JOURNAL OF CIVIL ENGINEERING AND MANAGEMENT
ISSN 1392-3730/eISSN 1822-3605
2016 Volume 22(2): 178–186
doi:10.3846/13923730.2014.897981

A BAYESIAN APPROACH FOR FORECASTING ERRORS OF BUDGET COST ESTIMATES

Seokyon HWANG

Reese Construction Management Program, Lamar University, P.O. Box 10565, Beaumont, TX 77710, USA

Received 07 Feb 2013; accepted 10 Apr 2013

Abstract. Accurate estimation of budget costs is critical for effective management of construction projects. The performance of various management functions is dependent on the accuracy of the estimates throughout the construction phase. These estimates, however, inevitably involve a considerable amount of error, which imperatively requires the evaluation of budget cost estimates and the measurement of errors associated with the estimates. Applying an analytical procedure, this study carried out a thorough statistical analysis of existing practice in the construction industry to identify limitations of the practice. As an alternative to the practice, a Bayesian approach was found to be more appropriate than the industry common practice to account for the probabilistic nature of estimates and to forecast errors associated with the budget estimates. A scenario-based example is included to demonstrate application of the analytical procedure for analysing historical cost performance data that are readily available in most construction companies.

Keywords: cost estimating, estimate errors, Bayesian approach, risk management, cost contingency.

Introduction

Accurate estimation of budget costs is critical for successful management of construction projects. Budget cost estimates that are prepared at the preconstruction stage and are updated when significant changes are made to a project, are referenced in decision-making for planning and control throughout the execution of a project. Thus, the accuracy of estimates greatly affects the quality of decisions. Inaccurate estimates naturally increase the risks of a project and become a barrier to effective management. Despite the importance of accurate budget estimates, the problem of accuracy remains unsolved (Trost, Oberlender 2003). Most construction projects experience a certain degree of deviation between estimated and actual costs – a detailed breakdown shows the greatest deviation in the construction phase (Ökmen, Öztaş 2010). Stevens (1995) reports that 67% of the total cost of all overruns occurs in the construction phase. A similar situation was observed from an investigation on construction cost data in the present study – 154 out of the 233 cases (66%) showed a percentage error greater than +20% or less than –20%.

Meanwhile, such deviation is virtually inevitable due to the nature of estimating process and the variables involved in the process. At the preconstruction stage, estimators make many assumptions relying on limited information available at the time of estimating and the assumptions involve a considerable degree of uncertainty (Soutos, Lowe 2005). In addition, the accuracy of cost estimation is highly dependent on many influential factors involved in the process – to name a few, the availability and accuracy of information used for estimating, the depth of estimators' knowledge and experience, and the infrastructure and systems utilized for estimating. Under these circumstances, estimators often prepare range estimates or contingency plans to account for the uncertainty and errors (Peurifoy, Oberlender 2001). Therefore, achieving accurate prediction of errors associated with budget estimates can bring significant improvement to managing risks in the course of project execution and determining contingency costs.

Research efforts in estimating to date have concentrated on the identification of factors that are believed to govern construction costs, resulting in estimating models that involve various factor variables. On the other hand, there is a lack of robust methods for evaluating the accuracy of budget cost estimates (Trost, Oberlender 2003). Estimators often apply a common industry practice that uses simple statistics such as sample means and standard deviations of historical cost estimates, actual costs, and errors selected from previous completed projects. For this, estimators carefully choose sample data to secure homogeneity in data based on the similarity of job characteristics at the activity and project levels. This paper addresses the limitations of this practice that is commonly applied in the industry, and proposes an alternative approach.

Corresponding author: Seokyon Hwang
E-mail: *hwang@lamar.edu*

Taylor & Francis
Taylor & Francis Group

1. Research questions and objectives

The accuracy problem requires estimators to determine how much potential error is associated with the prepared budget estimates, which is a challenging task at the planning stage. In an effort to support the task, this study seeks an effective way to measure estimate errors in a quantitative manner while minimizing the burden of dealing with various factors. It examines an existing common industry practice widely applied in the industry and also explores an alternative method. The specific objectives of this study are to answer three research questions by testing two null hypotheses (NH) and by comparing the accuracy of forecasted estimate errors, as listed below. This paper presents an analytical procedure for probabilistic measurement of estimate errors, and demonstrates an example of its application.

- Research question 1: Can the errors of budget cost estimates be explained by a relational model? Concerning this question, the NH to be tested is that there is a strong association between estimate errors and estimates or actual costs. If the NH is rejected, then the following two research questions to account for the probabilistic nature of estimate errors;
- Research question 2: Are the errors of budget cost estimates independent of the estimates? To answer this question, the present study tests a NH that the conditional probability distribution of error, given a sample of estimates, is statistically consistent (or, similar) with the distribution, given any randomly sampled subset of the sample estimates; and
- Research question 3: Can the errors of budget cost estimates be predicted more accurately by an alternative Bayesian approach? This question is to be answered by comparing the residuals of estimate errors that are yielded by the existing industry practice and an alternative Bayesian approach.

2. Methods for cost estimating

Various methods have been developed for estimating budget costs for the construction of capital projects. These methods can be loosely categorized into four groups based on the applied approach – statistical analysis of itemized costs, investigation of the impact of cost-influencing factors on construction costs, projection of the time-dependent cost trend, and integrated analysis of multi-objectives for cost optimization.

2.1. Statistical analysis of itemized costs

The construction industry has long utilized unit costs of previous construction work to estimate construction costs for future projects. Unit costs are normally used either at the activity level or at the project level. When applying this approach, estimators select historical projects (or, activities of the projects) that are similar to a new project (or, activities of the project), and use unit costs of the selected projects (or, activities) to prepare new estimates (Stevens 1995). While this approach is simple to apply, it involves an inherent limitation attributable to the characteristics and variable conditions of projects and judgment of similarity of those across projects. These highlight that the quality of an estimate is highly dependent on judicious selection of a sample from the population of historical data. In a similar context, some research has recognized the stochastic nature of costs, recommending probabilistic estimating as an alternative approach for accounting for uncertainty in construction costs (Flood 1997). Unlike deterministic common industry practices, these approaches incorporate probability distribution of cost into estimating costs.

2.2. Factor analysis models

Many previous studies have attempted to explain construction costs as a function of factors. Wilmot and Cheng (2003) developed a multivariate regression model to predict a long-term cost trend of highway construction by fitting various indexes and bid data to the estimated price of asphalt concrete work. Soutos and Lowe (2005) investigated the relationships between cost of building elements and building characteristics, such as structural type, square footage, and major systems, resulting in a regression model. Similar efforts have been made at the activity and work package levels. Zayed and Halpin (2005) created regression models to estimate pile construction costs by fitting factor variables, including site conditions, contractor's experience, and equipment condition. Hola and Schabowicz (2010) created a method for estimating earthwork costs by applying artificial neural networks. Simulations are often applied to explain the influence of factors. Ökmen and Öztaş (2010), for instance, proposed a simulation-based model to analyse the correlation between construction costs and risk-factors. While many relational models work well with highly correlated data, historical cost data often exhibit scattered patterns between estimates and factors.

2.3. Time-dependent cost trend projection

Some research has attempted to forecast cost change trends over time. Dawood and Molson (1997) explored a number of standard forecasting techniques, arguing that the decomposition model was the most accurate. Koppula (1981) developed and compared a Box-Jenkins model and a Holt-Winters smoothing model, using the construction cost index (CCI) and the building cost index (BCI). The research showed that while both univariate time series models yielded reasonably good results, the Holt-Winter model was slightly more accurate than the other. Williams (1994) incorporated two economic indexes – prime rate and new housing-starts – in addition to the construction cost index, which resulted in a neural network (NN) model to predict short-term growth of construction cost over time. Similarly, the consumer price index together with the aforementioned indexes have been analysed and have resulted in dynamic regression models (Hwang 2009) and univariate time series models (Hwang 2011) for forecasting short- and long-term changes in construction cost. These studies showed that forecasting models excluding economic index variables produced the most accurate forecasts.

2.4. Integrated analysis of multi-objectives

Unlike the aforementioned studies, a few previous studies have approached estimating costs from the perspective of optimization of multiple project objectives. A few simulation techniques were proposed to find a plan that optimizes cost and schedule for construction, using the probabilistic distributions of costs and durations (Rao, Grobler 1995; Isidore, Back 2001). Similar attempts have been made for the same purpose by using the genetic algorithms (GA) (Li, Love 1997; Que 2002) and by combining simulation techniques and GAs (Feng *et al.* 2000). Some studies further expanded the optimization by including more project objectives such as quality and safety. Kandil and El-Rayes (2005), for example, presented a multi-objective optimization method designed based on GAs to find an optimal solution with regards to cost, duration, and quality.

2.5. Evaluation of the accuracy of cost estimates

Noting that budget cost estimates inevitably involve errors, a few studies have attempted to find effective ways for analysing the uncertainty of the estimated budget costs. Most of these have focused on identifying what factors affect the accuracy of the estimates. Trost and Oberlender (2003) proposed a factor analysis model and a multivariate regression model by analysing 45 variables in eleven groups. Concerning cost overruns in reconstruction, Attalla and Hegazy (2003) fitted a total of 36 factors in seven categories; as a result, they developed a NN model and a regression model, and compared the two models. This study concluded that both models yielded similar accuracy while the NN model was more sensitive to a larger number of variables.

Some researchers have tackled the problem in a slightly different context. For the purpose of estimating contingency, Touran (2003) developed a probabilistic model that calculates the probability of cost overruns for change order costs, the number of change orders, and variation coefficients, and estimates contingency cost depending on a given contingency level. A few studies have adopted progressive updating of forecasted costs. Barraza *et al.* (2004) proposed a probabilistic S-curve that shows a range estimate of cost per predetermined period. The range estimate is updated as actual data becomes available. Similarly, a probabilistic cost forecasting method proposed by Kim and Reinschmidt (2011) makes use of performance data collected from an ongoing project to update cost estimates during the construction period. The method allows an adaptive combination of the information used to estimate a project and the actual performance data so as to revise cost estimates.

2.6. A common industry practice

Meanwhile, estimators in the industry often apply a few relatively simple approaches to account for potential estimate errors as they complete detail estimating. Most of the time, they rely on historical cost data (estimates and actual costs) collected from previous projects. For this, estimators make judicious selection of data through data sampling – they evaluate the similarity of project and activity characteristics between previous projects and a new project. For instance, let us assume $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are selected sample estimates and actual costs, respectively, and $\boldsymbol{X}_3$ is sample estimate errors ($\boldsymbol{X}_2 - \boldsymbol{X}_1$). Given the selected sample, they find its simple statistics – the sample average and standard deviation of sample. Using the statistics, they often decide a range of cost estimate or a contingency cost, given a desired confidence interval for the prepared cost estimate. Similarly, they often forecast estimate error by taking the sample average of $\boldsymbol{X}_3$. In this process, in-depth analysis of the probability distribution of sample that is selected from the population of historical data is rarely performed. Such a lack of understanding of the probability distribution of the sample used can lead to a misleading error calculation. This issue is investigated in depth in the present study.

3. Measurement of cost estimate errors

3.1. A statistics for measurement

Cost estimate error refers to the difference between estimated and actual costs. Budget costs are often updated for various reasons. In the event of having updated budget costs, the updated estimate is considered as the budget estimate in this study. A simple statistic is used in this study to quantify the amount of an estimate error in rate (error rate, *er*). Given an activity, its *er* is calculated by Eqn (1) where e_p and a_p represent an estimated budget cost and the actual cost at completion respectively for the activity of a project (p):

$$er_p = \frac{a_{p-}e_p}{e_p}, p = 1, 2, ..., P. \quad (1)$$

Thus, there are as many *er*'s for an activity of the same type as projects that have the activity. The negative value of er_p indicates an overestimate ($e_p > a_p$), whereas a positive value indicates an underestimate ($a_p > e_p$). Thus, it can be said that the smaller the absolute value of er_p is, the more accurate the estimate is.

Meanwhile, both cases can cause risks to project organizations. For example, if a contractor prepared an underestimated budget for a project and won the project at bidding, the contractor would suffer from cost overruns during construction of the project. On the contrary, if the contractor's budget cost was overestimated, the overestimated budget can bring a totally different result to the contactor at bidding where overestimated budget can prevent the contractor from winning the project.

3.2. A snapshot of estimate accuracy

As discussed earlier, it is hardly possible to predict budget cost estimates without errors. Figure 1, although not representative of the entire construction industry, gives a fairly good snapshot of the accuracy of budget estimates. The graph shows the calculated error rate (*er*) of 233 sets of unit cost data (124 formwork activities and 109 concrete pouring/finishing activities) that were collected from reinforced concrete work of medium size building projects that were constructed by a general contractor in a metropolitan area. Each set comprises an estimated cost and its corresponding actual cost at completion. In particular, the formwork activities were performed by the workers of self-performing group of the company.

As shown in Figure 1, the error rates are widely distributed while the majority falls in between ±50% error range. 117 cases were found to have an error rate less than or equal to 0.0 and 116 cases greater than 0.0. Out of the 233, only 78 cases fell in between ±20% error. Also, it is noticeable that the error rates loosely fit a normal distribution curve. In normal distribution, the shape of histogram to the left of the distribution peak is roughly a mirror image of the shape of the histogram to the right of the peak (Nolan, Speed 2000). This shape is observed from the data set in Figure 1 exhibiting an approximately symmetric distribution of frequency around 0.0. The cumulative probability (c.d.f.) curve in the graph follows a typical example of a distribution curve with short tails. This normality of sample allows convenience in statistical analysis of the sample, because many statistical analysis techniques are applied under the condition of normality (Box *et al.* 1978).

4. Research methodology

To answer the three research questions, this study has followed a procedure of statistical analysis. The following briefly describes the standard statistical techniques that are applied in the procedure.

4.1. Outlier tests

It is not unusual to observe outliers from a sample of data. Since outliers may greatly influence the result of the statistical analysis, it is necessary to remove them (Box *et al.* 1978). The Box-and-Whisker plot, also known as the Box-plot, is a generic statistical technique that is useful to identify outliers in a sample. The plot graphically shows location, dispersion, and outliers of a sample, and may indicate skewness and tail size of the distribution of a sample (Nolan, Speed 2000). The technique uses boundary values – lower quartile (Q_1, value of 25th percentile) and upper quartile (Q_3, value of the 75th percentile) – to determine outliers. Outliers are defined as data points falling beyond the range between upper limit ($Q_3 + 1.5 \times IQR$) and lower limit ($Q_1 - 1.5 \times IQR$) where inter-quartile range (*IQR*) is Q_3 minus Q_1.

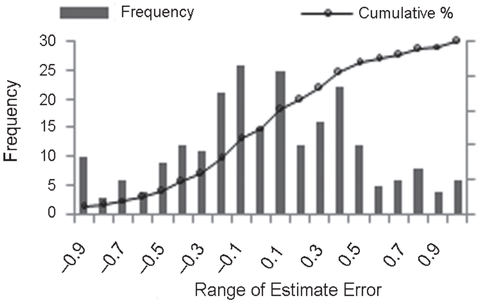

Fig. 1. A distribution of error rates (*er*) of the population of 233 unit costs

4.2. Correlation check

Following the outlier test, a correlation check between variables is needed. Given the random variables *X* and *Y*, it is necessary to examine if *X* is highly correlated with *Y*. The existence of correlation can be determined by measuring the correlation coefficient between the two variables that quantitatively represents the degree of association between those (Hogg, Craig 1995). The correlation coefficient between the two random variables *X* and *Y* is given by Eqn (2) where σ_X and σ_Y are standard deviations of *X* and *Y*, and σ_{XY} is covariance of *X* and *Y*. The parameter is greater than or equal to –1.0 and less than or equal to 1.0. High correlation indicates the evidence of strong association, so that the associated random variables can be explained by a relational model:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}. \qquad (2)$$

4.3. Distribution check

Using the three variables – budget cost estimates (***e***), actual costs (***a***), and error rates (***er***) – that are analysed in the present study, let us assume a sample space ***e*** that consists of subset e_k's, $k = 1,2,\dots K$. If the probability distributions of er_i's for e_k's are the same or similar, then, the conditional probability of er_i for a given subset e_k is equal to the probability of er_i for the sample space ***e***. The following hypothetical example illustrates the underlying logic. In this example, sample spaces ***e*** and ***er*** having 1000 samples, respectively, are randomly generated. Assume that that e_k's and er_i's are uniformly distributed from 1.0 to 5.0 and from –1.0 and 1.0, respectively. Given the samples, the probability that er_i is less than or equal to 0.0, $P(er_i \le 0.00)$ for a subset $(2.0 \le e_k \le 2.1)$, is compared to the probability of the same er_i, given the whole sample space ***e***. The former probability was found to be 0.412 (7 out of 17) and the latter probability was 0.414 (414 out of 1000). This indicates that if the probability distribution of error for each subset of ***e*** is consistent, the probability for a certain level of error will be statistically similar and follows that for the entire sample ***e***. In this case, the estimation of expected error can be measured based on the distribution of er_i, given ***e***; otherwise, the conditional probability of error rates needs to be considered.

4.4. Bayesian probability calculation

The Bayesian approach is an application of the calculation of conditional probability for which Bayes' theorem is used as the basis and the calculated probability is often called Bayesian probability (Box *et al.* 1978). Eqn (3) represents the Bayes' theorem that evolved from the definition of conditional probability and the law of total probability where the sample space is partitioned into *n* mutually exclusive and exhaustive events $C_1, C_2, ..., C_I$. Using the theorem, the conditional probability of a particular event C_j, given the event space C, is calculated from the probabilities of each event $C_1, C_2, ..., C_I$ and the conditional probabilities of C, given each event C_i, $i = 1, 2, ..., I$ (Hogg, Craig 1995):

$$P(C_j \mid C) = \frac{P(C_j)P(C \mid C_j)}{\sum_{i=1}^{n} p(C_i)p(C \mid C_i)}. \quad (3)$$

5. An analytical procedure for forecasting estimate errors and an application example

This section describes an analytical procedure that is built on well-established statistical techniques for probabilistic measurement of budget cost estimate errors by analysing historical estimates, actual costs, and estimate errors. The procedure explores and tests the existing common industry practice and an alternative Bayesian approach.

5.1. Preparation of sample data

The first step is to retrieve historical cost data – budget cost estimates (***e***) and actual costs (***a***) – from completed projects. A rule of thumb for such sampling is to select projects that share similar project characteristics, so as to maximize homogeneity in data. Meanwhile, quantities of activities of the same type can vary across projects, which results in different costs for the activities. Estimators normally select a unit cost and multiply the unit cost to the quantity of an activity. For these, it is reasonable to retrieve the cost data in terms of unit of work. Once ***e*** and ***a*** are retrieved, ***er*** can be calculated by using Eqn (1). Most construction companies store such data in project databases, thus, it is not difficult to acquire those. For an application example, an activity, forming concrete wall, was selected, applying the rule of thumb. Historical unit costs of the activity were collected from 30 similar projects from the 233 sets described earlier (Table 1).

5.2. Removal of outliers

A Box-plot test identified six outliers among 30 sets of data sample. The boundary and determinant values of Box-plot are as follows: 0.03 (Q_1), 0.44 (Q_3), 0.41 (I*Q*R), 1.06 (upper limit), and –0.58 (lower limit). The identified outliers are 9, 13, 18, 20, 27, and 29 (refer to the first column (*p*) of Table 1. Consequently, 24 sets of data are found to be valid.

Table 1. A sample of unit costs (* outlier)

p	e_p	a_p	er_p
1	4.66	4.17	–0.10
2	2.62	3.36	0.28
3	2.92	4.70	0.61
4	2.77	3.77	0.36
5	2.78	3.27	0.18
6	3.47	4.06	0.17
7	3.47	4.25	0.23
8	3.47	4.21	0.21
*9	3.47	7.39	1.13
10	2.96	4.35	0.47
11	3.08	2.66	–0.14
12	6.38	6.04	–0.05
*13	5.51	13.45	1.44
14	4.43	4.58	0.03
15	3.86	3.99	0.03
16	3.86	4.05	0.05
17	2.48	3.23	0.30
*18	2.48	7.31	1.95
19	2.78	3.64	0.31
*20	2.78	7.04	1.53
21	5.39	6.49	0.20
22	10.24	5.68	–0.45
23	4.02	3.61	–0.10
24	4.02	3.42	–0.15
25	4.02	4.51	0.12
26	7.48	9.00	0.20
*27	6.60	23.29	2.53
28	3.30	3.49	0.06
*29	3.30	8.46	1.56
30	4.87	2.05	–0.58

5.3. Measurement of the degree of correlation

Once outliers are removed from the sample, a correlation check is followed to answer the research question 1. This tests the null hypothesis that there are statistically significant correlations among the variables (***e***, ***a***, and ***er***) that are believed to be potentially related to each other. Given the sample (Table 1), the calculated correlation coefficient for each pair of variables is as follows: 0.66 for ***e*** vs. ***a***, –0.56 for ***er*** vs. ***e***, and 0.17 for ***er*** vs. ***a***. None of these coefficients suggests evidence of statistically significant relationships among the three variables. Meanwhile, such low correlation could be true only for the specific sample. In order to examine this, the correlation coefficients for the aforementioned 233 sets of data, from which the sample (Table 1) was selected, were additionally calculated. The calculated correlations were found to be 0.67 for ***e*** vs. ***a***, 0.20 for ***er*** vs. ***e***, and 0.18 for ***er*** vs. ***a***. Both results show that the probability of strong correlation among the variables is low. In principle, there can be strongly correlated data – in such cases, relational models generally can be pursued.

5.4. Examination of conditional dependency

The results of Section 5.3 leads to investigating how to account for the probability distribution of estimate errors when using historical data to estimate future work. The present study notes a common way of using historical data that estimators follow to select a cost estimate. As noted earlier, estimators consider homogeneity in data by comparing the similarity of conditions between future work and previously completed work. Consequently, activities sharing similar conditions tend to have similar estimates. A rationale behind this is that activities under similar conditions will have similar cost performance. From the rationale, a hypothetical belief – the estimate error of an activity tends to follow the behaviours of the errors of its similar estimates – can be established.

The above is a typical problem of statistical dependence between random variables which can be effectively examined by applying random sampling and conditional probability distribution theory (Box *et al.* 1978). With regard to the aforementioned beliefs, the dependency of $\boldsymbol{er}$ on $\boldsymbol{e}$ can be examined. To examine the dependency, let us assume a sample space that is partitioned into n mutually exclusive and exhaustive events er_i, $i = 1, 2, \ldots, n$, and an event space $\boldsymbol{e}$ (Fig. 2). Each subset er_i and the event space $\boldsymbol{e}$ in the diagram correspond to C_i and C in the above Eqn (3), respectively. The event space $\boldsymbol{e}$ in Figure 2 is also divided into a number of subset e_k's, $k = 1, \ldots, K$. If the conditional probability distributions of er_i, given $\boldsymbol{e} = e_k$, say, $P(er_i \mid \boldsymbol{e} = e_k)$'s are statistically consistent (or, similar), then it can be said that $\boldsymbol{er}$ (or, all er_i's) is independent of $\boldsymbol{e}$ where each e_k is a randomly selected subset from the entire event space; otherwise, $\boldsymbol{er}$ is dependent on e_k's.

For an application example, assume that $\boldsymbol{er}$ has six ranges, er_i, $i = 1, 2, \ldots, 6$ (Table 2). Two subsets of the event space $\boldsymbol{e}$ are also selected in terms of range – $3.0 < e_2 \leq 4.0$ and $4.0 < e_3 \leq 5.0$ (Table 3). Then, the conditional probability distributions are calculated for six er_i's, given e_2 and e_3, say, $P(er_i \mid \boldsymbol{e} = e_2)$ and $P(er_i \mid \boldsymbol{e} = e_3)$. A paired t-test can be effectively used to examine the congruency of the conditional probabilities. Given the alpha of 0.1 and 0.2 and the degree of freedom of 5, critical values of two-sided t-test are $\pm t_{0.1,5} = \pm 2.01$ and $\pm t_{0.2,5} = \pm 1.47$, whereas calculated t-value for $P(er_i \mid \boldsymbol{e} = e_2)$ vs. $P(er_i \mid \boldsymbol{e} = e_3)$ is –5.00. The test result shows that the conditional probability distributions of error rates, given e_2 and e_3, are significantly different. There is no reason to believe that the probability distributions of $\boldsymbol{er}$ for each subset e_k's are consistent and follow the distribution for the entire sample $\boldsymbol{e}$. There can be conditional dependency between estimates and estimate errors. Thus, the null hypothesis for the research question 2 should be rejected.

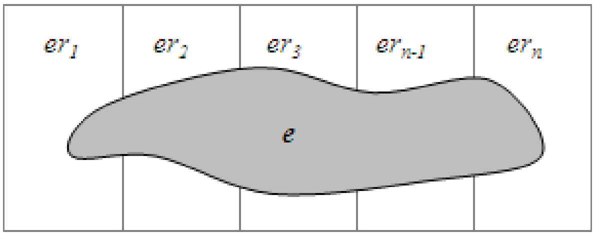

Fig. 2. A sample space and event space of events $\boldsymbol{er}$ and $\boldsymbol{e}$

Table 2. Probability of er_i, $P(er_i)$, for the entire event space $\boldsymbol{e}$

i	er_i	Occurrence	$P(er_i)$	c.d.f. $P(er_i)$
1	≤–0.50	1	0.04	0.04
2	–0.50~–0.25	1	0.04	0.08
3	–0.25~0.00	5	0.21	0.29
4	0.00~0.25	11	0.46	0.75
5	0.25~0.50	5	0.21	0.96
6	0.50<	1	0.04	1.00

Table 3. Event space e_k and probability of e_k, $P(e_k)$

k	e_k	Occurrence	$P(e_k)$	c.d.f. $P(e_k)$
1	≤ 3.0	7	0.29	0.29
2	3.0~4.0	7	0.29	0.58
3	4.0~5.0	6	0.25	0.83
4	5.0<	4	0.17	1.00

5.5. An alternative Bayesian approach

The results of Sections 5.3 and 5.4 support the choice of Bayesian approach for dealing with the nature of budget estimate errors. The following discusses the process of Bayesian probability calculation, the dependence of estimate errors on similar estimates, and a comparison of accuracy between the Bayesian approach and the existing common industry practice.

Given an estimate e_k for a future work we can calculate using Bayes' theorem $P(er_i \mid e_k)$ from $P(er_i)$, er_i, $i = 1, 2, \ldots, I$, and $P(e_k \mid er_i)$ for er_i, $i = 1, 2, \ldots, I$. In the context of Bayesian statistics, $P(er_i \mid \boldsymbol{e})$ is a prior distribution of the probability of estimate error, given all observations of a selected sample. On the other hand, $P(er_i \mid e_k)$ is a posterior distribution of the probability, given the observations in a subset (similar estimates) of the selected sample, with the likelihood $P(e_k \mid er_i)$.

For an application example, let us assume a scenario that an estimator has selected a budget cost estimate e_3 $(4.0 < e_3 \leq 5.0)$ for a formwork activity (Table 3). The estimator wants to know the probability that the selected estimate's er would fall between –0.25 and 0.00, $(-0.25 < er_3 \leq 0.00)$, say, $P(-0.25 < er_3 \leq 0.00) \mid 4.0 < e_3 \leq 5.0)$. In this scenario, the sample space $\boldsymbol{er}$ again is assumed to be divided into six ranges as shown in Table 2.

If the prior belief is true, then, by the statistical dependence theory, the probability distributions produced by two different approaches – an industry common practice and an alternative Bayesian approach – should be statistically similar. This was tested by comparing probabilities calculated by the approaches for the same estimates. Table 4 presents the calculated probabilities in terms of cumulative probability (c.d.f.) – "common-c.d.f." and "Bayesian c.d.f.". The common-c.d.f. is computed by the industry common practice using the entire 24 valid data sets. For example, let us assume that an estimator selects 4.66 as an estimate. In this case, the error rates of the estimate are computed by subtracting 4.66 from each actual cost of the 24 sets and dividing the difference by 4.66,

Table 4. Hypothesis test summary

e	Bayesian-c.d.f. (1)	Common-c.d.f.(2)	t-test
4.02	0.17	0.00	Number of pairs
	0.17	0.08	36
	0.67	0.46	
	1.00	0.83	Degree of freedom
	1.00	0.88	35
	1.00	1.00	
			Two given alphas
4.02	0.17	0.00	0.1, 0.2
	0.17	0.08	
	0.67	0.46	Critical values of
	1.00	0.83	two-sided tests,
	1.00	0.88	$t_{0.1,35}$ = 1.306,
	1.00	1.00	$t_{0.2,35}$ = 1.689
4.02	0.17	0.00	Calculated t-value
	0.17	0.08	of test,
	0.67	0.46	3.953
	1.00	0.83	
	1.00	0.88	
	1.00	1.00	
4.43	0.17	0.04	
	0.17	0.17	
	0.67	0.71	
	1.00	0.83	
	1.00	0.96	
	1.00	1.00	
4.66	0.17	0.04	
	0.17	0.25	
	0.67	0.79	
	1.00	0.88	
	1.00	0.96	
	1.00	1.00	
4.87	0.17	0.04	
	0.17	0.38	
	0.67	0.83	
	1.00	0.92	
	1.00	0.96	
	1.00	1.00	

(1) c.d.f. produced by a Bayesian approach.

(2) c.d.f. produced by the industry common practice.

which results in 24 estimated error rates. Given the 24 error rates, the probability of er_i, i = 1, 2, …, 6 (refer to Table 2), are computed to produce the common-c.d.f for the selected estimate 4.66. The same process is repeated for the remaining estimates considered in the scenario, e_3 (4.0 < e_3 ≤ 5.0). Meanwhile, the estimator can compute the Bayesian-c.d.f. for er_i, i = 1, 2, …, 6, given the same e_3, by repeating the aforementioned process of Bayesian probability calculation using Eqn (3) (refer to Section 4.4). Given the six estimates falling in the range 4.0 < e_3 ≤ 5.0 and the six ranges of er_i, i = 1, 2, …, 6, 36 cumulative probabilities can be computed as shown in Table 4. The result of two-sided paired t-test for the 36 sets is summarized in Table 4. Given the test result, there is no evidence to accept the null hypothesis for the research question 2. In addition to the test result in Section 5.4, taken with a larger sample, this test also confirms the risk of using the entire sample space without considering the dependence of estimate errors on similar estimates.

5.6. Application of the Bayesian approach

Applying the two approaches to the 24 valid data in Table 1, estimate errors are forecasted. The common industry practice, using all observations in the sample space, yields an average of the 24 estimate errors, $\hat{er}_S$= sample mean of er_p's, p = 1, 2, …, 24. Meanwhile, the alternative Bayesian approach gives an average of errors falling in each error range, $\hat{er}_R$ = sample mean of er_p's within a range, $p = 1, 2, \ldots, 24$. Given these estimate error forecasts, an approach that yields a smaller absolute residual of forecasted error should be regarded as a more accurate method. In a statistical sense, it can be determined by comparing the mean variance of the residuals from each method. The last two columns, %D_1 and %D_2, in Table 5 presents the residual percentage yielded from the two approaches. The averages of variances of %D_1, 100 x($\hat{er}_S - er_p$), and %D_2, 100 x ($\hat{er}_R - er_p$) are calculated to be 712.56 and 345.17, respectively. The result is in favour of the Bayesian approach. This means that the error of a new estimate that falls in a subset of a sample space will more likely follow the probability distribution of the errors of estimates in the subset. Considering the results of the hypothesis tests in Sections 5.4 and 5.5 and the accuracy comparison in Section 5.6, it is rational to believe that the whole sample space needs to be divided into subsets in accordance with the similarity of estimate values.

Table 5. Comparison of the accuracy of forecasted errors

e	e	a	er	$\hat{er}_S$	$\hat{er}_R$	%D_1	%D_2
≤3.0	2.48	3.23	0.30	0.09	0.36	–20.77	5.68
	2.62	3.36	0.28			–19.04	7.41
	2.77	3.77	0.36			–26.41	0.05
	2.78	3.27	0.18			–8.28	18.17
	2.78	3.64	0.31			–21.43	5.02
	2.92	4.70	0.61			–51.48	–25.03
	2.96	4.35	0.47			–37.76	–11.30
3.0~4.0	3.08	2.66	–0.14		0.09	23.21	22.61
	3.30	3.49	0.06			3.50	2.90
	3.47	4.06	0.17			–7.72	–8.32
	3.47	4.25	0.23			–13.29	–13.89
	3.47	4.21	0.21			–11.94	–12.54
	3.86	3.99	0.03			6.08	5.48
	3.86	4.05	0.05			4.35	3.75
4.0~5.0	4.02	3.61	–0.10		–0.13	19.67	–2.70
	4.02	3.42	–0.15			24.38	2.01
	4.02	4.51	0.12			–2.92	–25.28
	4.43	4.58	0.03			6.00	–16.37
	4.66	4.17	–0.10			19.79	–2.58
	4.87	2.05	–0.58			67.29	44.92
5.0<	5.39	6.49	0.20		–0.02	–11.00	–22.69
	6.38	6.04	–0.05			14.80	3.11
	7.48	9.00	0.20			–10.92	–22.61
	10.24	5.68	–0.45			53.88	42.19

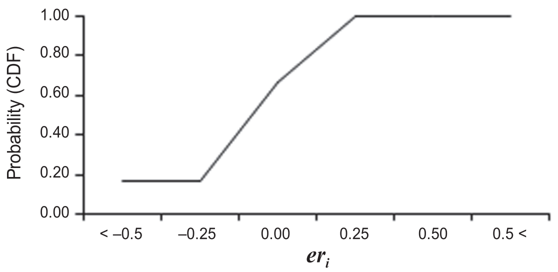

Fig. 3. Probability of potential errors, given an estimated unit cost *e3*, 4.0 < *e3* ≤ 5.0

The calculated Bayesian probability information is interpreted as the probability of expected error rate er_i, given an initial estimate e_k. In other words, the budget cost estimate e_k may involve as much as er_i with a probability $P(er_i \mid e_k)$. Using the information, estimators can predict in a probabilistic manner the expected actual costs at completion, given selected cost estimates. Figure 3 presents the cumulative distribution of the calculated Bayesian probability for the aforementioned scenario. The resulting probability is interpreted as follows. When an initial estimate e_3 is selected, there is a 67% chance that the selected estimate can have an error rate less than or equal to 0.00. This can be also said that there is a 67% chance that the actual cost at completion will be less than or equal to the estimated cost or there is a 33% chance that the actual cost can be greater than its estimated cost.

Conclusions

Budget cost estimates are important early decisions in managing construction projects. This study has addressed the limitations of the existing common industry practice for forecasting errors associated with budget cost estimates. A belief lies behind this practice – it is rational to forecast probability of estimate errors based on the distribution of all observed errors of a sample that is selected based on the similarity of conditions between future work and previously completed work. The statistical analysis on historical data has revealed the existence of conditional dependence of errors on similar estimates. The proposed approach incorporates the posterior information to compute the probability of estimate errors.

The findings of the present study suggest that a Bayesian probability approach is more effective than the existing practice for the application of historical cost data to predict estimate errors. The Bayesian approach is able to predict estimate errors more accurately than the current approach, taking into consideration the stochastic nature of estimate errors. It is envisioned that estimators can take advantage of the alternative approach to prepare reliable budget plans and contingency plans. Thereby, the approach can complement the existing methods for managing project uncertainty and risks due to inaccurate cost estimates.

The presented analytical procedure is easy to apply to examine the stochastic nature of historical estimates and estimate errors. Meanwhile, the procedure and the Bayesian approach can be more effectively applied when there is a large sample of historical cost data. A large sample that is more densely distributed can allow more precise forecasting with much shorter ranges of estimates and errors. While the distribution of error rates is unlikely dependent on the size of a subset, as shown earlier by the normality of the distribution of error rates, a large sample can reduce any potential influence of inconsistent size of each subset. It should also be noted that prior to generalization of the Bayesian approach, additional tests need to be conducted to verify the approach for more types of work other than those tested in this study, including data sets that are not highly homogeneous.

References

Attalla, M.; Hegazy, T. 2003. Predicting cost deviation in reconstruction projects: artificial neural networks versus regression, *Journal of Construction Engineering and Management* 129(4): 405–411.
http://dx.doi.org/10.1061/(ASCE)0733-9364(2003)129:4(405)

Barraza, G. A.; Back, W. E.; Mata, F. 2004. Probabilistic forecasting of project performance using stochastic S Curves, *Journal of Construction Engineering and Management* 130(1): 25–32.
http://dx.doi.org/10.1061/(ASCE)0733-9364(2004)130:1(25)

Box, G. E. P.; Hunter, W. G.; Hunter, J. S. 1978. *Statistics for experimenters: an introduction to design, data analysis, and model building*. New York: John Wiley & Sons. 653 p.

Dawood, N.; Molson, A. 1997. An integrated approach to cost forecasting and construction planning for the construction industry, in *Proc. of the 4th Congress on Computing in Civil Engineering,* 16–18 June 1997, Philadelphia, Pennsylvania, United States, 535–542.

Feng, C.; Liu, L.; Burns, S. A. 2000. Stochastic construction time-cost trade-off analysis, *Journal of Computing in Civil Engineering* 4(2): 117–126.
http://dx.doi.org/10.1061/(ASCE)0887-3801(2000)14:2(117)

Flood, I. 1997. Modelling uncertainty in cost estimates: a universal extension of the central limit theorem, in *Proc. of 4th Congress on Computing in Civil Engineering,* 16–18 June 1997, Philadelphia, Pennsylvania, United States, 551–558.

Hogg, R. V.; Craig, A. T. 1995. *Introduction to mathematical statistics*. 5th ed. New Jersey, Upper Saddle River: Prentice-Hall, Inc. 564 p.

Hola, B.; Schabowicz, K. 2010. Estimation of earthworks execution time cost by means of artificial neural networks, *Automation in Construction* 19(5): 570–579. http://dx.doi.org/10.1016/j.autcon.2010.02.004

Hwang, S. 2009. Dynamic regression models for prediction of construction costs, *Journal of Construction Engineering and Management* 135(5): 360–367. http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0000006

Hwang, S. 2011. Time series models for forecasting construction costs using time series indexes, *Journal of Construction Engineering and Management* 137(9): 656–662. http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0000350

Isidore, L. J.; Back, W. E. 2001. Probabilistic optimal-cost scheduling, *Journal of Construction Engineering and Management* 127(6): 431–437.
http://dx.doi.org/10.1061/(ASCE)0733-9364(2001)127:6(431)

Kandil, A.; El-Rayes, K. 2005. Multi-objective optimization for the construction of large-scale infrastructure systems, in *Proc. of the 2005 Construction Research Congress*, 5–7 April 2005, San Diego, California, United States, 1–11. http://dx.doi.org/10.1061/40754(183)99

Kim, B.; Reinschmidt, K. 2011. Combination of project cost forecasts in earned value management, *Journal of Construction Engineering and Management* 137(11): 958–966. http://dx.doi.org/10.1061/(ASCE)CO.1943-7862.0000352

Koppula, S. D. 1981. Forecasting engineering costs: two case studies, *Journal of Construction Division* 107(CO4): 733–743.

Li, H.; Love, P. 1997. Using improved genetic algorithms to facilitate time-cost optimization, *Journal of Construction Engineering and Management* 123(3): 233–237. http://dx.doi.org/10.1061/(ASCE)0733-9364(1997)123:3(233)

Nolan, D.; Speed, T. 2000. *Stat labs – mathematical statistics through applications*. New York: Springer. 282 p.

Ökmen, Ö.; Öztaş, A. 2010. Construction cost analysis under uncertainty with correlated cost risk analysis model, *Construction Management and Economics* 28(2): 203–212. http://dx.doi.org/10.1080/01446190903468923

Peurifoy, R. L.; Oberlender, G. D. 2001. *Estimating construction costs*. 5th ed. New York: McGraw-Hill. 512 p.

Que, B. C. 2002. Incorporating practicability into genetic algorithm-based time-cost optimization, *Journal of Construction Engineering and Management* 128(2): 139–143. http://dx.doi.org/10.1061/(ASCE)0733-9364(2002)128:2(139)

Rao, G. N.; Grobler, F. 1995. Integrated analysis of cost risk and schedule risk, in *Proc. of the 2nd Congress on Computing in Civil Engineering*, 5–8 June 1995, Atlanta, Georgia, United States, 1404–1411.

Soutos, M.; Lowe, D. J. 2005. ProCost – towards a powerful early stage cost estimating tool, in *Proc. of the International Conference on Computing in Civil Engineering,* 12–15 July 2005, Cancun, Mexico, 1–12. http://dx.doi.org/10.1061/40794(179)142

Stevens, J. D. 1995. *Cost estimating and forecasting for highway work in Kentucky*. Research Rep. KTC 95-12, Kentucky Transportation Center, University of Kentucky, Lexington, KY.

Touran, A. 2003. Probabilistic model for cost contingency, *Journal of Construction Engineering and Management* 129(3): 280–284. http://dx.doi.org/10.1061/(ASCE)0733-9364(2003)129:3(280)

Trost, S. M.; Oberlender, G. D. 2003. Predicting accuracy of early cost estimates using factor analysis and multivariate regression, *Journal of Construction Engineering and Management* 129(2): 198–204. http://dx.doi.org/10.1061/(ASCE)0733-9364(2003)129:2(198)

Williams, T. P. 1994. Predicting changes in construction cost indexes using neural networks, *Journal of Construction Engineering and Management* 120(2): 306–320. http://dx.doi.org/10.1061/(ASCE)0733-9364(1994)120:2(306)

Wilmot, C. G.; Cheng, G. 2003. Estimating future highway construction costs, *Journal of Construction Engineering and Management* 129(3): 272–279. http://dx.doi.org/10.1061/(ASCE)0733-9364(2003)129:3(272)

Zayed, T.; Halpin, D. 2005. Productivity and Cost Regression Models for pile construction, *Journal of Construction Engineering and Management* 131(7): 779–789. http://dx.doi.org/10.1061/(ASCE)0733-9364(2005)131:7(779)

Seokyon HWANG. Associate Professor in the Reese Construction Management Program at Lamar University (USA). He received his PhD in Civil Engineering at the Department of Civil and Environmental Engineering at the University of Illinois at Urbana-Champaign (2007). His research interests include risk analysis and decision-making for project planning/programming, integration of information technology into management of field operation, application of Building Information Modelling, and best practices and knowledge for sustainable organizational learning.