a case in the case-based library and the investigated problem case [38]. SI is normalized into a scale from ‘0’ to ‘1’ for easy comparison. Weights (W) of each CF can be assigned by AHP. **Similarity score** (SS) is calculated on the basis of the values of the CFs: numerical and nominal. For the nominal factor, the SS equals ‘1’ when the two values are identified, and ‘0’ otherwise,

\[
SS = \frac{1}{\sqrt{\sum_{i=1}^{SS} (x_{\text{case-based library}} - x_{\text{problem}})^2 + 1}} \quad - (4)
\]

For the numerical factor, SS is calculated by Eq. (4). In Eq. (4), \(x_{\text{case-based library}}\) represents the value of a factor for the cases stored in a case-based library. \(x_{\text{problem}}\) value corresponds to the target case for predicting progress costs. A more detailed classification method is applied to improve the accuracy in this study when decision-maker selects one of retrieved cases. It is possible to select the best matching case from the case-based library. Consequently, a new SS formula has been
developed and proposed here, which not only expresses the difference of compared cases but also enables the verification of the minimum and maximum relationship of the cases. The $SS'$ in the developed formula is referred to as $SS$ to distinguish it from the $SS$ used to retrieve similar cases. Finally, $SI$ is calculated due to Eq. (3).

3) **Case definition**: This step is used to fill in the case information for each case to be stored in the case-based library. A case collection interface is then applied to introduce data for the real building project cases into the case-based library. CF values that describe the cases must conform to the defined types. 57 building cases are included in the case-based library in the prepared CBR model.

4) **Rule definition**: Rules are used to compute $SI$ and to adapt a retrieved similar case to better meet the needs of the new problem. Rules are used to address the differences that exist between a new problem case (target case) and the retrieved similar case. The rules are applied to account for the differences and advise regarding what the plausible outcomes of a comparison might be. Rules can be used to change CF values based on comparison.

5) **Application interface**: After case retrieval is complete, the system returns a list of cases with $SI$ values indicating their similarity to the target case. Their scores indicate their relevance to the problem at hand. Decision-makers can apply the selected case to help decide how to solve the current problem. The selected case can then be adapted to better assist in making a decision.

6) **System validation**: To determine whether the predicted project cost provided by CBR-AHP is a good estimate of the problem case, three methods that have been reported by Yau and Yang [38], Ardit and Tokemir [19] and Koo et al. [39] are used. Each of these methods makes use of the overall case $SI$ for each retrieved case. These methods are as follows: (1) the problem case is compared to the characteristics of the retrieved case that has the highest overall $SS$; (2) the problem case is compared to the most frequent characteristics in the top ten retrieved cases, or fewer if ten are not available, that have an overall $SS$ greater than or equal to .75 (75%); and (3) the problem case is compared to the average characteristics of the top five retrieved cases, or fewer if five are not available, that have an overall $SS$ greater than or equal to .75 (75%). The average of the predicted condition is weighted using the overall $SS$ to magnify the importance of the retrieved cases that have higher $SS$.

According to the CBR concept, the case with the highest $SI$ in the case-based library may be considered to have the most similar project characteristics compared to the test case in this study. These results may also be used as references in the decision-making process.

5. Results and Conclusion

Collected cases are used to illustrate the method and evaluate its accuracy. Two test cases are used to demonstrate the process by which project cost of progress is produced, refined, and then rearranged to reflect the broader objectives of the contractor. Table 4 summarizes the results from the two test cases. To evaluate the accuracy of expenditure cost estimates obtained from the above presented progress matching method, two error measures are used to measure the accuracy of the results in terms of closeness of fit, as well as to
provide a basis for subsequent system performance evaluation: mean absolute error rate (MAER) and Cronbach’s alpha,

\[ MAER = \left[ \frac{1}{n} \sum \frac{|c_r - c_a|}{c_r} \times 100 \right] \quad (5) \]

Where \( c_r \) is the retrieved expenditure cost by CBR-AHP, \( c_a \) is the collected expenditure cost, \( n \) is the number of sets of test cases.

<table>
<thead>
<tr>
<th>Table 4. Results of the CBR-AHP model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
</tr>
<tr>
<td>----------</td>
</tr>
<tr>
<td>0% to 10%</td>
</tr>
<tr>
<td>11% to 20%</td>
</tr>
<tr>
<td>21% to 30%</td>
</tr>
<tr>
<td>31% to 40%</td>
</tr>
<tr>
<td>41% to 50%</td>
</tr>
<tr>
<td>51% to 60%</td>
</tr>
<tr>
<td>61% to 70%</td>
</tr>
<tr>
<td>71% to 80%</td>
</tr>
<tr>
<td>81% to 90%</td>
</tr>
<tr>
<td>91% to 100%</td>
</tr>
</tbody>
</table>

The MAER of cost estimation is then calculated by comparing the retrieved expenditure cost with the actual cost of the new case, MAER gives a direct measure of the average error in a percentage. Through the repetition of this process for multiple new cases, the expenditure cost of new cases is predicted, and MAER is calculated. The result for MAER is 10.89%, which is calculated using two cases multiplied by 100 time intervals per each case (2 cases x 100 time/case = 200 times). However, the standard deviation in the early rate of progress (0% to 20%) is high, and during this period it is an important point to reduce the standard deviation. Nevertheless, CBR–AHP is an accurate and reliable system. This error rate is deemed acceptable by the American Association of Cost Engineers (AACE), and is the expected error range of the detailed estimate presented by the Construction Industry Institute (CII). Therefore, the predicted expenditure cost by the system is considered reliable.

![Figure 4. MAER of each progress(i.e. Case 1)](image)

![Figure 5. Retrieved and actual progress(i.e. Case 1)](image)

Cronbach’s alpha is used as a measure of the internal consistency or reliability of test results. It is commonly used as a measure of the internal consistency or reliability of a psychometric test score for a sample of examinees. A widely accepted social science cut-off point is that alpha should be .60 or higher for a set of cases to be considered a scale. The reliability analysis is conducted using Cronbach’s alpha that is calculated by Statistical Package for the Social Sciences (SPSS) 20.0. The result for Cronbach’s alpha is .728. This means that the model developed in this study could control
the Cronbach’s alpha of the prediction accuracy, and is sufficiently optimized to obtain consistent prediction accuracy.

\[
\text{Cronbach's } \alpha = \frac{K}{K-1} \left(1 - \frac{\sum_{i=1}^{K} \sigma_{r}^2}{\sigma_{x}^2}\right) = \frac{K}{K-1} \left(1 - \frac{\sigma_{r}^2}{\sigma_{x}^2}\right)
\]  

(6)

Where \(K\) is the number of components (K-items or testlets), \(\sigma_{x}^2\) is the variance of the observed total test scores, \(\sigma_{r}^2\) is the variance of component \(i\) for the current sample.

This study highlights the importance of project cost expenditure pattern, and discusses the developed forecasting CBR–AHP model. This study combines preliminary cost of progress estimation using a hybrid CBR–AHP model and subsequent estimation by matching progress. The developed model is an empirical methodology whose estimate is based on past data and does not depend on a detailed schedule of activity times. In terms of application, the CBR model was intended to predict financing requirements and to obtain a quick cost of progress forecast before the start of construction. The results show that the presented model can consistently reduce errors. Therefore, the model can be used to prepare a preliminary progress estimate at the early stages when minimal project information is available.

The main limitations of the model lie with the case-based library. First, the case-based library should be created from projects with the same conditions of contract concerning progress measurement, so that case progress records are usable references for the matching progress method. Second, the case-based library should have sufficient cases that cover a variety of projects, so that both the CBR model and its extension, the integration model, have better estimation accuracy than the present models.

References

37. Huang G, Cost modeling based on support vector regression for complex products during the early design phases [dissertation], Virginia: Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University: 2007, p. 29–34.