Interaction between Theory and Practice in Civil Engineering and Construction
Edited by Komurlu, R., Gurgun, A. P., Singh, A., and Yazdani, S.

ISBN: 978-0-9960437-2-4

A BAYESIAN-BASED MODEL FOR ALLOCATING CONTINGENCY TO A PORTFOLIO OF CORRELATED CONSTRUCTION PROJECTS

PAYAM BAKHSHI[1] and ALI TOURAN[2]

[1]*Dept of Construction Management, Wentworth Institute of Technology, Boston, MA*
[2]*Dept of Civil and Environmental Engineering, Northeastern University, Boston, MA*

Despite the availability of various probabilistic and non-probabilistic methods for determining budget contingency, still many large capital projects are suffering from cost overrun. Furthermore, most of the effort has been concentrated on calculation of contingency for a single project rather than a group of projects. This paper introduces a new Bayesian-based model for allocating contingency budget to a portfolio of correlated construction projects. The proposed model enables an owner agency to define the individual project confidence level for contingency calculation taking into account the portfolio budget and the desired portfolio confidence level. The model recognizes the correlation between each pair of projects in the portfolio and calculates the required increase in budget in such a way to ensure adequate budgets with respect to individual projects and the portfolio. Using the information from newly completed projects, the Bayesian technique can be used to update the model parameters periodically so that more accurate contingency budget can be established for the portfolio. A numerical example is presented to show the application of the model on a portfolio of transit projects. The proposed model can be employed as an effective tool for the owner agencies in charge of funding a group of projects every year.

Keywords: Probabilistic model, Risk, Monte Carlo, Truncated normal distribution, Pearson correlation.

1 INTRODUCTION

In 2012, Bakhshi and Touran proposed a model for calculation of contingency in a portfolio of construction projects. The model assumes normal distribution for the cost overruns/ underruns and truncated normal distribution for the cost of each project in the portfolio. The promise of the model is to protect a portfolio of projects against cost overrun by adjusting their original budgets. The proposed model helps an agency find the level of confidence needed at the individual project level to ensure that the portfolio budget will meet the minimum level of confidence based on available funding and the agency's policy goals. A Bayesian approach is employed to update the model on regular intervals. As more information becomes available in the future, the required adjustment in portfolio budget will be reduced, because the accuracy of estimating the contingency is improved. However, the limitation of this model is that it is not considering the correlation among the projects in Bayesian approach when updating the model. This current research is a continuation of Bakhshi and Touran (2012) where

dependencies (Pearson correlation coefficient) between each pair of projects in the portfolio is recognized. This enhances the efficiency and accuracy of the model.

2 BASIS OF THE MODEL

The model presumes truncated normal distribution for the cost of each project in the portfolio. To form this distribution, it is assumed that the probability of experiencing underrun m is α as the discrete portion of distribution. m is an arbitrary number based on agency's objectives and α can be determined by reviewing the historical cost overruns/underruns. Let us assume that there is a database of construction projects comprising of $i = 1, \ldots, n$ projects with the initial budget of b_i. It is found through this historical data that there is α% chance to have m percent underrun and get the project done with $c_i = (1-m).b_i$. Figure 1 depicts the model steps, the parameters, and their definitions as used in the process. Interested readers can refer to Bakhshi and Touran (2012) for details of the model and how the equations were derived.

$$\eta = \Phi\left\{\Phi^{-1}(\alpha) + \left[(1-\alpha).\Phi^{-1}(\alpha) - \frac{e^{\frac{-[\Phi^{-1}(\alpha)]^2}{2}}}{\sqrt{2\pi}}\right].\left[\frac{\varphi.\sqrt{\sum_{i=1}^{n} c_i^2 + 2\sum_{i<j}^{n}\sum_{j}^{n}\rho_{ij}c_i c_j}.\Phi^{-1}(\gamma)}{(1-m).(1-\rho).B} - 1\right]\right\} \quad (1)$$

$$\frac{B^*}{B} = (1-m).\left[1 + \frac{(1-\rho).[\Phi^{-1}(\eta) - \Phi^{-1}(\alpha)]}{(1-\alpha).\Phi^{-1}(\alpha) - \frac{e^{\frac{-[\Phi^{-1}(\alpha)]^2}{2}}}{\sqrt{2\pi}}}\right] \quad (2)$$

3 BAYESIAN APPROACH FOR UPDATING THE MODEL WITH *k* CORRELATED PROJECTS

It is assumed that k recently completed and correlated projects need to be updated. To conduct the analysis with correlated cost overruns/underruns, the joint density function of project cost overruns/underruns is required; the probability distribution of each project's cost overrun/underrun is the marginal distribution. Knowing the marginal distributions of cost overruns/underruns is not sufficient to obtain their joint density function. Multivariate normal distribution is the special case in which the only information required other than marginal distribution of each random variable is the values of covariance among the variables (Rowe 2003). When there is a multivariate normal distribution, each of its marginal variables by itself is normally distributed. The converse, however, is not generally true (Kutner *et al*. 2005). Despite this, a simplifying assumption that the joint density function of the cost overruns/underruns to be a multivariate normal distribution. The multivariate normal PDF is:

$$f(\delta_1, \ldots, \delta_k) = \frac{\left|\mathbf{V}^{-1}\right|^{1/2}}{(2\pi)^{n/2}}.\exp\left[-\frac{1}{2}.\left(\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}\right)'.\mathbf{V}^{-1}.\left(\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}\right)\right] \quad (3)$$

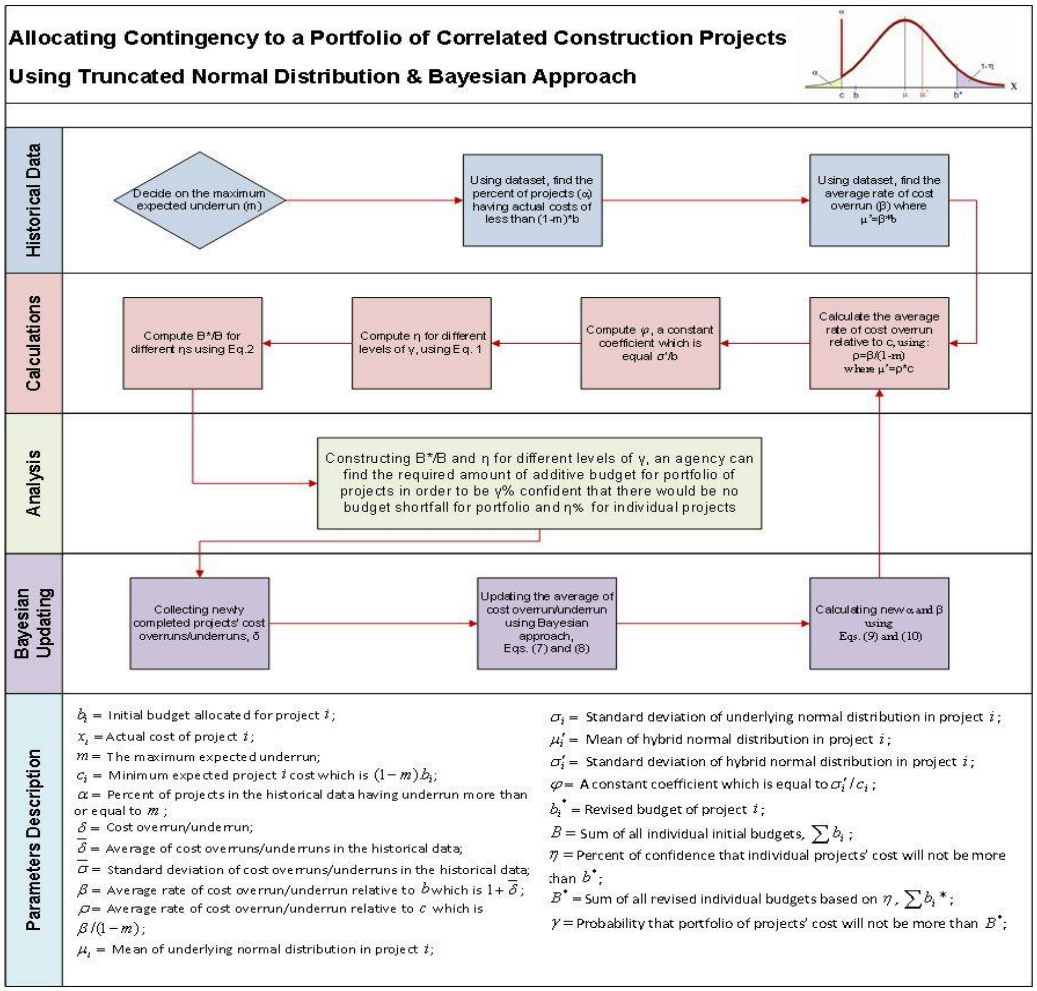

Figure 1. The model steps and parameters.

Where all bold letters represent a matrix/vector and: (1) $\boldsymbol{\delta} = (\delta_1, ..., \delta_k)^T$ are the cost overruns/underruns of k newly completed projects (T denotes the transpose of matrix). (2) $\bar{\boldsymbol{\delta}} = (\bar{\delta}_1, ..., \bar{\delta}_k)^T$ are the means of each cost underruns/overruns distribution. Since it was assumed that all these k projects are coming from a normal population with the mean $\bar{\delta}$, $\bar{\boldsymbol{\delta}}$ can be written as $\bar{\boldsymbol{\delta}} = (\bar{\delta}, ..., \bar{\delta})^T$. (3) $\mathbf{V}$ is the variance-covariance matrix which is a symmetrical ($k \times k$) matrix. (4) ρ_{mk} is the correlation coefficient between project m and k. (5) σ_j is the standard deviation of cost overrun/underrun of project j. Again, since it was assumed that all k projects are coming from a unique normal population with the standard deviation of $\bar{\sigma}$, all σ_j $(i = 1, ..., k)$ in the matrix $\mathbf{V}$ are equal to $\bar{\sigma}$. Since the standard deviation of the population is not known, $\bar{\sigma}$ is assumed to be the standard deviation of cost overruns/underruns of k observed projects σ. (6) $|\mathbf{V}^{-1}|$

is the determinant of matrix $\mathbf{V}^{-1}$ which is the inverse of matrix $\mathbf{V}$. One should note that Eq. (3) is the joint probability function of project cost overruns. In other words, assuming that $\bar{\boldsymbol{\delta}}$ and $\mathbf{V}$ are known, the probability of observing $\boldsymbol{\delta} = (\delta_1, ..., \delta_k)^T$ is found. With Bayesian updating, to find the likelihood of having $\bar{\boldsymbol{\delta}} = (\bar{\delta}, ..., \bar{\delta})^T$ as the parameter of distribution is needed when $\boldsymbol{\delta} = (\delta_1, ..., \delta_k)^T$ is observed. Thus, Eq. (3) can obtain the likelihood function of $\bar{\boldsymbol{\delta}}$ which is the function of just $\bar{\delta}$ as follows (Rowe 2003):

$$L(\bar{\delta}_1, ..., \bar{\delta}_k) = L(\bar{\delta}, ..., \bar{\delta}) = \frac{\left|\mathbf{V}^{-1}\right|^{1/2}}{(2\pi)^{n/2}} . \exp\left[-\frac{1}{2} . (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}})' . \mathbf{V}^{-1} . (\boldsymbol{\delta} - \bar{\boldsymbol{\delta}}) \right] \quad (4)$$

Having the likelihood function and assuming a prior distribution $f'(\bar{\delta})$ like Eq. (5), the posterior distribution $f''(\bar{\delta})$ is a normal shape given as Eq. (6).

$$f'(\bar{\delta}) \sim N(\delta', \sigma') = \frac{1}{\sqrt{2\pi} . \sigma'} \exp\left[-\frac{1}{2} . \left(\frac{\bar{\delta} - \delta'}{\sigma'} \right)^2 \right] \quad \& \quad f''(\bar{\delta}) = k . L(\bar{\delta}_1, ..., \bar{\delta}_k) . f'(\bar{\delta}) \sim N(\delta'', \sigma'') \quad (5 \ \& \ 6)$$

In the correlated case, to calculate mean and standard deviation (δ'', σ'') of $f''(\bar{\delta})$, only numerical approach is available and there is no close form formula. To this end, a range of possible $\bar{\delta}$ is selected. To be conservative, the range is assumed to be from -99.99% to 200% with the pace of 0.001. These values are input in Eqs. (4) and (5) to respectively calculate likelihood value, $L(\bar{\delta}_1, ..., \bar{\delta}_k)$ of each possible $\bar{\delta}$ and the correspondence prior PDF value, $f'(\bar{\delta})$. It should be noted that for the values of $\bar{\delta}$ outside the range of [-99.99%, 200%], the $L(\bar{\delta})$ and accordingly $f''(\bar{\delta})$ become too small so that they can be ignored from the analysis without any significant impact. Reviewing Eq. (4), it is found the term in the exponential function is the product of three $(1 \times k)$ and $(k \times k)$ and $(k \times 1)$ matrices which results in a polynomial function of $\bar{\delta}$. It means that Eq. (4), for any $\bar{\delta}$, gives a scalar likelihood value. Having variance-covariance matrix, $L(\bar{\delta}_1, ..., \bar{\delta}_k)$ can be easily calculated by any available mathematical package such as MATLAB. Multiplication of the prior PDF and likelihood values of each $\bar{\delta}$ gives $f''(\bar{\delta})$ which is the posterior PDF value of $\bar{\delta}$ before normalization (the area under the curve is not equal one). Both curves $L(\bar{\delta}_1, ..., \bar{\delta}_k)$ *vs.* $\bar{\delta}$ and $f''(\bar{\delta})$ *vs.* $\bar{\delta}$ need to be normalized. To calculate δ'' and σ'', the area under the curve of $f''(\bar{\delta})$ *vs.* $\bar{\delta}$ after normalizing is assumed to be divided to t narrow rectangles. The area of each rectangle is the probability of having the $\bar{\delta}_{(mid)j}$ (midpoint of the rectangle j). Then:

$$\delta'' = E(\bar{\delta}) = (\bar{\delta} : f''(\bar{\delta}) \ is \ Maximum) \quad (7)$$

$$\sigma''^2 = E(\bar{\delta}^2) - (E(\bar{\delta}))^2 = \sum_{j=1}^{t} \bar{\delta}^2_{(mid)j} . P(\bar{\delta}_{(mid)j}) - \delta''^2 = \frac{1}{\sum_{j=1}^{t} \left[\bar{\delta}_{(mid)j} . f''\left(\bar{\delta}_{(mid)j}\right) \right]} \left[\sum_{j=1}^{t} \bar{\delta}^2_{(mid)j} . \left(\bar{\delta}_{(mid)j} - \bar{\delta}_{(mid)j-1} \right) f''\left(\bar{\delta}_{(mid)j}\right) \right] - \delta''^2 \quad (8)$$

The term $\sum_{j=1}^{t}\left[\bar{\delta}_{(mid)j}.f''\left(\bar{\delta}_{(mid)j}\right)\right]$ in Eq. (8) is to normalize the posterior distribution and plays the role of k in Eq. (7).

4 UPDATING THE PRIMARY PARAMETERS OF THE MODEL

In Section 3, using Bayesian approach and having the information of newly completed projects the distribution of $\bar{\delta}$, the average of cost overruns/underruns, was updated and posterior distribution $f''(\bar{\delta}) \sim N(\delta'', \sigma'')$ was calculated. The mean δ'' and standard deviation σ'' of the posterior distribution is now used to update the primary parameters of the proposed model α, β, and ρ as follows:

$$\alpha_{new} = P(x < -m) = P(Z < \frac{-m-\delta''}{\sigma''}) = \Phi(\frac{-m-\delta''}{\sigma''}) \quad \& \quad \beta_{new} = 1+\delta'' \Rightarrow \rho_{new} = \frac{\beta_{new}}{1-m} \qquad (9 \ \& \ 10)$$

where Φ is the cumulative function for standard normal distribution. The proposed model is updated by α_{new}, β_{new}, and ρ_{new} values and becomes ready to be applied to any prospective set of projects which are in budget allocation process.

5 APPLICATION OF THE MODEL

The application of the model is demonstrated using cost data from 31 transit projects funded by the Federal Transit Administration (FTA) of the US Department of Transportation. These projects are divided into three groups: (1) a set of 22 project completed before 2004 named Historical Dataset for determining the model parameters of α, $\bar{\delta}$, β and ρ; (2) a set of 5 projects completed in 2004 and named First Dataset to see the effect of the model on cost overruns/ underruns and updating the model with actual costs of this projects; and (3) a set of 4 projects completed in 2005-06 and named Second Dataset to further investigate the effectiveness of the model and the Bayesian updating part. After verifying the normality of Historical Dataset using a test of goodness of fit, $m = 15\%$ was assumed as the maximum expected underrun defined by the FTA. Then from this dataset, it was calculated that $\alpha = 9.1\%$ and the average of cost underruns/overruns is $\bar{\delta} = 8.79\%$; thus $\beta = 1.0879$ and $\rho = 1.2799$. To determine the correlation among each pair of the projects in the First and Second Datasets, an approach called the Proposed Structured Guideline (PSG) was used to elicit correlation (Bakhshi 2011). The results of applying the proposed approach on aforementioned datasets are summarized in Table 1. Column "Actual Cost Overrun/ Underrun" depicts the actual mean and standard deviation of cost overruns/ underruns in three datasets. Column "Adjusted Cost Overrun/ Underrun" shows the mean and standard deviation of cost overrun/ underrun if the model had been applied to the data. The last Column "Updated Cost Overrun/Underrun" presents the mean and standard deviation of cost overruns/underruns after using the Bayesian updating which will prepare the model for the next application. The required adjustment in the value of factor B^*/B and cost overrun/underrun are diminished after each updating. For example, the First Dataset could have ended up with a 6.00% cost underrun instead of the actual 13.84% cost overrun by assigning increasing factor of 1.2111 to the budget and individual risk

assessment confidence level of 77%. For the Second Dataset, using the model could have reduced the cost overrun from 21.04% to 7.73% by increasing the budget by a factor of 1.0832 and individual risk assessment confidence level of 77%. Table 1 shows the improvement that can be gained by applying this model over a period of time. Due to the inherent characteristic of Bayesian updating, it is expected that the model gets more accurate as more projects become available and more updating occurs.

Table 1. Summary of the results from applying the proposed model on transit projects.

Data	α	β	For γ = 85%		Actual Cost Overrun/ Underrun		Adjusted Cost Overrun/ Underrun		Updated Cost Overrun/ Underrun	
			B^*/B	η	$\bar{\delta}$	$\bar{\sigma}$	$\bar{\delta}_{Adj.}$	$\bar{\sigma}_{Adj.}$	δ''	σ''
Historical Dataset	N/A	N/A	N/A	N/A	8.79%	0.2053	N/A	N/A	N/A	N/A
First Dataset	9.10%	1.0879	1.2111	77%	13.84%	0.1844	-6.00%	0.1522	1.30%	0.0965
Second Dataset	4.55%	1.0130	1.0832	77%	21.04%	0.4118	-7.73%	0.3139	1.10%	0.0888
Prospective Dataset	3.50%	1.0110								

6 CONCLUSION

To control the cost overrun, in this paper, a new probabilistic model was introduced for allocating contingency in a portfolio of correlated construction projects. The model assumes hybrid normal distribution for cost of projects and utilizes available historical data. Then, a Bayesian approach is employed to update the model as more projects are completed and new information becomes available. The proposed model first helps an agency to find the required portfolio's budget increase in order to have a certain confidence γ that the budget will be sufficient. Also, the model gives the required confidence level η to conduct risk assessment at individual project level to insure that the portfolio budget will not overrun with a probability of more than $1-\gamma$. The model can be updated with the information of newly completed projects where the dependencies between the projects are acknowledged and incorporated into the model.

References

Bakhshi, P. and Touran, A., A New Approach for Contingency Determination in a Portfolio of Construction Projects, *J.of Risk Analysis and Crisis Response*, 2 (4), 223-232, Dec, 2012.

Bakhshi, P., A Bayesian Model for Controlling Cost Overrun in a Portfolio of Construction Projects, PhD Dissertation, Northeastern University, Boston, March, 2011.

Kutner, M., Nachtsheim, C., Neter, J., and Li, W., *Applied Linear Statistical Models*, 5th Ed., McGraw-Hill, New York, NY, 2005.

Rowe, D., *Multivariate Bayesian Statistics: Models for Source Separation and Signal Unimixing*, Chapman & Hall/CRC Press, Boca Raton, FL, 2003.