行政院國家科學委員會專題研究計畫　成果報告

隨機性成本估價之風險評估模式

計畫類別：個別型計畫
計畫編號：NSC94-2211-E-032-020-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：淡江大學土木工程學系

計畫主持人：楊亦東

報告類型：精簡報告
報告附件：出席國際會議研究心得報告及發表論文
處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 16 日

行政院國家科學委員會專題研究計畫成果報告
隨機性成本估價之風險評估模式

Stochastic estimation model for uncertain cost elements

計　畫　編　號：NSC-94-2211-E-032-020
執行期限：94 年 8 月 1 日至 95 年 7 月 31 日
主持人：楊亦東　執行機構及單位名稱： 淡江大學土木工程系

一、摘要

本研究提出以蒙地卡羅方法從事不確定性營建專案估價之模式。研究方法為應用多變量隨機亂數產生法中之高斯關連結構來處理具相關特性之成本項目。為求突破其他模式之限制，本研究提出之模式將處理：(1)成本項目具有不同的統計分佈型式(有些為連續分佈、有些為間斷分佈；有些為對稱分佈、有些為偏態分佈)；(2)相關特性可以用傳統線性或等級相關係數加以描述；(3)成本項目具有複雜的相關情況；(4)自動化近似非半正定性相關值矩陣以滿足數學理論要求。模式流程為：首先近似相關值矩陣；再就可行之相關性調整具高斯分佈之隨機亂數；依據各式分佈之反函數將高斯隨機亂數轉換為各式成本項目之估計值；最後就所有估計值進行不確定性營建專案估價。本研究提出之模式已應用於實務案例，並就結果之統計相關性與原先設定加以比較驗證。模擬結果實證指出成本項目之相關特性對專案成本造成之顯著影響。

關鍵詞：專案估價、風險評估、蒙地卡羅方法、高斯關連結構、統計分析

Abstract

This study proposes a Monte Carlo method to incorporate correlations between cost elements in the process of cost estimation. The method being considered is the Gaussian copula in the field of multivatiate random number generation. The uniqueness of the proposed model lies in the capability to treat the situations when (1) distributions of cost elements have various types and shapes; (2) correlations are described by either Pearson or Spearman coefficients; (3) cost elements have complex correlations; and (4) a negative semidefinite correlation matrix shall be adjusted to the closest feasible one. The proposed method first checks the feasibility of the correlation matrix, adjusts it by an eigenvalue correction method, then uses the correlations to generate correlated multivariate random vectors, which are employed to model possible outcomes of the cost elements. The method has been applied to a practical dataset to indicate that the impact of correlations is significant and may cause serious problems if neglected. The result is also used to validate that the proposed method can capture the correlations with relatively small deviations.

Keywords: Cost estimation, Risk assessment, Monte Carlo method, Gaussian copula, Statistic analysis

二、研究目的

Cost estimation begins in the early stages of construction projects and may repeat frequently during the entire life cycle. The reliability of cost estimation is important to ensure the success of the project since it serves as the foundation for making critical financial decisions.

In construction projects, the prices of all the resources (material, equipment, and labor) are exposed to certain levels of uncertainty (Russell and Ranasinghe 1992). To manage such inherent uncertainty, Monte Carlo simulation methods have been widely applied for various types of projects, such as (Touran 1993; Elkjaer 2000). In Monte Carlo simulation, a mathematical model is constructed based on pre-specified probability distributions, which describes the possible outcomes of major cost elements involved in a project, and run to see what the overall project cost will be for each simulation replication. After a certain number of replications, the collected samples are used to derive the output distribution of the overall project cost.

An important enhancement of ordinary simulation methods has been directed to consider statistical correlations (dependencies) between cost elements. The correlation represents the co-movement of two cost elements; when one is more expensive, the other tends to cost more as well (or cost less for a negative correlation). Arguments and evidences for the existence of correlations and their profound impact on simulation results have been presented in the literature (Diekmann 1983; Wall 1997). To treat the correlations, various approaches have been proposed, such as (Touran and Wiser 1992; Wang 2002).

The goal of this project is to develop a simulation-based method, which incorporates correlations between cost elements with more modeling capabilities. The present method fulfills the following requirements:

1. To allow the distributions (i.e., marginal distributions) of individual cost elements to be of

different types. Namely, some of them may only be expressed with discrete and finite options whereas others can be expressed as continuous functions. In addition, those continuous distributions may come from different families (e.g., some are lognormal while some are beta).

2. To provide an automatic procedure to check the feasibility (a mathematical definition will be given later) of a correlation matrix and adjust it if infeasible.

三、研究內容

The proposed method takes two sets of input: marginal distributions of the cost elements (measured in unit cost, for example £/m^2) and a correlation matrix between these elements. The method is composed of two stages, which will be explained below.

Setup stage

The proposed method starts with a check on the feasibility of the original correlation matrix. If it is already positive semi-definite, one can immediately begin the simulation steps described in the next section; otherwise, we adopt the eigenvalue correction method from Ghosh and Henderson 2003) to approximate the infeasible correlation matrix into a feasible one.

The setup stage consists of the following steps:

1. Decompose the correlation matrix M into a diagonal vector D of the eigenvalues and a full matrix V whose columns are the corresponding eigenvectors so that

$$MV=VD \tag{1}$$

2. Locate the negative eigenvalues and change them to a tiny positive number ε to yield a new diagonal vector $\overline{D}$.

3. Adjust the correlation matrix M by

$$M = V\overline{D}V^T \tag{2}$$

4. Take the diagonal elements of M and store their inverses as the diagonals in a full matrix

E:

$$\mathrm{E}_{ij} = \begin{cases} 1/\sqrt{\mathrm{M}_{ij}} & \forall\, i = j \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

5. Normalize the diagonal elements to ensure unit diagonals (fundamental requirement for a correlation matrix):

$$\overline{\mathrm{M}} = \mathrm{EME} \tag{4}$$

where $\overline{\mathrm{M}}$ is the new (approximated) correlation matrix.

A preliminary test is performed to ensure the effectiveness of the setup stage. We use a documented 19×19 correlation matrix (Chau 1995) as an example. This correlation matrix has been proved to be non-positive-semidefinite (Ranasinghe 2000). Our tasks here are to apply the proposed steps and to check if the differences between the approximated values and the original specifications are small enough. The differences are quantified in two metrics: L_{ave} (average) and L_{max} (maximum):

$$L_{ave} = \frac{\sum_{i>j} |\mathrm{M}^* - \mathrm{M}|}{\frac{n(n-1)}{2}} \tag{5}$$

where M is the specified correlation matrix and M^* is the approximated one; n is the number of cost elements.

$$L_{\max} = \max_{i>j} |\mathrm{M}^* - \mathrm{M}| \tag{6}$$

After performing the correction steps, L_{ave} is 4.8954×10^{-18} and L_{max} is 1.7208×10^{-15}. This shows empirically that the setup stage can adjust the infeasible correlation matrix to a feasible one with ignorable changes.

The setup stage is to treat possible infeasibility, which may result from either erroneous input or inconsistent estimation. In other words, the setup stage would not be of any good if the correlation coefficients are incorrect or inconsistent. Thus a careful review is critical to

ensure the correlation coefficients can reflect the true behavior of the correlation relationships.

Simulation stage

The fundamental concept of the simulation stage is to generate a vector of correlated normal variates, transform them into uniform variates by the aid of the cumulative normal probability function, and then map the variates into their individual marginal distributions by the inverse transform method. The generated random variates are used to model the cost elements with the desired correlation structure. The procedure described here incorporates ideas from a new correlated multivariate generation technique (Normal To Anything, NORTA) (Cario and Nelson 1997). In what follows, we enumerate all the steps and provide computational guidelines.

1. Apply the Cholesky decomposition to the correlation matrix so that $M=CC^T$ where C represents the Cholesky triangular.
2. Generate an IID (independent and identically distributed) unit scaled uniform random vector, $Y=(Y_1, Y_2, \ldots, Y_n)$ where *n* is the number of cost elements.
3. Translate Y into a standard-normal random vector $P=(P_1, P_2, \ldots, P_n)$.
4. Transform P into a correlated standard-normal random vector $Z=(Z_1, Z_2, \ldots, Z_n)$.
5. Compute $U_i = \Phi(Z_i)$ for $i = 1, 2, \ldots n$, where $\Phi(.)$ denotes the standard normal cumulative distribution function (CDF).
6. Compute $X_i = F_i^{-1}(U_i)$ for $i = 1, 2, \ldots n$, where $F_i^{-1}(U_i)$ represents the inverse of the *i*th marginal CDF.
7. Return X_i as the estimate for cost element *i*.
8. Compute the total project unit cost by summing up all the cost elements.
9. Repeat Steps 2 through 8 for each simulation replication, $j = 1, 2, \ldots, m$.
10. Return summary statistics on all simulation replications.

For Step 1, there exist several efficient algorithms to perform the Cholesky decomposition. The generation of a uniform random vector in Step 2 is a standard feature supported by almost all the popular computer languages (such as C++, Java, Visual Basic, FORTRAN). The transformation in Step 3 can be approximated by the following equation:

$$P_i=(Y_i^{0.135}-(1-Y_i)^{0.135})/0.1975 \quad (7)$$

The transformation in Step 4 is

$$Z_i = \sum_{j=1}^{i} c_{ij} P_j \quad \text{for} \quad c_{ij} \in \text{C} \quad (8)$$

Step 5 involves the following integral

$$U_i = \int_{-\infty}^{Z_i} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \quad (9)$$

, which can be approximated by

$$U_i = \begin{cases} (1/\sqrt{2\pi} \times \exp(-Z_i^{\,2}/2)) \times (b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5) & \text{if } Z_i < 0 \\ 1-(1/\sqrt{2\pi} \times \exp(-Z_i^{\,2}/2)) \times (b_1 t + b_2 t^2 + b_3 t^3 + b_4 t^4 + b_5 t^5) & \text{if } Z_i \geq 0 \end{cases}$$

where $t = 1/(1+0.2316419\times|Z_i|)$

$$b_1 = 0.319381530 \quad (10$$

$$b_2 = -0.356563782 \quad)$$

$$b_3 = 1.781477937$$

$$b_4 = -1.821255978$$

$$b_5 = 1.330274429$$

The marginal CDF in Step 6 can be of any type of distribution as long as their inverses can be calculated either directly or via approximation. This is why the proposed method is able to treat different kinds of distributions simultaneously. It is much easier when the marginal distribution has a closed-form inverse (such as uniform or triangular). Otherwise, one has to rely on numerical approximation algorithms to find the inverses of the commonly-used

distributions, such as beta, gamma, and normal distributions. Step 7 is self-explanatory.

By using X_i as the estimate of the *i*th cost element, Step 8 is simply

$$E_j = \sum_{i=1}^{n} X_i \quad (11)$$

where E_j is the *j*th observation of the unit cost of the project. To make the addition meaningful, the estimates of different cost elements should be converted into the same unit of measure, such as £/m^2.

四、結果與討論

The proposed method is applied to the British data set described in (Wall 1997) to demonstrate its practical use. The data set is drawn from 216 office buildings built between 1980 and 1994 and consists of 8 major cost elements. The dataset has been standardized based on the times and locations the buildings were built.

All the cost elements and their marginal distributions are shown in Table 1. The value of each cost element is expressed as £/m^2. Here a cost element represents a relatively large work package, which may consist of several tasks. For example, “superstructure” involves formwork, steelwork, and concrete pouring. This level of granularity is suitable for higher level estimation. Moreover, the measure of £/m^2 can be changed to reflect the usual unit for progress measurement, if the proposed method is applied to other construction projects. For instance, a reasonable measure of cost elements for a highway project may be £/m while that for a residential community project may be £/house.

In the example, we consider three families of distributions, i.e., lognormal, beta, and discrete. The lognormal distributions are used because they fit the data better as argued by Wall. The use of the other two is based on a pragmatic situation when a cost estimator prefers not using historical data but rather using a discrete distribution to describe possible outcomes of “fitting and furnishings”, and beta distributions (three points) to estimate the distributions

of “services” and “external works”. These arrangements have been justified in previous sections.

Table 2 shows the rank correlation coefficients between the cost elements of the full data set. Before applying the proposed method, the rank correlation coefficients are reviewed and adjusted to verify (1) if they can reflect the actual behavior of the correlations and (2) if they, derived from past data, are suitable for the current project. This process is based on practical judgments and can complement pure mathematic analysis. In this example, the rank correlation coefficients between “external works” and other cost elements are adjusted to be zero.

A simulation experiment is designed to implement the proposed method and to evaluate the impact of correlations between cost elements. In the experiment, every simulation replication leads to a sample of the project cost by simply summing up cost elements drawn from individual distribution. The output statistics can then be used to assess the behavior of the true project cost.

After 1000 simulation replications, Table 3 lists the descriptive statistics for the unit cost of the project. To assess the impact of correlations, we compare two scenarios: including and excluding correlations. Fig. 1 is a box-and-whisker plot which is used to visually compare the distributions of the two scenarios. The first observation is that both distributions are skewed to the right because the mean (shown as the cross) is larger than the median.

The second observation is that the scenario of “including correlations” has a much longer tail to the right than that of “excluding correlation”. This indicates the former has a larger variability (uncertainty) than the latter. This conclusion is unsurprising because the former has a much greater standard deviation that the latter (149.55 versus 108.92, a 37% difference). Consequently, the 95% confidence interval of the former is much wider than that of the latter.

Fig. 2 plots the CDFs of both scenarios. A practical use of the chart is to estimate the

unit cost of the project with a certain probability. Taking correlations into consideration, the unit cost with a 0.90 probability is 958.50 £/m^2, which would be profoundly underestimated as 903.52 £/m^2 if the correlations are neglected. The difference of 54.98 £/m^2 is greater than the cost of "substructure" (with a mean of 47.2 £/m^2 in Table 1). In other words, by neglecting the correlations, the error can be as serious as doing the substructure for free.

The proposed method is an approximation because of the following reasons. First, it is assumed that the correlation between X_i and X_j in Step 7 (denoted by M_X) is close to the correlation between Z_i and Z_j in Step 4 (denoted by M_Z). Theoretically, to find a proper M_Z that leads to the desired M_X requires solving n(n-1)/2 nonlinear equations but the computation can be cumbersome (Chen 2001). Second, Steps 3, 5, and 6 require numerical approximation.

Since the proposed method is an approximation, it is necessary to check the aggregated difference between the original specified correlation matrix and the generated one on the aforementioned metrics: L_{ave} and L_{max}. For this particular application, L_{ave} is 0.018 and L_{max} is 0.051. Moreover, the standard deviation of the differences is 0.015. Thus the confidence limits for the mean of the difference is estimated to be 0.018 ± 0.0059 at a two-tailed significance level of 0.05 with 27 degrees of freedom. The results provide us confidence that the proposed method, despite being an approximation, can model the desired rank correlations with relatively small deviations and thereby can help assess the true impact of correlations on cost estimation.

五、計畫成果自評

The proposed method is more general than previous approaches because (1) it can treat different types of marginal distributions (discrete or continuous, different families distributions) for cost elements in one framework and (2) it can automatically adjust an infeasible correlation matrix into a close and feasible one very efficiently. Details of the proposed method have been published in (Yang 2005).

The modeling capabilities of the proposed method are empirically validated by an application to a modified British data set consisted of 216 office buildings. With the modeling capabilities, the proposed method helps cost estimators assess the true impact of correlations between cost elements on the project unit cost. The impact has been shown significant and should be considered with caution.

六、參考文獻

Cario, M.C. and Nelson, B. L. (1997). “Modeling and generating random vectors with arbitrary marginal distributions and correlation matrix.” Tech. Rep., Department of Industrial Engineering and Management Sciences, Northwestern University.

Chau K. W. (1995). “Monte Carlo simulation of construction costs using subjective data.” Construction Management and Economics, 13, 369-383.

Chen, H. (2001). “Initialization for NORTA: generation of random vectors with specified marginals and correlations.” INFORMS Journal on Computing, 13, 312-331.

Diekmann, J. E. (1983). “Probabilistic estimating: mathematics and applications.” Journal of Construction Engineering and Management, ASCE, 109(3), 297-308.

Elkjaer, M. (2000). “Stochastic budget estimation.” International Journal of Project Management, 18(2), 139-147.

Ghosh, S. and Henderson S. G. (2003). “Behavior of the NORTA method for correlated random vector generation as the dimension increases.” ACM Transactions on Modeling and Computer Simulation, 13(3), 276–294.

Ranasinghe, M. (2000). “Impact of correlation and induced correlation on the estimation of project cost of buildings.” Construction Management and Economics, 18, 395-406.

Russell, A. D. and Ranasinghe, M. (1992). “Analytical approach for economic risk quantification of large engineering projects.” Construction Management and Economics, 10, 277-301.

Touran, A. (1993). “Probabilistic cost estimation with subjective correlations.” Journal of Construction Engineering and Management, ASCE, 119(1), 58-71.

Touran, A. and Wiser, E. P. (1992). “Monte Carlo technique with correlated random variables.” Journal of Construction Engineering and Management, ASCE, 118(2), 258-272.

Wall, D. M. (1997). “Distributions and correlations in Monte Carlo simulation.” Construction Management and Economics, 15, 241-258.

Wang, W. C. (2002). “Simulation-facilitated model for assessing cost correlations.” Journal of Computer-Aided Civil and Infrastructure Engineering, 17(5), 368-380.

Yang, I-Tung (2005). “Simulation-based estimation for correlated cost elements.” International Journal of Project Management, 23(4), 275-282.

七、圖表

Table 1. Descriptive estimates for cost elements (distributions and parameters)

Cost Elements	Descriptive Estimate (in £/m^2)
Substructure	Lognormal (47.2,30.9) [a]
Superstructure	Lognormal (263.6,82.4) [a]
Internal finishes	Lognormal (63.2,24.4) [a]
Fittings and furnishings	Discrete (7,0.2; 8,0.5; 9,0.2; 10,0.1)[b]
Services	Beta (150,180,220) [c]
External works	Beta (70,85,120) [c]
Preliminaries	Lognormal (76.4,47.3) [a]
Contingencies	Lognormal (21.2,13.2)[a]

[a] Lognormal (mean, standard deviation); the lognormal distributions are estimated by the historical approach based on 216 buildings.
[b] Discrete (outcome, probability); the discrete distribution is subjectively specified
[c] Beta (minimum, mode, maximum); the three parameters are subjectively specified

Table 2. Rank correlation coefficients between cost elements [a]

	Substructure	Superstructure	Internal finishes	Fittings and furnishings	Services	External works	Preliminaries	Contingencies
Substructure	1.00							
Superstructure	0.33	1.00						
Internal finishes	0.26	0.52	1.00					
Fittings and furnishings	0.10	0.26	0.28	1.00				
Services	0.28	0.57	0.64	0.33	1.00			
External works	0.00[b]	0.00[b]	0.00[b]	0.00[b]	0.00[b]	1.00		
Preliminaries	0.35	0.37	0.44	0.18	0.39	0.00[b]	1.00	
Contingencies	0.23	0.28	0.34	0.21	0.29	0.00[b]	0.36	1.00

[a] Correlations above 0.10 significant at 95% confidence

[b] Subjective correlations

Table 3. Statistics of two scenarios: including and excluding correlation (in £/m^2)

Statistics	Excluding correlations	Including correlations
Mean	759.21	756.88
Standard Deviation	108.92	149.55
Minimum	514.50	470.50
Q1 (25% percentile)	680.99	647.07
Q2 (Median)	759.21	756.88
Q3 (75% percentile)	823.71	843.35
Maximum	1147.20	1393.30
95% C.I. lower bound	590.30	522.00
95% C.I. upper bound	1024.00	1091.00
Estimate with 0.9 Probability	903.52	958.50

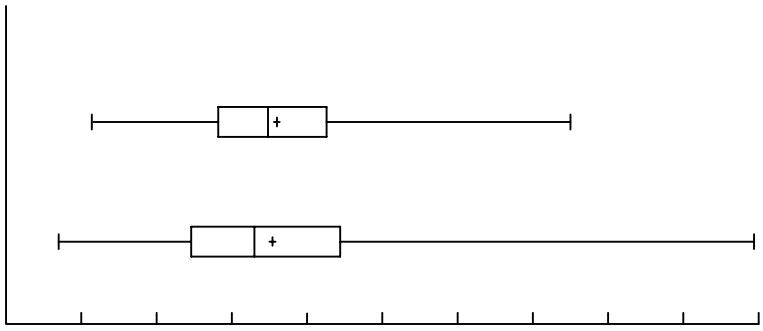

Fig. 1. Box-and-whisker plot for comparison between two scenarios

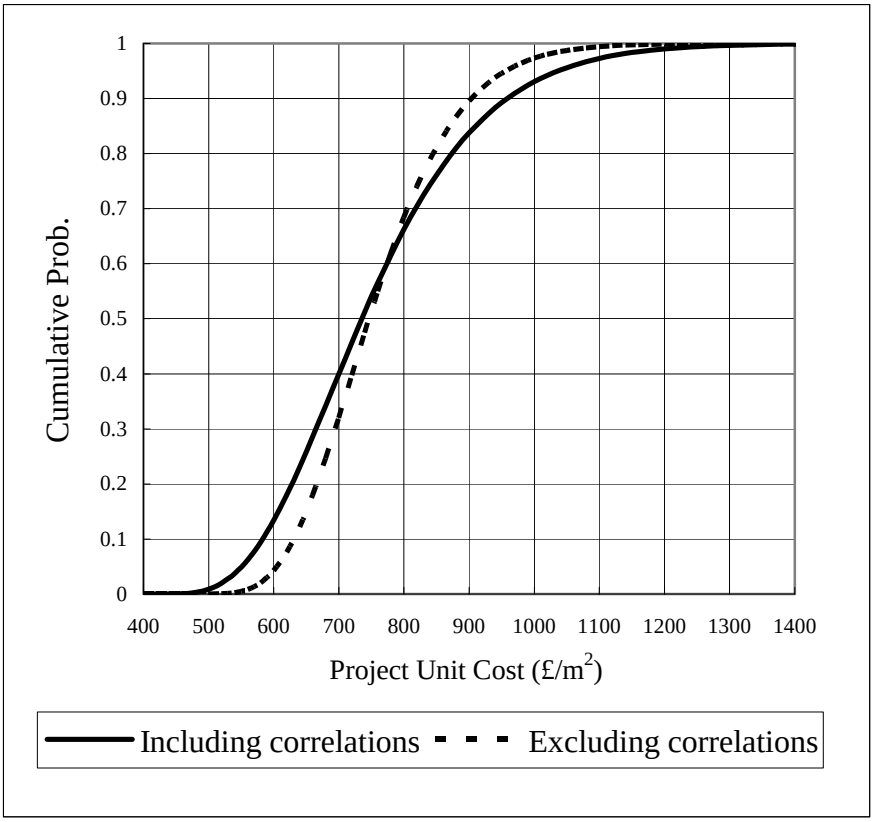

Fig. 2. Comparison on cumulative distribution functions of two scenarios