Developing Risk-Based Cost Contingency Estimation Model Based on the Influence of Cost Overrun Causes

Fahad Saud Allahaim,
Faculty of Engineering and IT, The University of Sydney
(email: fall5762@uni.sydney.edu.au)
Li Liu,
Faculty of Engineering and IT, The University of Sydney
(email: li.liu@sydney.edu.au)
Xiaoying Kong,
Faculty of Engineering and IT, University of Technology, Sydney
(email: xiaoying.kong@uts.edu.au)

Abstract

Cost overrun on infrastructure projects is widespread and represents significant financial risks to stakeholders. The large number of possible causes makes the planning and management of projects challenging. A survey of 160 project managers of infrastructure projects in Saudi Arabia was conducted to elicit the cost overrun causes. After cluster analysis, the causes were reduced to four dimensions: scope changes, market and regulatory, inadequate planning and control, and unforeseen circumstances. These four dimensions were then used to develop a risk-based cost contingency estimation model (RBCCEM) to improve the accuracy of cost forecasting and then validated using a bootstrapping approach. The accuracy of cost estimation measures was used to compare RBCCEM with fixed cost contingency (10%), reference class forecasting (RCF P50 & P90), and hybrid (it is a combination of RBCCEM & RCF P50). The comparison suggested that the RBCCEM could be more accurate as the error decreased by 10%. Therefore, by considering the actual impact of cost risk of similar projects, the results show that cost contingency was improved and the model delivered a better result compared to RCF.

Keywords: Cost overrun, cost overrun causes, infrastructure projects, classification and cost contingency estimation

1. Introduction

A significant proportion of large infrastructure projects have experienced substantial cost overrun which has led to financial or fiscal distress to project stakeholders and resulted in the deferral or cancellation of other projects (Flyvbjerg, 2014). Cost overruns in infrastructure projects are common around the world, as identified by Flyvbjerg et al (2003). Controlling project cost within budget is important for most if not all projects. The focus on cost performance is even stronger for infrastructure projects because of their high costs. Therefore, it is critical that causes of cost overrun are identified and effectively managed to minimize cost overrun.

Studies have identified a wide range of factors that lead to cost overruns, with two main schools of thought on the causes of cost overrun: technical and strategic causes. Technical causes include mistakes in design, overall price fluctuations, inaccurate estimations, government regulations, project size, quality of the contractor management team, plan changes, priority on construction deadlines, completeness and the project information timelines, the lack of experience of the estimators, certain bidding conditions, project characteristics, and lack of past data on similar types of projects (Koehn et l., 1978; Shash and Al-Khaldi, 1992; Lowe and Skitmore, 1994; Al-Harbi et al., 1994; Flyvbjerg et al., 2002; Memon, et al., 2011). The strategic causes considered optimism bias, which encapsulates the systematic propensity of decision makers to be over-optimistic about the outcomes of planned actions, as the main culprit of cost overruns for infrastructure projects (Flyvbjerg et al., 2002). However, the rhetoric seems to have shifted towards strategic misrepresentation as the main cause of cost overrun, which refers the use of deceptive means in order to win the project or obtain project funding (Liu and Zhu, 2007).

One of the techniques of reducing the impact of project cost overruns is the use of project cost contingencies—usually as a fixed proportion of the project total estimated cost and most recently estimated produced using sophisticated approaches such as reference class forecasting (RCF) or risk-based estimating (RBE) (Liu et al., 2010), but each method has its limitations. By taking into consideration the actual impact of cost risk of similar projects this paper develops and validates a cost contingency estimation model. A cross-sectional survey was conducted in Saudi Arabia to identify the causes of cost overrun of infrastructure projects in Saudi Arabia and the causes identified form the basis of the new cost contingency estimation model.

The structure of the paper is as follows: literature on causes of cost overrun was reviewed and the research design was explained. Cluster analysis is used to classify the causes of cost overrun into clusters. Subsequently, a cost contingency estimation model was developed by regressing project cost overruns on the clusters of causes. The model was then validated using the split sample. Further validation was conducted by comparing the accuracy of RBCCEM with those produced by the fixed cost contingency (10%), RCF (P50 & P90) and hybrid method, respectively. Finally, implications were discussed, future research directions were outlined and conclusions were drawn.

2. Literature review

This section examines the concept of the classification of cost overrun causes. Then, cost contingency estimation in infrastructure projects was also reviewed, followed by a discussion on using a classification approach in a cost contingency estimation model to improve cost contingency estimation accuracy.

2.1. Classification of causes of cost overrun

Cost overrun occurs in infrastructure projects (Memon, et al., 2011), and the causes are various. Classifying or grouping the large number of causes of overrun that may share similar patterns of impact can help manage causes during planning and construction.

Based on a survey of project managers on high-rise construction projects in two Indonesian cities, (Kaming et al., 1997) grouped seven causes of cost overruns into three groups using factor analysis: inflationary increases in material cost, inaccurate material estimating and project complexity. In Vietnam, Le-Hoai et al. (2008) categorized 21 causes of cost and time overrun for the construction industry using factor analysis and identified seven groups of causes: slowness and lack of constraint, incompetence, design, market and estimates, financial capability, government and worker factors. In Malaysia, Abdul Rahman et al. (2013) modelled 35 causes of cost overrun in large construction projects with a partial least squares-structural equation modelling approach and categorized the cost overrun conceptually in seven groups: contractor's site management related factors, design and documentation related factors, financial management, information and communication, human resource, non-human resources, project management and contract administration. These classification attempts have shown that homogenous groups of causes of cost overrun exist which aggregate the effect of causes within the same dimension.

Flyvbjerg has published various widely cited papers on causes of cost overrun for infrastructure projects. Flyvbjerg (2006) proposed a conceptual categorisation of cost overrun based on four main types of explanations that are claimed to account for cost overrun: technical, economical, psychological and political. Flyvbjerg et al. (2003) and Flyvbjerg (2008) acknowledged the technical explanations for cost overrun such as project size and location, but they concluded that the political-economic explanation of strategic misrepresentation and the psychological explanation of optimism bias are the main causes of cost overrun.

In brief, many causes significantly overlap, with relationships between multiple causes contributing to the final cause of cost overruns. There is a need to understand how the diversity of causes share similar patterns and how these causes impact on cost overrun, how causes can be mitigated, and the techniques or tools to ameliorate or eliminate cost overrun.

2.2. Estimation of infrastructure project cost

Accurate cost estimation tools can help reduce or eliminate the uncertainties of cost overrun. Accurate cost forecasting of large project costs is based on the availability and the level of professional knowledge and the historical cost data quality (Liu and Zhu, 2007). Available information, however, might be limited in the early stage of a large project. This may mean the quantity surveyor must make assumptions about the design; a detail of a project that may not eventuate as the life cycle of the project evolves (Liu et al., 2010).

The most critical feature of effective cost estimation is its potential for accuracy. Classic cost estimates consist of a base estimate, accounting for all physical quantities of materials and labour, and an additional risk contingency quantifying the underlying levels of uncertainty associated with the base estimate (Liu et al., 2010). Accuracy in forecasting costs and risks is valuable for decision makers to make rational decisions. Research has shown that cost forecasting errors are not unique to any specific industry or to project type with estimate inaccuracy in transport (Flyvbjerg et al., 2002), roads (Odeck, 2004), general construction (Liu and Zhu, 2007) and industrial projects (Merrow and Yarossi, 1990). Many studies have found, however, that there has not been noticeable improvement in estimation accuracy despite continued research (Flyvbjerg et al., 2002; Liu and Zhu, 2007).

The dominant methods of cost contingency estimation used in infrastructure projects can be classified into three categories: conventional contingency approach, risk-based estimation (RBE) and reference class forecasting (RCF) (Liu et al., 2010). The conventional contingency approach is to add a percentage, such as 10%, to the most likely estimate of the known works (Burger, 2003) based on the estimator’s experience, which may be prone to optimism biases and could lead to cost overrun (Yeo, 1990; Newton, 1992; Mok et al., 1997). The cost contingency technique is acceptable under stable conditions and simple projects, however, it is inappropriate for large and complex projects (Newton, 1992). As a result, it is a less evidence-based approach and a reason for many projects having cost overrun (Hartman, 2000).

The other two methods, RCF and RBE, have been shown to increase the accuracy of cost contingency estimation (Liu et al., 2010). The RBE model is the cost of individual components with base estimates and stochastic or random risk contingencies. Summing the stochastic cost components determines the distribution or probability of the overall project cost (Shaheen et al, 2007). The RBE method identifies inherent risks that directly relate to the internal behaviour of a project; as well as contingent risks derived from external events that may or may not occur (Aspinall and Trueman, 2006). It requires large amounts of expert time and expense (Liu et al., 2010) especially for large and complex projects.

RCF developed by Flyvbjerg, which only takes into account a project’s class (the outcome of cost overrun), even when other project factors might impact upon estimate accuracy. RCF utilises a database of previous project performance, from which a subsample of similar projects is selected, and adds a contingency to the total project cost (Liu et al., 2010). As RCF’s aim is to mitigate either optimism bias or strategic misrepresentation, it does not specifically address other causes of cost overrun such as technical causes and does not forecast events which may influence the project.

Since cost contingency accounts for the unforeseen cost risks, it is likely the estimating model based on the actual impact of cost risk of similar projects could produce more accurate cost contingency estimates. As a result, in the paper a risk based estimation method was adopted, by including ‘cost overrun causes classification scheme’, the accuracy of cost estimation could be improved. Supporting this proposal, (Liu et al., 2010) showed how RBE had excellent predictive validity as 90% projects having an actual cost within the range of the risk-based estimate in which they estimated the risk contingency of every single components of the project but they did include the cost overrun causes of similar projects. Therefore, this study designs and validates RBCCEM for infrastructure projects.

3. Research design

To develop the RBCCEM model, the survey data and the risk factors identified by the authors in a separate paper (Allahaim and Liu, 2015) is used to demonstrate development and validation process. First, the data collection and the identified causes of cost overrun for infrastructure projects in Saudi Arabia were summed up. Then, the use of cluster analysis to reduce the dimensionality of the risk factors was explained in preparation for the subsequent multiple regression analysis. The regression analysis derives the RBCCEM which was validated using bootstrapping analysis. Finally, the estimates produced by RBCCEM was compared those by alternative approaches such as the fixed cost contingency (10%), RCF (P50 & P90) and hybrid (RBCCEM & RCF P50) approaches.

4. Data collection

A survey of infrastructure project managers in Saudi Arabia was conducted to collect data from key infrastructure project professionals in three groups: owners exposed to project cost overrun, consultants supervising the projects, and contractors delivering the projects. The survey asked about the frequency of the 41 causes of cost overrun most frequently identified from 25 selected studies, as shown in Table 1. Respondents used five Likert-scale response anchors to assess the frequency of each cause in Saudi Arabia, based on their own professional experience. For more information about the data collection please refer to the survey data that conducted by the authors in a separate paper (Allahaim and Liu, 2015).

5. Data analysis and results

Based on the clusters identified, the scores for each cluster in each case was derived by aggregating the scores of each cause within each cluster. Subsequently, the cost overruns of projects were regressed on the four clusters identified to develop a risk-based cost contingency estimation model. R project software (version 3.0.2) and IBM SPSS 19 were used for statistical computing and graphics in the cluster analysis, model building and validation of the model.

5.1. Cluster analysis

Cluster analysis was used for dimension reduction (Everitt et al, 2011). The steps in cluster analysis of the data include preparing the data, determining the number of clusters, testing the cluster solution and finally validating clusters.

In Figure 1, there is an extreme “elbow” in the plot suggesting that solutions over four clusters do not have a substantial impact on the total SSE, which indicates that four clusters are appropriate. The next step tested hierarchical cluster analysis with the selected number of four clusters (Everitt et al, 2011). The Euclidean distance method was used to measure the dissimilarity distance based on the information values and the nature of the variables describing the objects to be clustered. Figure 2 shows the hierarchical cluster (tree) generated from R software, where the cause numbers (C1, C2... C41) refer to the causes listed in Table 1.

In the analysis, 10,000 bootstrap resamplings were used to reduce the error (Suzki and Shimodaira, 20016). In Figure 2, four rectangles have an AU *p*-value of 99 (0.99), therefore, for a cluster with AU p-value ≥95 (0.95), the hypothesis is rejected with significance level 0.01 for one cluster and 0.00 for three clusters, which indicates how strongly four clusters are appropriate as each cluster group contains objects which have a relationship with each other (Figure. 2).

5.1.1. Results – four cluster classification

The four cluster groups were defined based on causes of cost overrun and the literature as scope changes, market and regulatory uncertainty, inadequate planning and control and unforeseen circumstances. Table 1 shows how each of the 41 causes in Table 1 is allocated to one of the four clusters.

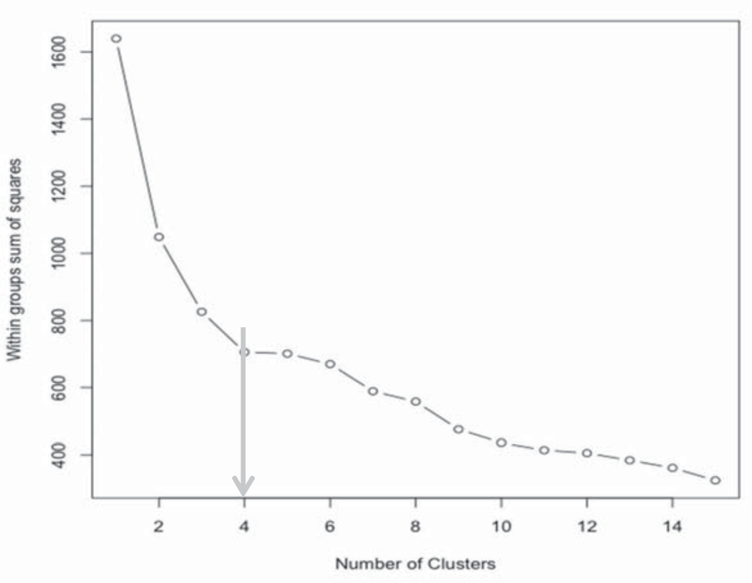
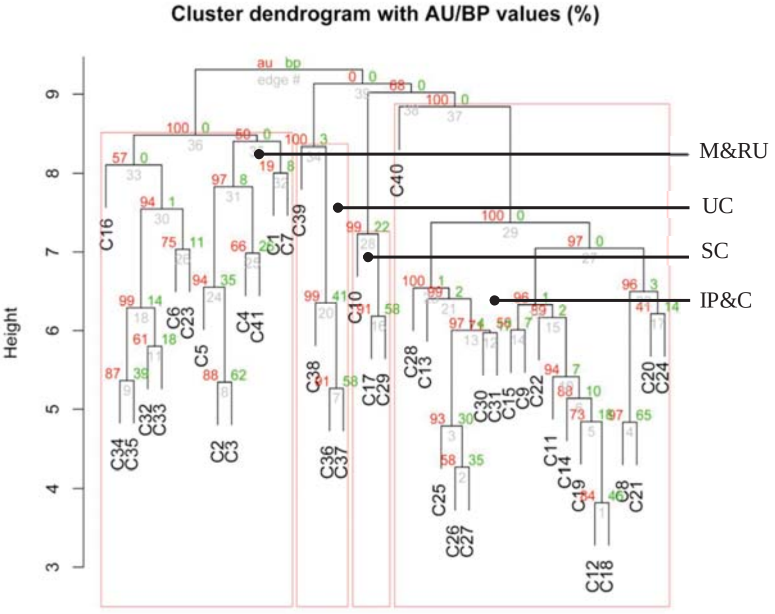


Figure 1: Elbow plot for the cluster determination



Note 1: Values at branches are AU p-values (left-red), BP values (right-green), and cluster labels (bottom). Clusters with AU ≥ 95 are indicated by the four red rectangles.
Note 2: Inadequate planning and control (IP&C), Market and regulatory uncertainty (M&RU), Scope changes (SC) and Unforeseen circumstance (UC).

Figure 2: Hierarchical clustering with four cluster solution

As shown in Table 1, the first cluster group is scope changes (SC), which represents the causes of cost overrun due to design changes, additional work and rework, and change in the scope of the project. The causes in this cluster are related to time, that is the urgency of the project, namely how much time there is to complete the job. Forcing the project team to take short-cuts or to work on tasks which clash with other tasks and working on concurrent tasks and projects are known to cause delays and cost overrun. The second cluster group is market and regulatory uncertainty (M&RU), which includes causes of cost overrun that relate to the chance or speculation changing costs, whether directly or indirectly. The third cluster group is inadequate planning and control (IP&C), it represents the causes of cost overrun which relate to project planning and control, which comprise the most critical causes of cost overrun in large projects in Saudi Arabia. Inadequate planning and control dimension is referring to the factors that could increase the complexity and thus difficulty of controlling the project cost. The last cluster group unforeseen circumstance (UC). The causes of this cluster relate to environment issues, as well as social and cultural impacts. These issues increase the pressure to find a solution to these problems associated with the project site. For example, the increase of environmental requirements has a significant impact on construction operations, which leads to technical uncertainty that relates to the physical difficulty of completing a project.

Table 1: Four-cluster classification scheme for causes of cost overrun

Classification clusters	Key	Causes of cost overrun	Relationship to cost overrun
Scope changes	C17	Design changes*	Unclear project scope forces project team to take short-cuts, crashing tasks, concurrent tasks/projects, which are known to cause delays and cost overrun (Shenhar and Dvir, 2007).
	C10	Additional work and rework*	
	C29	Change in the scope of the project*	
Market and regulatory uncertainty	C5	Market conditions (materials and labour)*	Increases the volatility of input costs and thus chances of overrun (Pindyck, 1993).
	C41	Practice of assigning the contract to the lowest bidder*	
	C4	Slow payment of completed works*	
	C3	Cash flow during construction	
	C33	Obstacles from government	
	C1	Inflation	
	C35	Laws and regulatory frameworks	
	C16	Failure to price in some risks	
	C2	Monthly payment difficulties from agencies (e.g. contractor, owner)	
	C34	Political complexities	
	C7	Deficiencies in cost estimates prepared by public agencies	
	C32	Fraudulent practices	
	C23	High interest rate charged by bankers on loans	
	C6	Fluctuation in money exchange rate	
Inadequate planning and control	C40	Delays (decision making, in approval of drawings, material delivery)*	Increases the complexity of coordination of parties and tasks, thus making it harder to meet present targets (Baccarini, 1996).
	C21	Design error*	
	C8	Deficiencies in the infrastructure*	
	C20	Changes in material specification and type*	
	C13	Shortage of site workers	
	C18	Incorrect planning and scheduling by contractors	
	C24	Inadequate specifications	
	C14	Unrealistic contract duration and requirements imposed	
	C11	Lack of experience of project manager (e.g. location, type)	
	C28	Lack of constructability	
	C15	Strategic misrepresentation	
	C22	Project size	
	C12	Contractor's poor site management and supervision skills	
	C19	Late delivery of materials and equipment	
	C25	Waste on site	
	C9	Labour, insurance, work security or workers' health problems	
	C27	Poor financial control on site	
	C26	Equipment availability and failure	
	C31	Optimism bias	
	C30	Inadequate modern equipment (technology)	
unforeseen circumstances	C37	Site constraints	Increases the uncertainty of tasks and
	C36	Weather conditions	

	C38	Social and culture impact (e.g. problems with neighbours)	outcome, thus making planning and estimating difficult (Ofori, 1992).
	C39	Heritage material discovery	

Note: (*) ranked in the top ten causes

5.2. Development of the risk-based cost contingency estimation model

To build the model using multiple linear regressions (MLR), as there are 160 cases and 41 causes, the data was randomly split into two data sets with two-thirds of the data as the training set (100 cases=62.5%) for model building and the remaining one-third (100 cases=37.5%) as the test set to ensure more reliable results and also to reduce bias in the validation (Kothari, 1985). For model validation, bootstrap resampling of multiple linear regressions was employed. Then, the estimates by RBCCEM were compared with those produced by alternative models such as fixed contingency and RCF. Error indices such as mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE) and root mean square error (RMSE) were used to compare the accuracy of estimates produced (Han and Kamber 2006). Then, RBCCEM, fixed cost contingency (10%), RCF (P50 & P90) and hybrid (is a combination of two models which used to increase the accuracy of cost contingency estimation as Liu et al., (2010) identified) were compared based on the distribution of the means of adjusted cost overrun as the smallest dispersion with the shortest distance between the mode and median indicates the most accurate model (Rothwell, 2005; Lawrence, 2007).

5.2.1. Risk-based cost contingency estimation model building

The regression results using the training subsample are reported in Table 2. The results in Table 2 show all four clusters have significant impact on cost overrun. It is worth noting that Table 2 shows that the R-squared is 32% and adjusted R is 30%, which indicates that the four clusters explains about 30% of variance in cost overruns (Chambers, 1992). The interpretation here is that the observed variation in cost overrun is also explained by other factors beyond those captured in the equation. It is not the intention of this paper to delve into which other factors explain overrun, as it focused on the four clusters (*p*-value < 0.05) as all variations that were categorised based on 41 causes that are frequently identified in the literature are significant. Therefore, the RBCCEM is represented by the equation 1.

$$\boldsymbol{BCCEM = \%Cost\ overrun = 0.652 + 0.157(scope\ changes) + 0\ .089(market\ \&\ regulatory\ uncertainty) + 0.047(inadequate\ planning\ \&\ control) + \ 0.024(unforseen\ circumstnaces)}$$

Equation 1

Note: the value of inadequate planning and control, market and regulatory uncertainty, scope changes and unforeseen circumstances ranges from 1-5 as 1 has low risk and 5 has major risk.

Table 2: Residuals, coefficients and p-values of regression analysis

Residuals	**Min**	**1Q**	**Median**	**3Q**	**Max**
	-2.109362147	- 0.515626813	0.00002042	0.47228	1.52539
Coefficients		**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
	(Intercept)	0.65196	0.08967	2.35254	0.00027***
	Scope changes	0.15733	0.10251	0.46178	0.00452**
	Market and regulatory uncertainty	0.08865	0.11366	0.77998	0.03734*
	Inadequate planning and control	0.04728	0.06818	0.54674	0.05218
	Unforeseen circumstances	0.02431	0.11406	2.13199	0.03558*
Residual standard error: 0.0159 on 95 degrees of freedom Multiple R-squared: 0.3249 Adjusted R-squared: 0.3048 F-statistic: 9.463 on 4 and 95 DF					

Note: $p < 0$ '***', $p < 0.001$ '**', $p < 0.01$ '*', $p < 0.05$ '.'

5.2.2. Models validation

The bootstrapping method was used for validating the RBCCEM. 5,000 bootstrap samples were created to validate the predictive ability of the proposed RBCCEM (multi-linear regression (MLR) model). According to all three measures (Table 3), the coefficient was statistically significant (the default for boot.ci is a 95% confidence interval) as they are centred to the normal, which indicated the model was valid. As reported in Table 3, the mean values of the four regression coefficients estimates from the RBCCEM bootstrapping were close to the proposed RBCCEM (Tables 2 and 3). Also, the standard error values of the four-parameter estimates from the RBCCEM bootstrapping were close to the proposed RBCCEM (Tables 2 and 3). The similarity of estimates of the RBCCEM model from both split samples suggests that RBCCEM is valid and robust.

Table 3: RBCCEM bootstrapping

Ordinary nonparametric bootstrap		**Estimate**	**bias**	**std. error**
	(Intercept)	0.57073	- 0.01413	0.092570
	Scope changes	0.14501	0.00348	0.128087
	Market and regulatory uncertainty	0.07840	- 0.00275	0.101091
	Inadequate planning and control	0.03820	- 0.00481	0.083566
	Unforeseen circumstances	0.03298	0.00148	0.100938
Bootstrap Confidence interval calculations	**Level**	**Normal**	**Percentile**	**BCa**
	95%	(0.2662, 0.9035)	(0.2243, 0.8637)	(0.2453, 0.8803)

Note: Calculations and Intervals on Original Scale

5.2.3. Models evaluation using estimation accuracy measures

In the models evaluation we used two methods. The first method was by using estimation error (error indices). The second method was comparing the adjusted cost overrun percentage means by using independent samples t-test. The following section will discuss these two methods and the results are delivered.

5.2.3.1. Model evaluation using measures of forecast accuracy (error indices)

To further validate the model, the estimates produced by RBCCEM were compared with alternative methods such as RCF. RCF uses a database of actual performance of comparable past projects within a given reference class to provide an objective reference point for the cost forecast of a current project (Flyvbjerg, 2006). For a particular project, reference class forecasting requires the following three steps (Flyvbjerg, 2006, p. 8): (a) identifying a relevant reference class of past projects as the base, (b) establishing a probability distribution for the selected reference class and (c) comparing the specific project with the reference class distribution.

To compare the models, the accuracy of each model was measured. Forecast accuracy measurements were based on the distributions of absolute errors (|E|) or squared errors (E^2), taken over the number of observations (n), which are the most commonly used measures to compare the performance of predictive models (Hyndman and Koehler, 2006). These include mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE) and root mean square error (RMSE) (Swanson et al., 2011). Table 4 presents MAPE MAE, MSE and RMSE where values of 0 indicate a perfect fit (Singh et al., 2013). Table 4 shows that the RBCCEM has comparable error indices to that of RBCCEM bootstrapping. In contrast, the error indices for RCF are much higher, suggesting RBCCEM is more accurate.

Table 4: The MAE, MAPE, MSE, and MAPE of MLR and error estimates of the four-clusters model: Proposed RBCCEM, RBCCEM bootstrapping and RCF model

	MAE	MAPE	MSE	RMSE	RMSE -MAE
Proposed RBCCEM	0.473786	15.78 %	0.426456	0.653036	0.18
RBCCEM bootstrapping (5,000 bootstrap of test data)	0.428924	14.68 %	0.358026	0.598353	0.17
RCF model	0.872376	25.19 %	1.190669	1.091178	0.22

5.2.3.2. Models evaluation by comparing the means of adjusted cost overrun percentage of the models

As discussed, β1 to β4 were estimated based on a sample of 100 and the model was validated using bootstrapping based on the 60 samples, which shows the model was valid and accuracy was improved compared with RCF. In this section, the mean of the adjusted cost overrun percentage of the RBCCEM was compared with those estimated using alternative approaches, such as RCF and fixed contingency (10%), using the split sample. The result in Tables 5 and 6 shows the adjusted cost over. The comparison results reported in Tables 5 and 6 showed that the adjusted cost overrun of RBCCEM results are significantly lower than adjusted cost overrun of fixed (10%) cost contingency, RCF and hybrid approach.

Table 6 shows that the means of adjusted cost overrun percentage using the fixed cost contingency, RBCCEM, RCF P50, RCF P90 and hybrid (RBCCEM + RCF P50), respectively, are all significantly different from each other (*p*-value <0.00). Considering the negative mean (under budget) in the adjusted cost overrun percentage mean of RBCCEM (-0.11451), the results suggested that the mean of adjusted cost overrun percentage using the RBCCEM approach is preferable to that using the RCF P50, RCF P90 and hybrid.

Moreover, Table 6 reports that the adjusted cost overrun using RCF reduce the overrun significantly (*p*-value <0.05) as the mean differences for P50 is (0.1049) and for P90 is (0.0140) (Table 5 and 6). Despite the fact that RCF P50 and RCF P90 have lower mean differences, it should be noted that the RCF estimates are subject to the acceptable risk of cost overrun which RBCCEM does not. In addition, the RBCCEM model tends to underrun budget while the RCF model tends to overrun budget. The dispersions of the RCFs are higher than RBCCEM (Figure 3) suggesting RBCCEM produces more consistent results. In sum, using estimates of contingencies by RBCCEM results in slight average cost underruns with more consistent and accurate estimates than that from RCF.

In addition, the variance of adjusted cost overrun percentage using the fixed cost contingency, RBCCEM, RCF P50, RCF P90 and hybrid, respectively, are significantly different (*p*-value <0.035) (Table 6). Considering that the variance of adjusted cost overrun percentage using the RBCCEM is lower than those using fixed cost contingency, RCF P50, RCF P90 and hybrid (Table 6), respectively, the results indicates RBCCEM produces the most consistent estimates for cost contingency Therefore, RBCCEM is the preferable method for estimating cost contingency for infrastructure projects.

Table 5: Descriptive statistics for mean of adjusted cost overrun of two models

	N	**Mean**	**Std. Deviation**	**Std. Error Mean**
Fixed cost contingency 10%	60	**0.3319**	0.22739	0.02935
RBCCEM	60	**– 0.1145^**	0.15616	0.02016
RCF uplift P50	60	**0.1049**	0.17155	0.02214
RCF uplift P90	60	**0.0140**	0.15604	0.02014
Hybrid of RBCCEM and RCF uplift P50	60	**–0.0096^**	0.32646	0.04215

Note: ^ A negative value denotes under and a positive figure indicates over budget.

Table 6: Test results for the equality of means and variances

	Leven's Test for Equality Variance		**t-test for Equality of Means**				
	f	**sig**	**t**	**df**	**Sig. (2-tailed)**	**Mean Difference**	**Std. Error Difference**
Fixed cost contingency 10% *vs.* RBCCEM	3.289	0.035	6.54	104.52	0.000***	0.44650	0.13561
Fixed cost contingency 10% *vs.* RCF uplift P50	7. 659	0.005	9.174	109.73	0.000***	0.23677	0.03117
Fixed cost contingency 10% *vs.* RCF uplift P90	7.394	0.006	8.931	104.48	0.000***	0.31795	0.04360
Fixed cost contingency 10% *vs.* Hybrid of RBCCEM and RCF uplift P95	6.616	0.010	6.650	105.342	0.000***	0.34157	0.05136

Note: Significance codes: $p < 0$ '***', $p < 0.001$ '**', $p < 0.01$ '*', $p < 0.05$ '.'

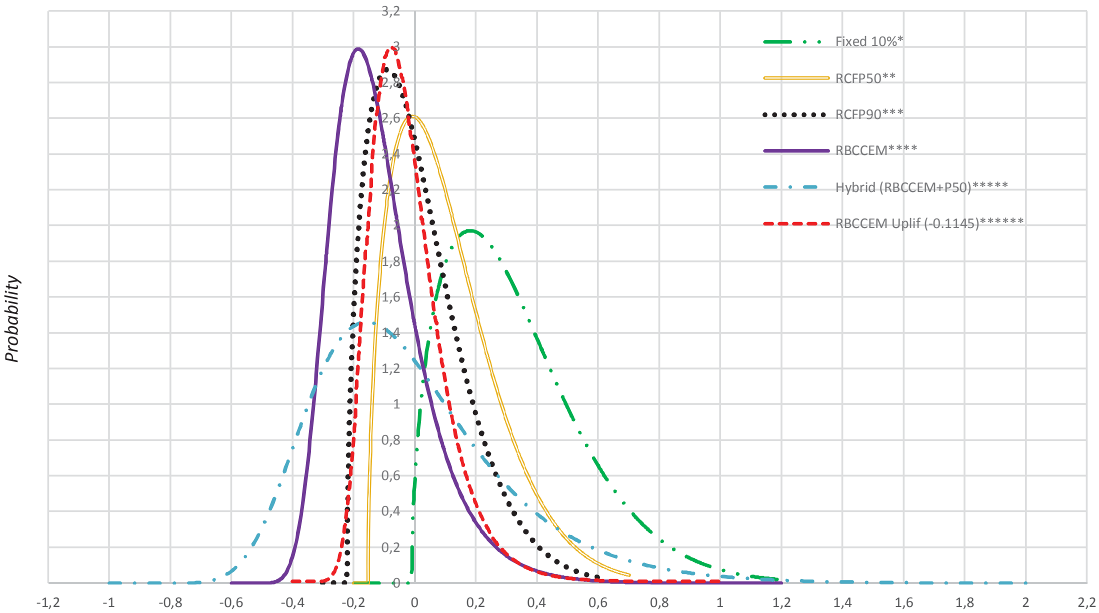
Figure 3 shows the mean distributions fitting to the data. As can be observed that all distributions are skewed to the left and the tails of the distributions are longer in the right. Therefore, the mode is the peak of each distribution, and the median and mean come after it in the right. Furthermore, the distribution of adjusted cost overrun percentage for RCF P50 and P90 are fitted to Weibull distribution as can be observed in the Figure 3 by the orange doubled line and black dotted line, respectively.

In addition, the fixed 10% cost overrun (green long dashed dot dot line) is fitted to Weibull distribution and hybrid (RBCCEM + RCF P50) (blue long dashed dot line) is fitted to exact value distribution. The adjusted cost overrun of RBCCEM distribution (purple solid line) is fitted to exact value distribution which is biased toward under-budget (mean adjusted cost overrun=-0.1145). Comparing the dispersions of the distribution curves presented in Figure 3, RBCCEM has the narrowest dispersion, supporting the above conclusion that RBCCEM produces the most consistent estimates for cost contingency of infrastructure projects.

Further, the small negative mean of adjusted cost overrun of using RBCCEM can be offset by adding an amount equal to 0.1145As a result, the distribution of RBCCEM +0.1145 (red dashed line) shifted around zero and has the narrowest dispersion compared to the other distributions.

6. Conclusions

Based on a cross-section survey of managers involved in infrastructure projects in Saudi Arabia, cluster analysis was used to reduce 41 causes of cost overruns to four clusters; scope changes, market and regulatory uncertainty, inadequate planning and control, and unforeseen circumstances. Using multiple - regression analysis, the risk-based cost contingency estimation model (RBCCEM) was developed by regressing project cost overrun on the four clusters. Then, validity of the RBCCEM was validated by multiple regression bootstrapping using the remaining split sample (sample size of 60) and by comparing the cost overrun outcomes of using cost contingency estimates from RBCCEM to those of using alternative estimating approaches such as RCF and fixed contingency. The validation analysis



*Note: * distribution fitting of fixed % cost contingency of cost overrun (Weibull distribution)*
*** distribution fitting of adjusted % cost overrun based RCF on 50% percentile (Weibull distribution)*
**** adjusted % cost overrun based RCF on 90% percentile (Weibull distribution)*
***** distribution fitting of adjusted % cost overrun based on RBCCEM (proposed model) (Exact value distribution)*
****** distribution fitting of Hybrid model (RBCCEM + RCF P50) (Exact value distribution)*
******* distribution fitting of adjusted % cost overrun based on RBCCEM (RBCCEM + uplift (-0.1145) (Exact value distribution)*

Figure 3: The fitting distribution curves of fixed cost contingency, RBCCEM, RCF (P50 & P90), hybrid and RBCCEM uplifted (RBCCEM-0.1145)

showed that the degree of dispersion of the cost overrun and the mean of the cost overrun after including the cost contigency is the lowest for RBCCEM, in which it is prefered method of estimation for cost contingency. However, the accuracy of cost contingency could be improved further by offsetting the negative mean of cost overrun using hybrid approach, i.e. by deducting the mean from the cost contingency produced by RBCCEM. Such an adjustment uplifts the means of cost overrun to zero while the degree of dispersion remains unchanged.

To apply RBCCEM, an organization needs to ascertain a comprehensive list of the risks to the cost overrun of similar projects through using questionnaire survey. The questionnaire should be based on the questionnaire used in this study and tailored to the project at hand. Assuming the survey responses of at least 30-40, then the organization can proceed to categorizing the risks by conducting clustering analysis. Subsequently, construct scores can be derived by aggregating the scores of individual risks within each category. Finally, regression analysis of cost performance on risk categories is conduced to drive the cost contingency model which will be used to predict cost contingency for the project. Bootstrapping using a holdout sample is a useful validation of the cost contingency estimation model.

The findings are based on a cross-sectional survey of managers involved in infrastructure projects in Saudi Arabia. Therefore, caution should be exercised when generalizing to other contexts. Future