Basic Technical Paper Template
(Rev. 2010-04-05)

(William) Craig Boudreau
66421

Using Historic Data to Improve Monte Carlo Prediction of Project Outcomes

May 17, 2016

Table of Contents

List of Tables

List of Figures

List of Equations

ABSTRACT

This paper presents a model using historic data from past projects to predict the final costs for similar projects through the use of Monte Carlo simulation. This model uses job-to-date measurements for ten (10) key performance indicators (KPIs) combined with known historic progression of these KPI's to estimate the probable range of project cost at completion. Three different Monte Carlo models are developed, which vary based on inclusion or exclusion of input documents such as construction schedules and indirect staffing plans. To understand the historic progression of the ten (10) KPIs used in the Monte Carlo simulation, eleven (11) projects in the Western Canadian heavy industrial construction sector are examined.

This approach shows promise using a limited sample size of projects. Future work is recommended to increase the sample size used to derive the Monte Carlo model as well as the number of projects used to test the model.

INTRODUCTION

Contractors are undertaking more risk than ever due to increasingly complex projects and more onerous contract terms. This added risk makes accurate cost forecasting extremely important to effectively manage projects.

In order to sanity-check cost forecasts, some organizations implement Monte Carlo simulations. This approach involves specifying ranges for input variables and, through numerical simulations, generates a range of possible project outcomes. Understanding the probability of possible project outcomes can validate remaining contingency or spot problems with the project team's cost forecast.

This paper explores using historic data from past projects to predict the final costs for similar projects through the use of Monte Carlo simulation. A model is presented where job-to-date measurements for ten (10) key performance indicators (KPIs) are combined with known historic progression of these KPI's to estimate the probable range of project cost at completion.

The Monte Carlo process provides valuable feedback but can suffer from two weaknesses. The first weakness is it often relies upon subject matter experts (SMEs) to provide ranges for input variables rather than distributions based on past data [6]. These ranges are often based on professional judgement and subject to biases and potential under-estimation of outlying events. The second weakness of the Monte Carlo process is that it often assumes all modeled variables are independent and does not account for cross-correlations. For example, it assumes that the hourly rate for direct labor is independent from the hourly rate for indirect labor. In reality, these two factors would most likely be correlated due to market conditions.

In order to increase the accuracy of this Monte Carlos forecasting process, this paper uses *performance of past projects* to determine the probability distributions of Monte Carlo input variables used in the models. This paper also explores the most sensitive inputs to this Monte Carlo simulation and what other forecasting technics are available to help improve forecasting of the most sensitive variables. Three (3) model types are presented to address the most sensitive

variables. The differences between the models are based on using either history or external documents (i.e. staffing plans and schedule) to calculate direct labor hours and indirect labor hours at completion and the convergence of individual KPIs to their at-completion values is examined to provide guidance on when each KPI's becomes useful in forecasting. Finally, benefits of this approach are discussed as they relate to assessing risk and contingency management

This approach is broadly similar to one proposed by Mohammad Libeig and Dr. Gholamreza Heravi [2] for using Monte Carlo analysis and historic data to forecast project performance. The main difference is this analysis focuses on estimation of final project costs as opposed to a wider assessment of project success (ignoring metrics for quality, cashflow and safety). This model also forecasts more granular financial KPIs which can then be individually cross-checked with line items in the project team's forecast.

METHOD

Step 1 - Model Development

First, ten (10) key performance indicators (KPI's) are identified that can be used to calculate the project's final cost. These KPI's are listed individually in Table 1 – KPI Definitions, including descriptions, mathematical formulae and unique equation numbers 1 through 10. From here forward each KPI will be denoted individually for example, Eq. 1 will be denoted as KPI 1.

KPI	Description	Formula	
KPI 1	Hourly rate for direct labor	Direct Labor Cost / Direct Labor Hours	Eq. 1
KPI 2	Hourly rate for indirect labor	Indirect Labor Cost / Indirect Labor Hours	Eq. 2
KPI 3	Ratio of indirect hours to direct hours	Indirect Labor Hours / Direct Labor Hours	Eq. 3
KPI 4	Hourly rate for equipment	Equipment Cost / Direct Labor Hours	Eq. 4
KPI 5	Hourly rate for subcontractors	Subcontractor Cost / Direct Labor Hours	Eq. 5
KPI 6	Hourly rate for material	Material Cost / Direct Labor Hours	Eq. 6
KPI 7	Hourly rate for small tools and consumables (ST&C)	ST&C Cost / Direct Labor Hours	Eq. 7
KPI 8	Hourly rate for travel and subsistence	Travel Cost / Direct Labor Hours	Eq. 8
KPI 9	Hourly cost for "other" costs	(Total Project Cost - Direct Labor Cost - Indirect Labor Cost - Equipment Cost - Material Cost - Subcontractor Cost - ST&C Cost - Travel Cost)/ Direct Labor Hours	Eq. 9
KPI 10	Direct Hours Expended	Direct Labor Hours	Eq. 10

Table 1 – KPI Definitions

Note that the inclusion of KPI 9 for "other" costs ensures the categories are mutually exclusive and exhaustive for all project costs. Also note that hourly metrics for equipment (KPI 4) and subcontracts (KPI 5) are measured per direct labor hour, not per equipment or subcontract hour. This is done out of convenience for how data was captured in the accounting system.

Next, using Equation 11, the *at-completion* value of these KPI's is combined to calculate the *at-completion* project cost. The subsequent discussion uses the terminology "at-completion" to

denote the final value of variables at project completion and "to-date" to denote the value of variables measured at some mid-point during project execution.

$$Project\ Cost_{at-Completion} = KPI\ 10_{at-completion} * \left[(KPI\ 3_{at-completion} * KPI\ 2_{at-completion} + KPI\ 1_{at-completion} + KPI\ 4_{at-completion} + KPI\ 5_{at-completion} + KPI\ 6_{at-completion} + KPI\ 7_{at-completion} + KPI\ 8_{at-completion} + KPI\ 9_{at-completion})\right] \quad \text{Eq 11}$$

A relationship is then developed to combine historic performance on past projects with performance to-date on a current project to forecast the at-completion KPI performance for the current project. In the formulae below, all variables labelled MC are treated as Monte Carlo input variables, where their distribution is determined based on historic data from past projects, which will be described below in *Step 2 - Analyzing Past Performance for use in Monte Carlo Calculations*.

So, Equation 11 allows the calculation of project estimate at completion (EAC) costs based on the knowledge of the KPI's *at-completion*. Equation 12 describes the relationship between each KPI's to-date value to that KPI's at-completion value.

$$KPI\ n_{At-Completion} = KPI\ n_{To-Date} * MC\ n \quad \text{Eq 12}$$

Where:

KPI n $_{At\text{-}Completion}$= *Value of n^{th} KPI at completion*

KPI n $_{To\text{-}Date}$ = *Value of n^{th} KPI measured on project to-date*

MC n = *Monte Carlo input variable representing known historic progress of n^{th} KPI from to-date value to at-completion value.*

Equation 11 is then modified by the general relationship described in Equation 12 to produce Equation 13 which allows the calculation of a project's at-completion costs based on the knowledge of KPI's *to-date* combined with knowledge of historic progression of these KPI's. This in turn allows the probabilistic modeling of project final costs based on these known to-date inputs.

$$Project\ Cost_{At-Completion} = \frac{KPI\ 10_{to-date}}{MC\ 10} * [(KPI\ 3_{to-date} * MC\ 3 * KPI\ 2_{to-date} * MC\ 2 + KPI\ 1_{to-date} * MC\ 1 + KPI\ 4_{to-date} * MC\ 4 + KPI\ 5_{to-date} * MC\ 5 + KPI\ 6_{to-date} * MC6 + KPI\ 7_{to-date} * MC\ 7 + KPI\ 8_{to-date} * MC\ 8 + KPI\ 9_{to-date} * MC\ 9)\ ]$$ Eq 13

Where:

MC 1= *Monte Carlo variable representing percent change of KPI 1 (Hourly rate for direct field labor) from to-date value to at-completion value.*

MC 2= *Monte Carlo variable representing percent change of KPI 2 (Hourly rate for indirect staff labor) from to-date value to at-completion value.*

MC 3= *Monte Carlo variable representing percent change of KPI3 (Ratio of indirect hours to direct hours) from to-date value to at-completion value.*

MC 4= *Monte Carlo variable representing percent change of KPI 4 (Hourly rate for equipment) from to-date value to at-completion value.*

MC 5= *Monte Carlo variable representing percent change of KPI 5 (Hourly rate for subcontractors) from to-date value to at-completion value.*

MC 6= *Monte Carlo variable representing percent change of KPI 6 (Hourly rate for material) from to-date value to at-completion value.*

MC 7= *Monte Carlo variable representing percent change of KPI 7 (Hourly rate for small tools and consumables (ST&C)) from to-date value to at-completion value.*

MC 8= *Monte Carlo variable representing percent change of KPI 8 (Hourly rate for travel and subsistence) from to-date value to at-completion value.*

MC 9= *Monte Carlo variable representing percent change of KPI 9 (Hourly cost for "other" costs) from to-date value to at-completion value.*

MC 10= *Monte Carlo variable representing change of percent of direct hours spent to-date to at-completion value.*

Note that the standard deviation of Monte Carlo variables described in Equation 13 tends to be higher at the beginning of the project and tends to narrow as the project progresses. Because of this, it is expected that the standard deviation of output Project Cost $_{\text{at Completion}}$ variable would also narrow as the project progresses. This intuitively makes sense, as the range of possible outcomes would be expected to narrow as the project progresses and less uncertainty remains.

Step 2 - Analyzing Past Performance for use in Monte Carlo Calculations

The second step is selecting past representative projects and using these to understand the historic evolution of project KPIs. This knowledge is then used in the Monte Carlo analysis. To do this, 11 projects were selected based on having been completed in the prior 2 years and having total installed cost between $5 million and $100 million (CDN). Contract types for projects that met these criteria include four (4) reimbursable projects, five (5) fixed price projects and two (2) mixed contract projects.

Data was retrieved from the organization's accounting system, JD Edwards (JDE). Using the coding available in the accounting system, data is extracted and analyzed for the ten KPIs described in *Step 1 – Model Development* and described in *Table 1 –KPI Definitions.*

The cumulative value for these KPIs is calculated for each month the project was in execution. A monthly interval was chosen to due to the timing of the accrual process, thus ensuring that all costs were included in the analysis. The physical percent complete is also recorded for each monthly interval. This results in data samples for each KPIs at varying completion percentages during the project.

These KPIs are then normalized based on their known at-completion values. For example, if KPI 3, the ratio of indirect hours to direct hours, is 48% when the project is reported to be 10% complete and KPI 3 finished at 41% when the project was 100% complete, the normalized metric for KPI 3 at 10% complete is calculated as 0.48/0.41 or 1.17. In other words, at this point in the project, KPI 3 (the indirect ratio) is 17% higher than it was at completion. This normalization was repeated for each 5% interval of reported percent complete to generate a curve showing the evolution of each KPI throughout the project life cycle. An example of a generated curve is shown in Figure 1. Note that these curves by definition will always converge to 1 at completion. The 5% interval was chosen as tradeoff between granularity and computational efficiency. Because percent completes and KPI values were only measured at the end of each month, most KPI values for the 5% progress intervals were linearly interpolated between monthly measurements. For example, if one month a KPI measured 1.10 and the project was reported as 50% complete and at the subsequent month,

the same KPI was measured at 1.00 and the project was reported as 60% complete, the value of this KPI at 55% completion would be interpolated between these two months, which in this example would result in a value of 1.05.

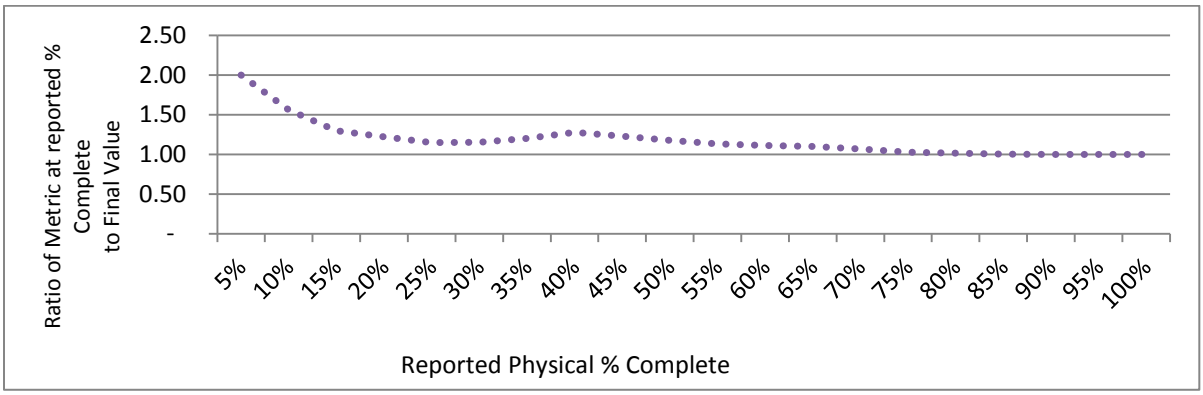

Figure 1 – KPI Evolution across a Single Project (KPI 3; Indirect Ratio)

This process is repeated across each sample project to generate an average evolution for each KPI. A normal distribution is assumed and this allows averages and standard deviations to be calculated for each KPI at each 5% interval of physical percent completion. Figure 2 shows an example of the evolution of a KPI (KPI 3; Indirect Ratio) across all projects and overlays the mean and curves for ± one standard deviation. Although the sample size of projects was low (n = 11), analysis was performed to confirm that normal distributions were acceptable to model historical KPIs. See Appendix D – Confirming Normal Distribution of Historical KPI's.

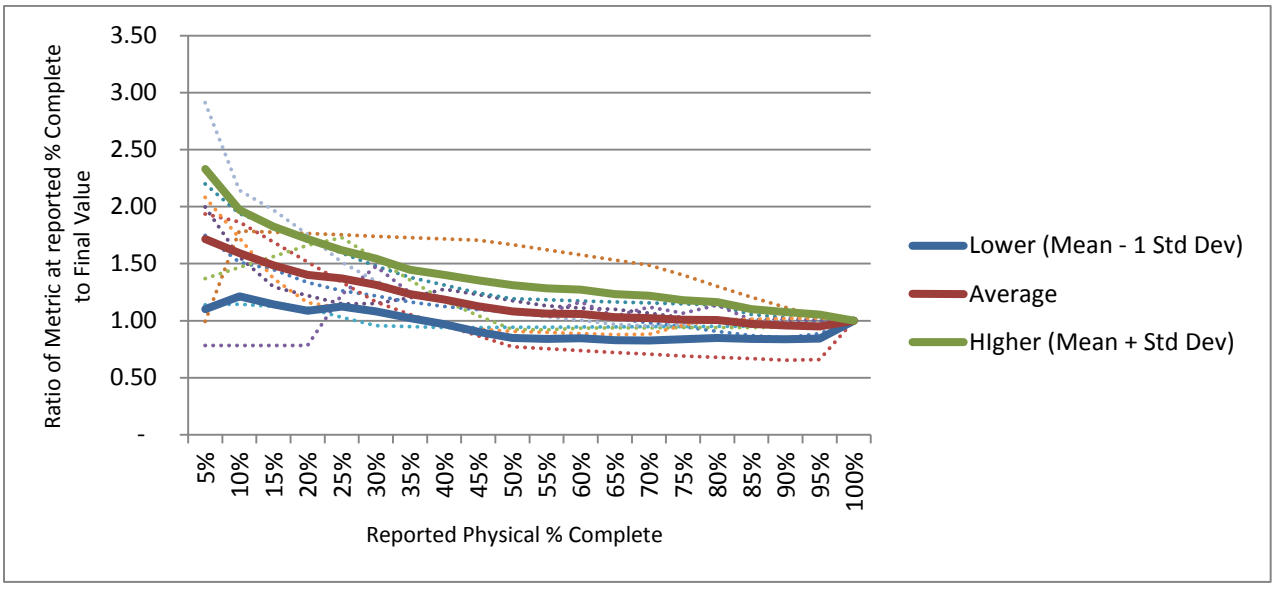

Figure 2 - KPI Evolution across All Projects (KPI 3; Indirect Ratio)

This process is then repeated for all ten (10) KPIs, so that the average historic evolution can be understood for each KPI. The results are shown below in Figure 3 for KPI 1 (Direct Labor), including means and curves for ± one standard deviation. Graphs for the remaining 9 KPI's are shown in Appendix C and are generally similar to KPI 1. For most KPIs, standard deviations are highest early in the project and converge as the project progresses.

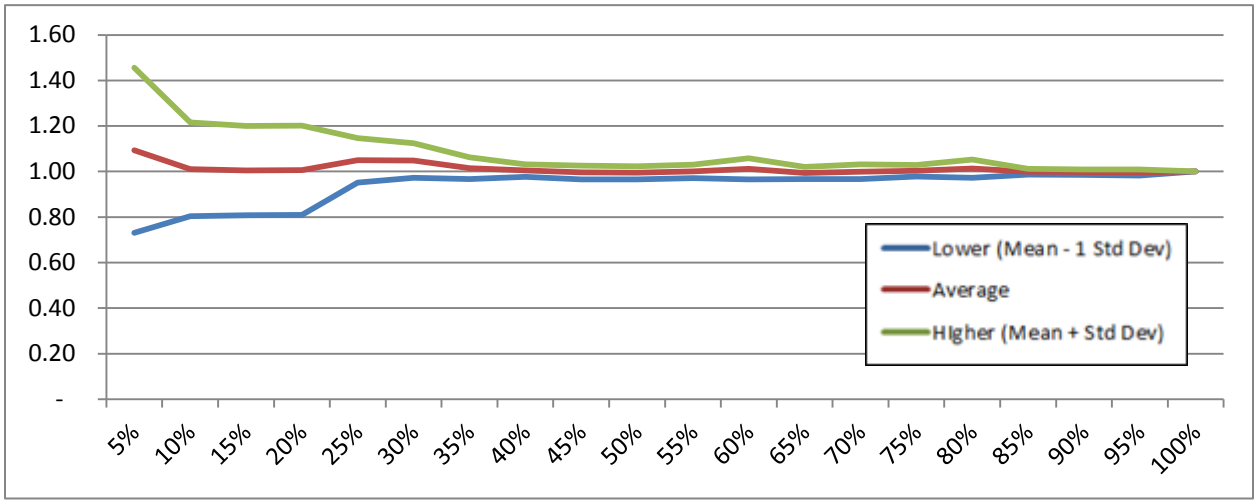

Figure 3 - Normalized KPI (KPI 1; Direct Labor)

Based on analysis of KPI 10; Percent of Direct Hours Spent , it was determined a triangular distribution would be more appropriate than a normal distribution for the associated Monte Carlo input variable MC 10. For a detailed discussion please see Appendix A – Determining Distribution of KPI 10 Hours Spent. The curve for minimum, maximum and average values used in the triangular distribution for KPI10 (percent of direct hours spent) is shown in Figure 4.

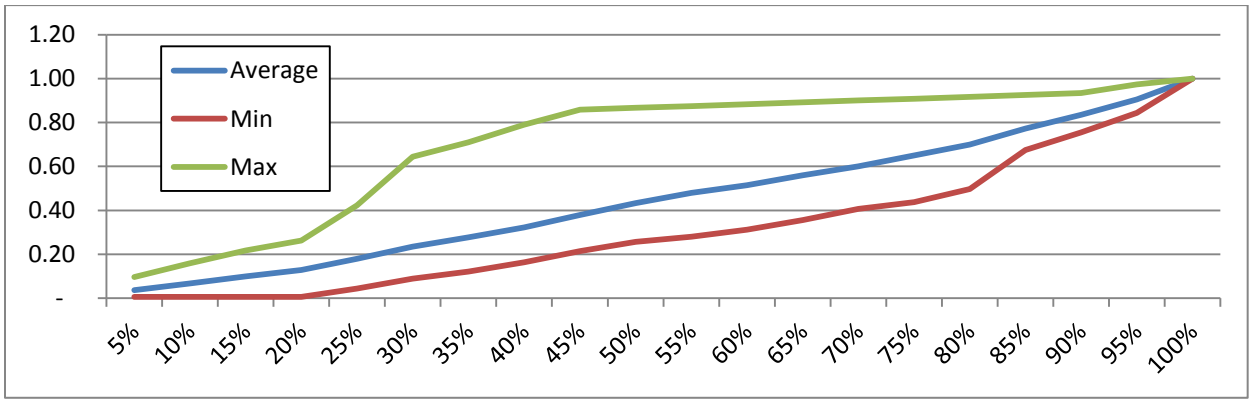

Figure 4 – KPI 10 Evolution

In addition to providing an understanding of the individual probability distributions for each KPI, correlation coefficients between KPIs are also calculated throughout the progression of the project. For example, at 80% complete, the correlation coefficients between KPIs 1 through 10 are shown in Table 2.The correlation coefficients measure the degree of statistical relation between two variables and the correlation coefficients are calculated using the commonly used Pearson Product-Moment method, which is below shown in Equation 14 [3].

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{Eq 14}$$

Where:

x and y represent the two datasets being tested for correlation

r is the degree of correlation between the datasets

For example, at 80% complete, the correlation coefficients between KPIs 1 through 10 are shown in Table 2:

	Cost per Hour for Direct Labour	Cost per Hour for Indirect Labour	Cost per Hour for Equipment	Cost per Hour for Material	Cost per Hour for Subcontracts	Cost per Hour for STC	Cost per Hour for Travel	Other Cost Per Hour	Indirect Ratio	Percent of Direct Hours Spent
Cost per Hour for Direct Labour		0.90	0.42	0.19	0.46	-0.32	0.36	-0.23	0.26	-0.14
Cost per Hour for Indirect Labour			0.37	0.22	0.41	-0.14	0.35	-0.15	0.27	-0.26
Cost per Hour for Equipment				0.08	0.51	-0.85	0.52	-0.05	-0.02	-0.39
Cost per Hour for Material					0.71	0.07	0.24	-0.38	0.04	-0.62
Cost per Hour for Subcontracts						-0.44	0.46	-0.62	-0.37	-0.42
Cost per Hour for STC							-0.15	0.17	0.14	0.31
Cost per Hour for Travel								0.19	-0.33	0.18
Other Cost Per Hour									0.19	0.29
Indirect Ratio										-0.45
Percent of Direct Hours Spent										

Table 2 – Example Correlation of KPIs @ 80% Physically Complete

The calculation of these correlations is repeated at each 5% progress interval within the project.

Step 3 - Determining Relative Significance of Modelled KPIs

The next analytical task is to determine which modelled KPIs have the greatest impact on the final EAC forecast cost. This is done by examining the regression coefficients of the various modelled KPIs. Through this analysis it is determined that KPI 10 (Percent of Direct Hours Spent) was the most significant KPI. The next most significant KPI is KPI 3 (Ratio of Indirect Hours to Direct Hours). The full results are shown in Figure 5 – Modelled Variables Ranked by Order of Importance. This intuitively made sense, as all costs in the model are calculated as *an hourly rate* multiplied by *a forecast number of hours* and both KPIs identified above drive the EAC hours used in these calculations.

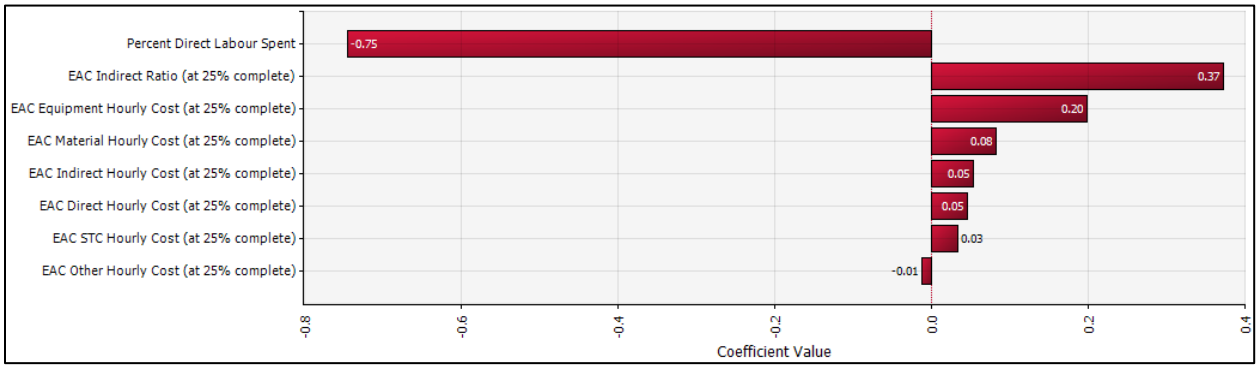

Figure 5 – Modelled Variables Ranked by Importance [1]

Knowing that KPI 10 (Percent of Direct Hours Spent) and KPI 3 (Ratio of Indirect Hours to Total Hours) have the greatest impact on the final forecast value, different approaches to determine the values of these KPIs at completion are explored.

Step 4 – Determining Alternative Methods for Calculating EAC Hours

This investigations initial goal was to forecast a current project's outcome by relying only on historic performance from past project and to-date data on a current projects available from an accounting system. This approach ignores other sources of forecast data that may be available on the current projects. Deliverables such as construction schedules, staffing plans and other estimates provide important additional forecasting inputs which can increase the overall model's accuracy, especially at early stages of the project.

[1] Figures 5-8 and 11 are generated from Palisade @RISK®, Version 6.3.1.

Based on an analysis of relative sensitivity of modelled KPIs, KPI 10 (Percent of Direct Hours Spent) and KPI 3 (Ratio of Indirect Hours to Total Hours) are determined to be the most sensitive to EAC total cost. These KPIs in turn drive the EAC direct and indirect labor hours and account for the greatest amount of variability in the forecast. To attempt to reduce the variance caused by these KPIs, three model types are analyzed to explore how alternate means of determining EAC direct and indirect labor hours can be incorporated into the Monte Carlo analysis.

Model Type 1

Model Type 1 uses historic performance to determine both EAC direct and indirect labor hours and represents what has been discussed above. Recall that…

$$KPI\ n_{At-Completion} = KPI\ n_{To-Date} * \ MC\ n \qquad \text{Eq 12}$$

… so knowledge of current values of KPI 10 (To Date Direct Labor Hours) and KPI 3 (Indirect Ratio) combined with their known historic performance for similar projects (represented by MC 10 and MC 3, respectively) would allow estimation of the at-completion direct labor hours and indirect labor hours. Again, this is represented by Equation 13, which has been previously discussed:

$$Project\ \ Cost_{At-Completion} = \ \frac{KPI\ 10_{to-date}}{MC\ 10} * [(KPI\ 3_{to-date} * MC\ 3 * \ KPI\ 2_{to-date} * MC\ 2 + KPI\ 1_{to-date} * MC\ 1 + \ KPI\ 4_{to-date} * MC\ 4 + \ KPI\ 5_{to-date} * MC5 \ + \ KPI\ 6_{to-date} * MC\ldots . KPI\ 7_{to-date} * MC\ 7 + KPI\ 8_{to-date} * MC\ 8 + KPI\ 9_{to-date} * MC\ 9)\ ] \qquad \text{Eq 13}$$

The problem with this model is that the Monte Carlo input variables MC 10 and MC 3 have large variances early in the project, and thus the resulting forecast EAC total cost also has a large variance. This model type requires the fewest inputs but offers the least accuracy.

Model Type 2

Where in Model Type 1, KPI $10_{\text{At-Completion}}$ (EAC Direct Labor Hours) is calculated based on past performance of similar projects, Model Type 2 assumes that KPI $10_{\text{At-Completion}}$ can be estimated more accurately by some external means. Generally, this would be either a labor-loaded schedule or some other form of labor estimate. In other words, KPI 10 is not estimated based on past performance but rather some alternate method. This formula for Model Type 2 is shown below (Equation 15), with differences from Equation 13 (Model Type 1) bolded.

$$Project\ Cost_{At-Completion} = \boldsymbol{EAC\ Direct\ Labour\ Hours} * [(KPI\ 3_{to-date} * MC\ 3 * KPI\ 2_{to-date} * MC\ 2 + KPI\ 1_{to-date} * MC\ 1 + KPI\ 4_{to-date} * MC\ 4 + KPI\ 5_{to-date} * MC5 + KPI\ 6_{to-date} * MC6 + KPI\ 7_{to-date} * MC\ 7 + KPI\ 8_{to-date} * MC\ 8 + KPI\ 9_{to-date} * MC\ 9)\ ] \quad \text{Eq 15}$$

Where:

EAC Direct Labor Hours = External estimate of final Direct Labor Hours

When evaluating the Type 2 Model, since the accuracy is evaluated by testing the model using data on a project that had already been completed, the EAC Direct Labor Hours are already known. To account for the inherent uncertainty resulting from a contemporaneous estimate of EAC Direct Labor Hours at a given point (percent complete) in the project, a degree of uncertainty is modelled for the Estimate-to-Complete (ETC) component of Direct Labor Hours. Recall that:

$$EAC\ Direct\ Labor\ Hours = To\ Date\ Direct\ Labor\ Hours + ETC\ Direct\ Labor\ Hours \quad \text{Eq 16}$$

ETC Direct Labor hours are calculated based on known values of EAC Direct Labor Hours (known due to testing a completed project with known final hours) subtracted from To Date Direct Labor Hours. The ETC Direct Labor hours are then treated as a modelled Monte Carlo variable with accuracy of +/- 10% for the purposes of testing the model. These modelled ETC Direct Labor Hours are then added to the To Date Direct Labor Hours to produce an estimate of EAC Direct Labor Hours that incorporates the uncertainty in the contemporaneous estimate of EAC Direct Labor Hours at a particular stage of the project. The formula for EAC Direct Labor Hours is shown below, where the ETC Direct Labor Hours takes into account the stated uncertainty. This model type is a mid-point of the three models for required number of inputs and accuracy.

Model Type 3

Where Model 1 and 2 assume different methods of calculating EAC Direct Labor Hours, they both use past performance to calculate EAC Indirect Labor Hours. Model 3 assumes that EAC Indirect Labor Hours are calculated from external forecasting documents, such as an indirect staffing plan, rather than past performance. The formula for Model Type 3 below is shown below (Equation 17), with differences from Model Type 2 bolded.

$$Project\ \ Cost_{At-Completion} = \ \ EAC\ \ Direct\ Labour\ Hours * \left[\left(\frac{\boldsymbol{EAC\ Indirect\ Labour\ Hours}}{\boldsymbol{EAC\ Direct\ Labour\ Hours}} * \right.\right. \quad \text{Eq 17}$$

$$KPI\ 2_{to-date} * MC\ 2 + \ KPI\ 1_{to-date} * MC\ 1 + \ KPI\ 4_{to-date} * MC\ 4 + \ KPI\ 5_{to-date} * MC5 \ +$$

$$\left.\left. KPI\ 6_{to-date} * MC6 + \ KPI\ 7_{to-date} * MC\ 7 + KPI\ 8_{to-date} * MC\ 8 + KPI\ 9_{to-date} * MC\ 9\right)\right]$$

Where:

EAC Direct Labor Hours = External estimate of at-completion Direct Labor Hours

EAC Indirect Labor Hours = External estimate of at-completion Indirect Labor Hours

The methodology for including contemporaneous uncertainty in the forecast of EAC Indirect Labor Hours is the same used for EAC Direct Labor Hours in Model Type 2. As with Model Type 2, the ETC component of the EAC Indirect Labor Hours is extracted and modelled assuming an accuracy of +/- 10%. The ETC component is then in turn is used to model the EAC Indirect Labor Hours which will now include variance to account for the contemporaneous uncertainty in the forecast.

In Model Type 3, EAC Direct Labor Hours are also calculated based on an assumed knowledge of the final forecast value, in the same manner described in Model Type 2. This model type requires the most inputs but offers the most accuracy.

All three model types are summarized in the Table 3 – Summary of Model Type characteristics:

Model Type			1	2	3
Model Type Description			Model both EAC direct and indirect labor hours based on past performance	Assume knowledge of EAC direct labor hours and model EAC indirect labor hours based on past performance	Assume knowledge of EAC direct *and* indirect labor hours
Model Metric	EAC Direct Labor Hours (KPI 10)	Calculation Method for Direct Labor	Based on historic performance	Based on assumed knowledge of final EAC Direct Labor Hours using external estimates.	
		Historic Metric Used	Percent of Direct Hours Spent (KPI 10)	None	
		Accuracy	Based on history; Generally less accurate	+/- 10% of ETC Direct Labor Hours	
		Required Additional Input	None	Labor-loaded schedule	
	EAC Indirect Labor Hours (KPI 3)	Calculation Method for Indirect Labor	Based on historic performance		Based on assumed knowledge of final EAC Indirect Labor Hours based on external estimate.
		Historic Metric Used	Ratio of Indirect Hours to Direct Hours (KPI 3)		None
		Accuracy	Based on history; Generally less accurate		+/- 10% of ETC Indirect Labor
		Required Additional Input	None		Staffing plan
Accuracy			Low	Medium	High
Required Inputs			Least	Medium	Most

Table 3 – Summary of Model Type characteristics

RESULTS

Results of Model Prediction (Ignoring Correlations)

The three model types described above are tested against three additional projects to evaluate their respective performance. These three projects have average costs of $15.0M with the smallest being $7.9m and the largest being $20.9m. All are lump sum contracts.

The results of the three model types' predicted EAC Total Cost are shown below for a single project in Figures 6, 7 and 8. This analysis is based on retrospective analysis of performance at 50% complete. Analysis at other percent completes is discussed in the next section. An accuracy range of within +/- 10% of the final project cost is highlighted on the probability distribution curves in Figures 6, 7 and 8 to determine the confidence for which each model predicts final project costs to within this level of accuracy. This process is actually a Bernoulli trial to determine whether a sample outcome of the Monte Carlo trial falls within the specified accuracy range of the known final cost (+/- 10%). The resulting probability of success in these Bernoulli trials is used to estimate the confidence of forecasting final cost for each Model Type to the specified level of accuracy.

For Model Type 1, based on retrospective analysis, the model predicts with 18.3% confidence that the final cost will fall within 10% of the actual observed final cost when the project is 50% complete. The distribution for EAC Total Cost is shown below in Figure 6. The main cause for the EAC Cost distribution curve shifting *lower* appears to be in hind-sight, the percent of hours spent was lower than the historical average for projects at 50% complete (38.3% on this project vs. 43.3% historical average). This could be due to an over-reporting of physical percent complete.

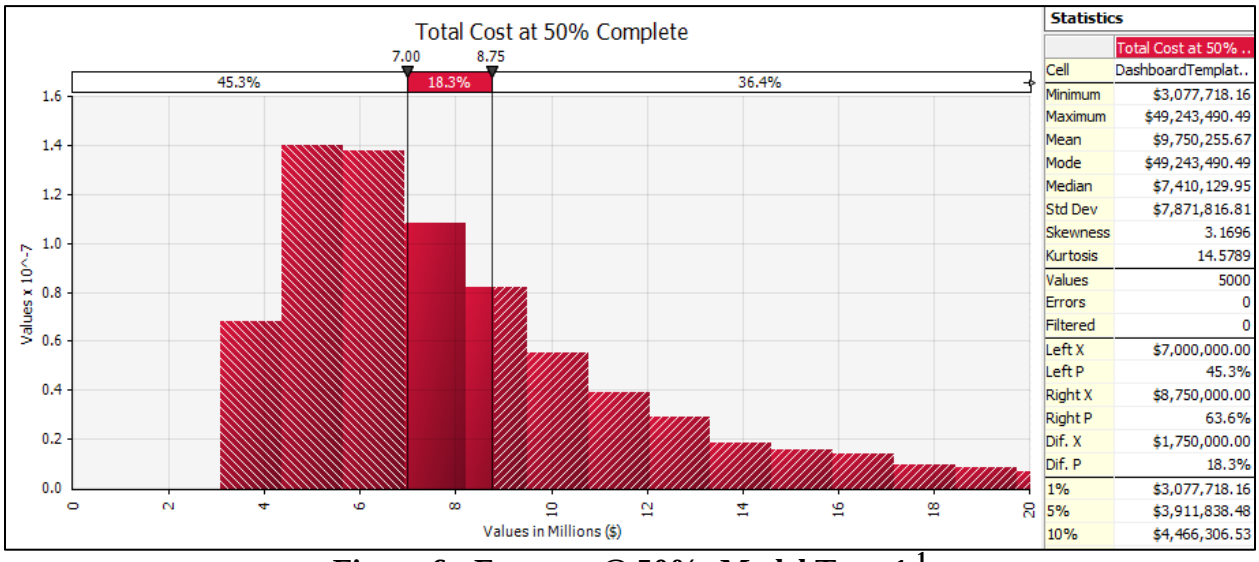

Figure 6 - Forecast @ 50%; Model Type 1 [1]

For Model Type 2, the model predicts with 57.6% confidence that the final cost will fall within 10% of the actual final cost. The distribution for EAC Total Cost is shown below in Figure 7. The main cause for the EAC Total Cost distribution curve shifting towards *higher* projected final costs appears to be in hind-sight, the indirect ratio fell further on this project than it had historically at 50% complete This particular project had a 22% reduction in the indirect ratio at completion versus an 11% historic average. This could be due to a quicker ramp down of indirect staff than historically has taken place.

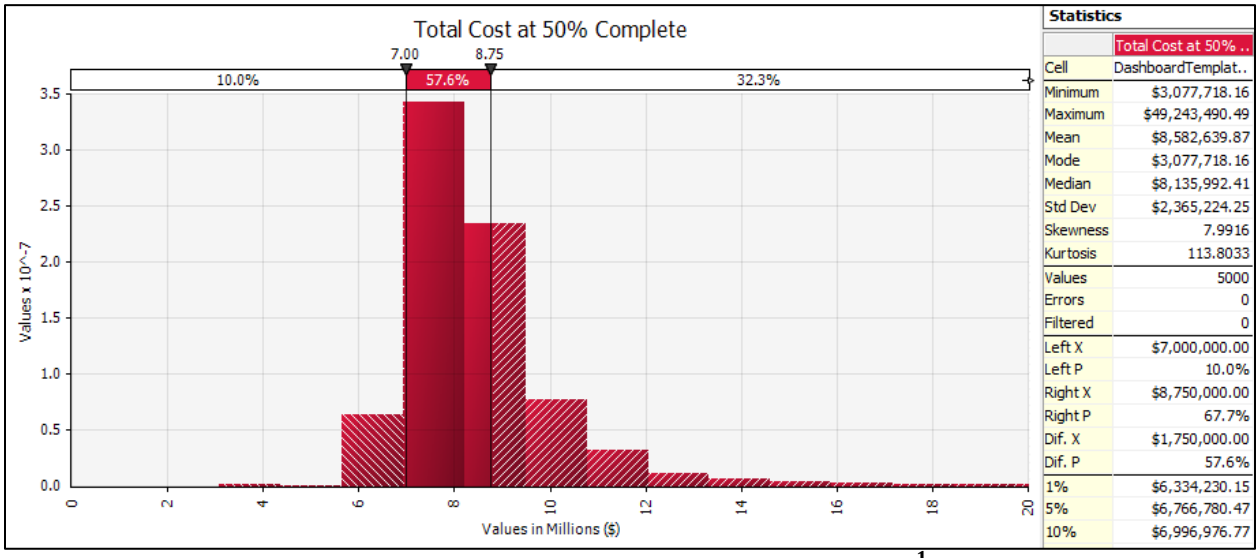

Figure 7 - Forecast @ 50%; Model Type 2 [1]

For Model Type 3, the model predicts with 66.8% confidence that the final cost will fall within 10% of the actual final cost. The distribution for EAC Total Cost is shown below in Figure 8.

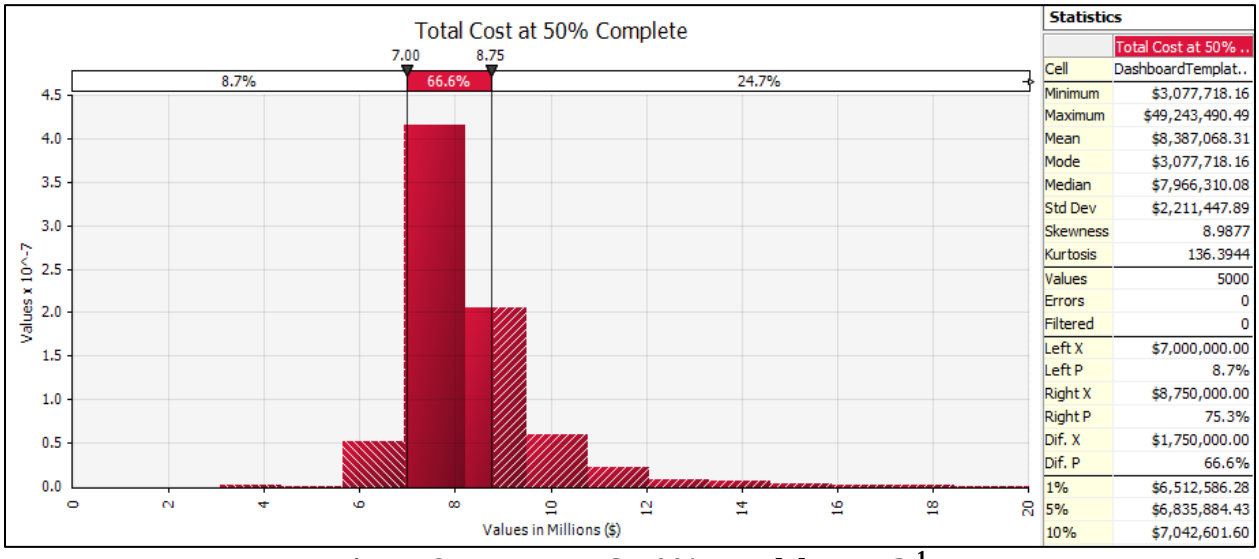

Figure 8 - Forecast @ 50%; Model Type 3 [1]

The above probability distributions show the confidence of achieving a forecast of +/- 10% of the final known cost at 50% complete for a single sampled project. Figure 9 expands this analysis to show the evolution of each models' accuracy in predicting the final EAC cost to within +/-10%, averaged across the three back-tested projects. As expected, Model 3 outperforms all other models.

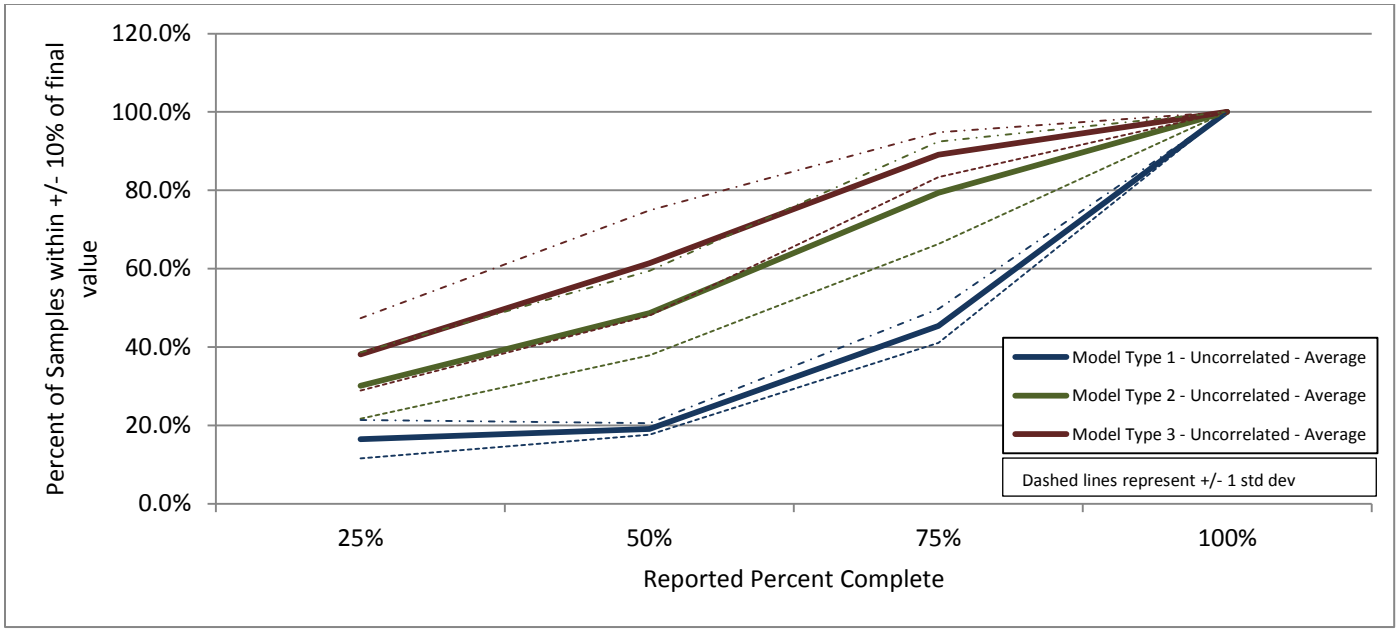

Figure 9 – Accuracy for Final Cost Forecast

Based on linear interpolation of the above analysis, Table 4 below summarizes the required percent complete on a project to accurately forecast final cost to within +/- 10% of the actual EAC Total Cost using each model. Required percent completes are shown to achieve both a 50% (P50) and 90% (P90) confidence for this level of accuracy.

	Percent Complete Required for accuracy of +/-10% of Final Total Cost	
	P50	P90
Model Type 1	77%	95%
Model Type 2	51%	87%
Model Type 3	38%	77%

Table 4 –Analysis of Model Convergence

As shown in Table 4, Model Type 1 does not appear to provide accuracy within +/- 10% until late in the project. As such, it is suitable for only the most basic of sanity checks at earlier stages. Model Type 3 begins to show reasonable levels of confidence (P50) for this level of accuracy when the

project is roughly half complete and achieves high confidence levels (P90) when the project is 87% complete. Model Type 3 performs the best and provides an estimate for +/- 10% of Final EAC Cost at reasonable levels of confidence (P50) very early in the project (38%) and achieves a 90% confidence interval at 77% physically complete. As describe above, Model Types 2 and 3 assume the ETC hours can be estimated to accuracy within 10%, which is an important assumption worth repeating.

Results of Model Prediction (Accounting for Correlations)

In the above analysis, each modelled variable is assumed to be independent from the other KPIs. This can lead to under estimation of cases where extreme events for two KPIs occur simultaneously due to correlation of these KPIs. This in turn underestimates the likelihood of extreme cases in the forecast EAC Total Cost. One example of a correlated variable is KPI 1 (*Direct* Labor Cost per Hour) and KPI 2 (*Indirect* Labor Cost per Hour). These two factors will likely be affected by the same market pressure and thus move together. Treating them as independent variables would understate the potential for sharp changes in market conditions occurring for these KPIs simultaneously. Table 3 confirms this correlation, showing a positive correlation of 0.9 between these two KPIs.

In order to model the impact of the historic correlations between all KPIs, the three Model Types were rerun taking into account KPI correlations. Random samples for the Monte Carlo variables are generated that reflects historic correlations and known distributions of these KPIs (see Appendix B).

Figure 10 shows the percent of modeled samples that fall within +/- 10% of the final EAC cost for each model type. Curves for treating KPIs as either independent or correlated are both shown. Accounting for correlations results in all three models taking longer (i.e. additional physical completion) to achieve the desired P50 confidence. This is as expected as correlations result in larger standard deviations in the models since outlying events are now more likely, as discussed above. Hence, these models take longer to converge.

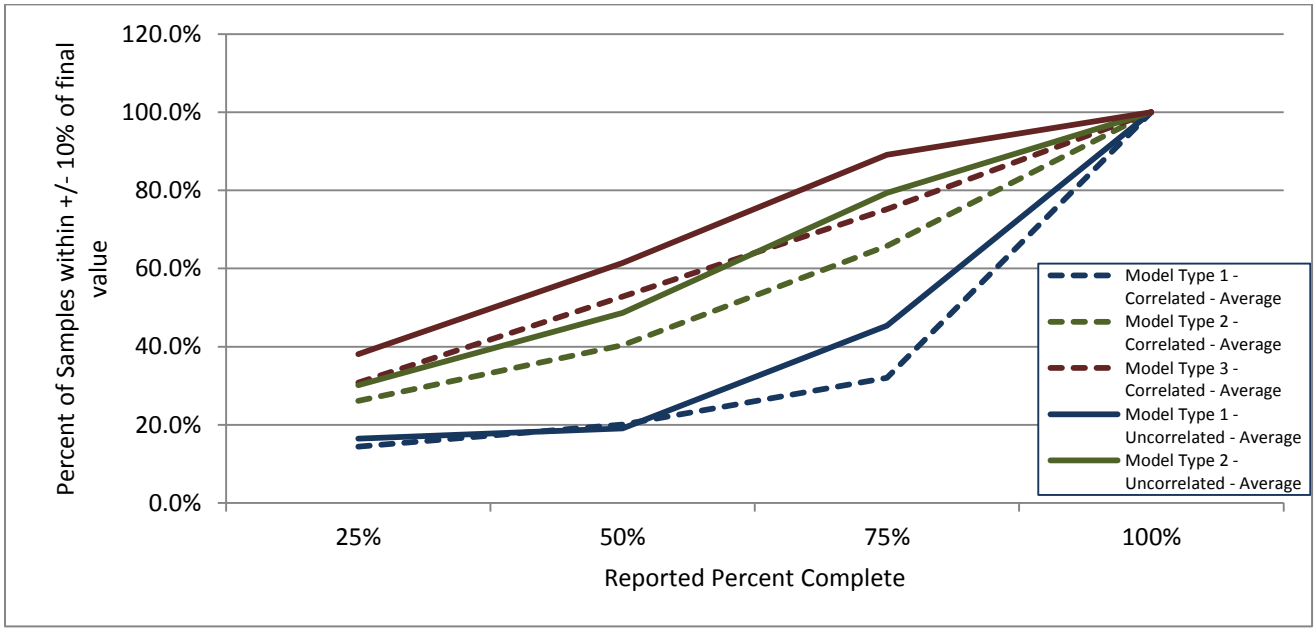

Figure 10 – Model Comparison, Accounting for Correlations

A comparison of required physical completion to achieve desired accuracy and confidence (P50, +/- 10% EAC) are shown below in Table 5. Note that increased physical completion requirements range from an additional 3% (for Model Type 1) to 17% (for Model Type 3). The average increased requirement for completion is 7.3% across the three models.

	Percent Complete Required to Achieve Confidence of +/-10% Accuracy		Difference Between Correlated and Independent Required Percent Complete
	P50 - Independent	P50 - Correlated	
Model Type 1	77%	82%	4.5%
Model Type 2	51%	59%	8.3%
Model Type 3	38%	47%	9.0%
Average			7.3%

Table 5 – Effect of KPI Correlations on Convergence

Discussion on Conversion of KPI's

Some KPIs converge more quickly to their final EAC values than others. Since the KPI's can also be used as inputs into a manual, deterministic cost forecasting exercise, understanding when convergence occurs for each KPI can provide guidance on when it is appropriate to use that KPI in a manual forecasting excercise. Convergence of KPIs was analyzed at various points of project completion to determine when a desired level of accuracy and confidence is achieved for each KPI.

For this analysis the confidence intervals are calculated for as 50% confidence and 90% confidence (P50 and P90, respectively). Accuracy ranges are analyzed based on being within +/- 20%, +/- 10% and +/- 5% of the final EAC number. The required physical percent complete in order to achieve these accuracies and confidence intervals are shown below for each metric. For example, to know within +/- 5% what the final *cost per hour for direct labor (KPI 1)* is, with 90% confidence, at least 25% physical completion is needed in order to achieve this level of accuracy and confidence. KPIs that converge earliest in the project are more appropriate to be used as inputs into a deterministic cost forecast. It is noted that hourly cost for both Direct (KPI 1) and Indirect labor (KPI 2) are the first values to converge and offer a 90% confidence of being within 5% of the final value when the project is at 25% and 50% physically complete, respectively. The Indirect Ratio also converges relatively early, offering accuracy within 10% of the final value at 25% physically complete. Other KPIs converge later in project and for these, alternate methods than hourly KPIs may be more appropriate to use in a deterministic cost forecast. For example, the estimate of at-completion subcontractor costs may be more accurately determined from a review of cost commitments of purchase orders rather than using KPI 5. Similarly, at-completion equipment costs may be best forecast from an equipment log, rather than KPI 4. Convergence of all KPIs is shown below in Table 6.

Able to predict EAC KPI value within…	**+/- 5%**		**+/- 10%**		**+/- 20%**	
Confidence interval	**P90**	**P50**	**P90**	**P50**	**P90**	**P50**
KPI 1 - Cost per Hour for Direct Labor	25%	10%	15%	5%	5%	5%
KPI 2 - Cost per Hour for Indirect Labor	50%	5%	5%	5%	5%	5%
KPI 3 - Indirect Ratio	95%	25%	25%	5%	5%	5%
KPI 8 - Cost per Hour for Travel & Subsistence	85%	35%	40%	25%	30%	20%
KPI 9 - Other Cost Per Hour	95%	40%	45%	30%	30%	10%
KPI 6 - Cost per Hour for Material	85%	70%	75%	5%	25%	5%
KPI 4- Cost per Hour for Equipment	95%	55%	60%	25%	40%	10%
KPI 10 - Percent of Direct Hours Spent	85%	75%	75%	50%	55%	10%
KPI 5 - Cost per Hour for Subcontracts	95%	80%	85%	45%	50%	35%
KPI 7- Cost per Hour for STC	NA	NA	NA	NA	95%	85%

Table 6 – Individual KPI Convergence

Benefits to Risk Management by Using Historic Data in Monte Carlo Forecasting

The use of historical data in Monte Carlo simulation improves risk managementby providing a quantitative approach to assessing remaining project contingency. Knowledge of the probability distribution for project final costs enables an evaluation of a deterministic forecast generated by the project team to assess the probability of achieving this forecast performance. If the probability is

assessed to be too low, additional contingency could be added to the project forecast to increase the final cost and thus the probability of completing the project within this amount. With future research, this same concept can be extended so that knowledge of KPI evolution could be combined with initial estimates for KPI's to provide guidance on contingency requirements prior to project sanctioning.

Discussion on Data Capture

It is worth noting that there is a variation in the quality of data captured across various KPIs. Cost "types" are very well segregated (i.e. labor, equipment, material and subcontractors) and expended hours are also accurately captured. Further differentiation of costs (i.e. small tools and consumables, directs vs. indirect costs) rely upon cost coding which can be modified by the project teams and is therefore subject to more uncertainty. Having the same rigor in cost segregation across all KPIs reduces the variance in these KPIs and thus the variance in the overall model. This in turn allows better prediction of project outcomes earlier in the project. It is recommended that more controls be put in place to standardize the use of cost codes across the organization, such as standardized definitions for each cost code. One particular metric that significantly affects the accuracy of this model is percent complete. Standardized rules of credit will help ensure reported percent complete is accurate and comparable across projects.

Another challenge is the sample size for completed projects is relatively low (n = 11). Increasing this sample size would improve the handling of outlying results in the model. It would also allow selection of historic data to better match the project characteristics of the project being forecast such as project size, project location, project scope and contract type. This may help to increase confidence levels and provide more accurate forecasts sooner in the job.

The use of more granular models with greater than ten (10) KPIs could also be explored. Other KPIs that could be relevant include performance factors, scope growth, percentage of local versus non-local labor, percentage of staff versus contractors and percentage of overtime. Additionally some of the ten (10) KPIs could be subdivided such as modeling travel, subsistence, equipment types and facility costs separately. As the number of modelled KPIs increases, it further emphasizes the need to increase the number of sampled projects and sharpen the capture of data on these projects. Finally, additional validation of the model is recommended by back-testing results across additional projects. This would provide additional statistical evidence as to the accuracy of the overall model.

CONCLUSION

This analysis shows that Monte Carlo simulation can provide reasonable estimates of final project outcomes given knowledge of historic project performance. Based on the review of model tests, reasonable estimates for final project costs (+/- 10% of final cost, P50 confidence) can be achieved as early as 38% physically complete if accurate schedules and staffing plans are available (Model Type 3). With schedules but no staffing plans, reasonable estimates can be achieved at 51% complete (Model Type 2). Without either a schedule or a staffing plan and instead relying solely on historic data, reasonable estimates can be achieved at 77% complete (Model Type 1). This shows that the best modelled forecasts come from a combination of historic performance data and project-specific planning documents such as schedules and staffing plans. Accounting for correlations between modelled KPIs requires approximately 7.3% additional physical completion before reasonable estimates can be established. This is as expected due to the increased probability of outlying events with correlated variables and thus a larger standard deviation in the modelled outcome. Accounting for correlations has the advantage of more realistically accounting for simultaneous outlying events.

This analysis also identifies the individual KPI's that are best suited as inputs into a deterministic cost forecasts. For example, KPI 1 (hourly rate for direct labor) converges to within 5% of its at-completion values when the project is only 25% complete (90% confidence) and KPI 2 (hourly rate for indirect labor) converges to within 5% of its at-completion value when the project is 50% complete. Early convergence makes these two KPIs well suited for use as inputs into a deterministic forecast. Other KPIs such as KPI 4 (Hourly Rate for Equipment) and KPI 5 (Hourly Rate for Subcontractors) do not convergence until much later in the project. Thus, deterministic estimates for equipment and subcontract costs are best estimated by other means such as a review of cost commitments or an equipment plan. The benefit of using historical data to model project costs is it enables a quantitative assessment and reevaluation of remaining contingency.

Although these results are promising, it is worth re-iterating the limited sample size (n=11) analyzed in this paper and it is thus recommended that more sample projects be analyzed to increase the statistical rigor of these conclusions.

BIBLIOGRAPHY

No.	Description
1	Clark, D.E. 2001 **Monte Carlo Analysis: Ten Years of Experience.** Cost Engineering Vol 43, No. 6
2	Ilbeigi, Mohammad; Heravi, Gholamrezaq 2010 **Forecasting Construction Project Performance Using Monte –Carlo Simulation Approach.** 2010 AACE International
3	Pearson, E. S. and Hartley, H. O 1972 **Biometrika Tables for Statisticians** Cambridge University Press. pp. 117–123, Tables 54, 55 ISBN 0-521-06937-8.
4	Upton, G and Cook, I. 2008 **A Dictionary of Statistics (2nd Edition)** Oxford University Press ISBN 9780199541454
5	Vrijland, M 2005 **Correlation of Variables in Monte Carlo Simulation** 2006 AACE International Transactions
6	Whiteside, J.D. 2008 **A Practical Application of Monte Carlo Simulation in Forecasting.** 2008 AACE International Transactions

APPENDIX A – DETERMINING DISTRIBUTION OF KPI 10 HOURS SPENT

Initial simulations resulted in huge standard deviations of projected final cost. This was traced to using a normal function to model the Percent of Direct Hours Spent input variable (labelled MC 10 in Equation 13). Logically the Percent of Direct Hours Spent variable would need to fall between 0 and 1. This, however, did not always occur when this variable was modelled as a normal distribution based on means and standard deviations derived from past project data.

The problem occurs because EAC Labor Direct Hours are calculated as follows…

$$EAC\ Direct\ Labour\ Hours\ = \frac{To\ Date\ Direct\ Labour\ Hours}{Percent\ of\ Direct\ Hours\ Spent} \quad \text{Eq A.1}$$

…and thus as the Percent of Direct Hours Spent variable approaches zero, the EAC Direct Labor Hours will approach infinity. Conversely, if the Percent of Direct Hours Spent exceeds 1, then the forecast EAC Direct Labor Hours will be less than the To Date Direct Labor Hours, which logically is impossible.

Unfortunately, early in the project variance for Percent of Direct Hours Spent is high and this can result in Monte Carlo samples that occur outside this logical bound. Although the limits of the Percent of Direct Hours Spent output could be truncated to avoid these extreme values, it was observed that modelled inputs close to these logically impossible values would still result in unrealistic cost estimates and substantially inflate the standard deviations in the final model.

Because of this, it was chosen to use a triangular distribution, rather than a normal distribution to estimate Percent of Direct Hours Spent. A comparison between the two methods is shown in the following graph. Note that the triangular distribution reduces outlying samples and also allows a skewing of the samples to better reflect past observations.

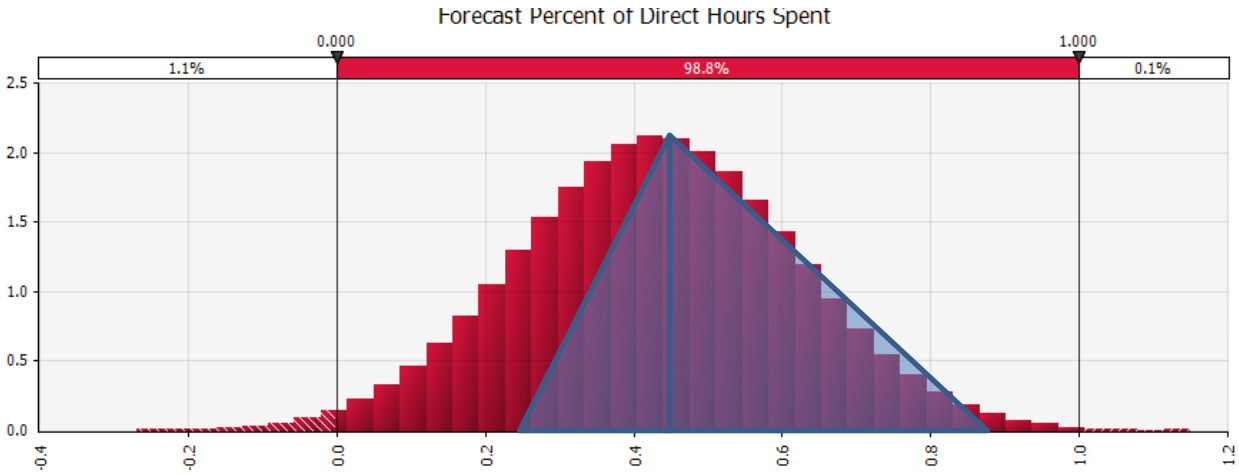

Figure A.1 - Comparison of Distributions for KPI 10

Use of the triangular distribution for Percent of Direct Hours Spent resulted in standard deviations of the overall EAC Forecast Cost dropping from erratic values (standard deviations 20x the mean) to more reasonable values that were the same order of magnitude of the mean. Details for a particular modeled project are shown below. This shows that the triangular distribution gives more realistic values for standard deviations for the final EAC Forecast Cost and is thus recommended for use in future models and used throughout the analysis in this paper.

Reported % Complete	*EAC Forecast Cost*				
	Using Normal Distribution for Percent of Direct Hours Spent		Using Triangular Distribution for Percent of Direct Hours Spent		Percent Reduction in Standard Deviation using Triangular Distribution
	Standard Deviation	Mean	Standard Deviation	Mean	
25%	$2,765,000,000	$20,006,781	$5,140,128	$9,418,403	99.8%
50%	$178,801,500	$8,794,852	$1,906,322	$6,447,976	98.9%
75%	$6,654,754	$8,442,503	$1,006,526	$7,820,124	84.9%

Table A.1 - Comparison of Forecasts for Different Distributions

APPENDIX B – HANDLING CROSS-CORRELATIONS

In Excel it is trivial to generate samples that conform to a given normal distribution with known mean and standard deviation. It is more difficult to generate samples across a number of correlated variables that conform to both the normal distribution of the individual variables *and* the cross correlation between each pair of variables [5].

In order to generate random data for a Monte Carlo simulation that corresponded to the identified cross correlations, a genetic algorithm was utilized. Genetic algorithms operate by optimizing a target function based on trial and error selection of input data in a manner that mimics natural selection. The details of this algorithm are described below.

The first step in designing the genetic algorithm is to design a fitness function. This function is shown below and the goal of genetic algorithm will be to minimize this value. The idea is to compare differences in *means, standard deviations and correlations* of the *observed values on historic projects* and *calculated values based on generated samples*.

The fitness function is:

Eq B.1

$$Error = K_1 \, x \sum Error\ in\ Means + K_2 \, x \sum Error\ in\ Standard\ Deviations + \sum Error\ in\ Correlations$$

Where

$X_1, X_2 \ldots X_n$ = KPIs being modelled (i.e. percent spent, hourly rate for material, etc).

K_1, K_2 = Constants to scale the relative value of mean error, standard deviation error and correlation error in calculating the overall error.

$$\text{Error in Means} = (\mu \text{ of } X_{1,\,Observed} - \mu \text{ of } X_{1,\,Generated\ Sample\ Data})^2 + (\mu \text{ of } X_{2,\,Observed} - \mu \text{ of } X_{2,\,Generated\ Sample\ Data})^2 + (\mu \text{ of } X_{n,\,Observed} - \mu \text{ of } X_{n,\,Generated\ Sample\ Data})^2$$

$$\text{Error in Standard Deviations} = (\sigma \text{ of } X_{1,\,Observed} - \sigma \text{ of } X_{1,\,Generated\ Sample\ Data})^2 + (\sigma \text{ of } X_{2,\,Observed} - \sigma \text{ of } X_{2,\,Generated\ Sample\ Data})^2 + (\sigma \text{ of } X_{n,\,Observed} - \sigma \text{ of } X_{n,\,Generated\ Sample\ Data})^2 + \ldots$$

$$\text{Error in Correlations} = (corr(X_{1,\,Observed} \,\&\, X_{2,\,Observed}) - corr(X_{1,\,Generated\ Data} \,\&\, X_{2,\,Generated\ Data}))^2 + (corr(X_{1,\,Observed} \,\&\, X_{3,\,Observed}) - corr(X_{1,\,Generated\ Data} \,\&\, X_{3,\,Generated\ Data}))^2 + (corr(X_{n,\,Observed} \,\&\, X_{n-1,\,Observed}) - corr(X_{n,\,Generated\ Data} \,\&\, X_{n-1,\,Generated\ Data}))^2$$

Once the target function had been established, 100 sample data sets were populated. Each of the sample data sets contained 200 values for each of the KPIs. This was analogous to having 100 data sets, each containing samples for 200 theoretical projects.

The initial sample data was populated based on observed mean and standard deviation of each KPI but no consideration was initially taken for correlations. This represented the first generation of data. Each of the sample data sets was scored based on the fitness error function (Equation 18). The subsequent generations of sample data sets were determined as follows:

- The top 25% of sample data sets (based on score of fitness function of the previous generation's data) were held over from the previous generation.
- 25% new random sample data sets were added
- 25% of the best scoring sample data sets from the previous generation had between 10% to 90% of their data randomly mutated (changed)
- 25% of the best scoring sample data sets from the previous generation were randomly cross bred together.

This process created the next generation. Because this generation contained the top scoring samples from the previous generation it would at a minimum score equally well on the target function. However, by adding random mutation and "cross breeding" the best samples, occasionally even higher scoring samples would emerge. This was repeated over several thousand generations to produce sample data that closely resembled the observed data for mean, standard deviation and correlations for all variables. A graph showing the improvement of the scoring function over many generations is shown below in Figure B.1.

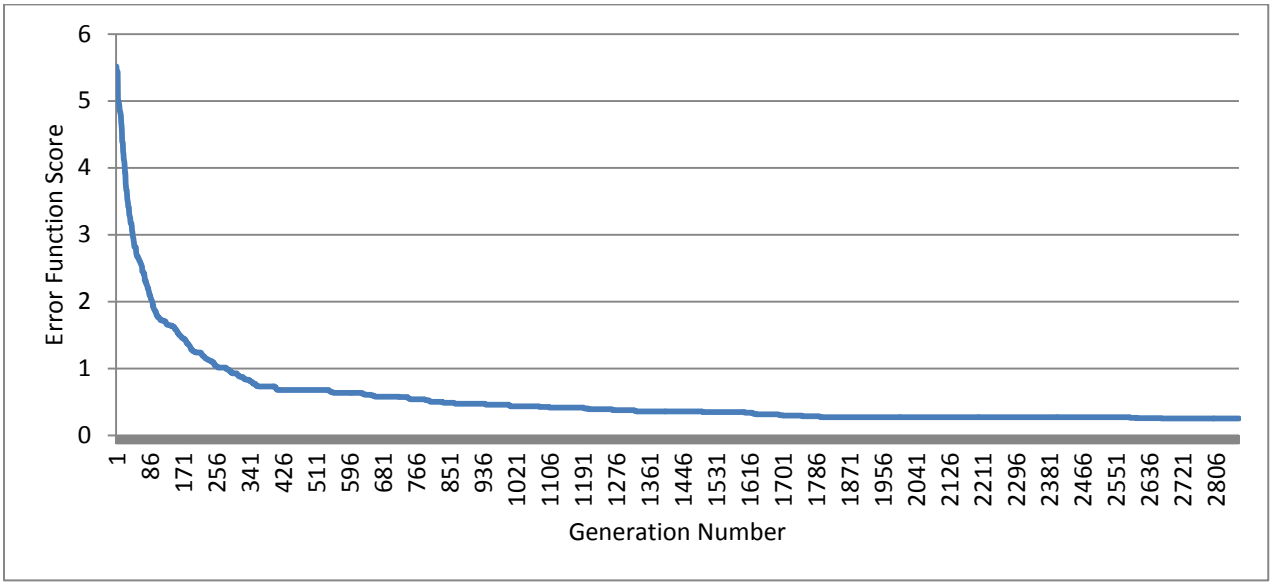

Figure B.1 – Fitness Function per Generation

Using this technic, 200 sample data points were generated for use in Monte Carlo simulations for projects at 25%, 50% and 75% complete. This data was then used to model project performance *taking into account cross correlations of data*. Examples of observed vs. generated sample data statistics are shown on the following page in tables B.1 and B.2.

	Cost per Hour for Directs	Cost per Hour for Indirects	Cost per Hour for Equipment	Cost per Hour for Material	Cost per Hour for Subcontracts	Cost per Hour for STC	Cost per Hour for Travel & Subsistence	Other Cost Per Hour	Indirect Ratio	Percent of Direct Hours Spent
Normal Distribution Data										
Mean	0.99	1.06	0.85	1.51	0.99	2.76	1.11	1.00	1.12	0.43
Standard Deviation	0.03	0.10	0.27	0.48	0.27	3.21	0.19	0.11	0.23	0.19
Cross Correlations										
Cost per Hour for Directs		0.25	-0.01	0.55	0.14	-0.14	0.21	-0.55	0.25	0.16
Cost per Hour for Indirects			-0.13	0.38	-0.27	0.45	-0.21	-0.09	0.71	-0.26
Cost per Hour for Equipment				0.07	0.53	-0.43	0.80	-0.04	-0.03	-0.01
Cost per Hour for Material					0.48	-0.22	0.26	-0.54	0.64	-0.67
Cost per Hour for Subcontracts						-0.45	0.67	-0.62	-0.18	-0.45
Cost per Hour for STC							-0.20	0.36	-0.09	0.15
Cost per Hour for Travel & Subsistenc								-0.09	-0.17	0.01
Other Cost Per Hour									0.01	0.25
Indirect Ratio										-0.52
Percent of Direct Hours Spent										

Table B.1 - Characteristics of Generated Sample Data

	Cost per Hour for Directs	Cost per Hour for Indirects	Cost per Hour for Equipment	Cost per Hour for Material	Cost per Hour for Subcontracts	Cost per Hour for STC	Cost per Hour for Travel & Subsistence	Other Cost Per Hour	Indirect Ratio	Percent of Direct Hours Spent
Normal Distribution Data										
Mean	0.99	1.05	0.83	1.52	0.96	2.74	1.12	1.01	1.12	0.45
Standard Deviation	0.03	0.10	0.25	0.46	0.27	3.49	0.20	0.11	0.21	0.19
Cross Correlations										
Cost per Hour for Directs		0.22	-0.03	0.45	0.18	-0.11	0.14	-0.49	0.21	0.03
Cost per Hour for Indirects			-0.16	0.30	-0.21	0.35	-0.19	-0.10	0.59	-0.25
Cost per Hour for Equipment				0.14	0.47	-0.37	0.60	-0.07	-0.04	-0.05
Cost per Hour for Material					0.42	-0.22	0.21	-0.51	0.48	-0.55
Cost per Hour for Subcontracts						-0.44	0.57	-0.50	-0.10	-0.34
Cost per Hour for STC							-0.25	0.35	0.00	0.15
Cost per Hour for Travel & Subsist								-0.15	-0.13	-0.02
Other Cost Per Hour									-0.04	0.23
Indirect Ratio										-0.45
Percent of Direct Hours Spent										

Table B.2 - Characteristics of Historic Observed Data

APPENDIX C – EVOLUTION OF KPI's

Additional progressions of various KPIs are shown below.

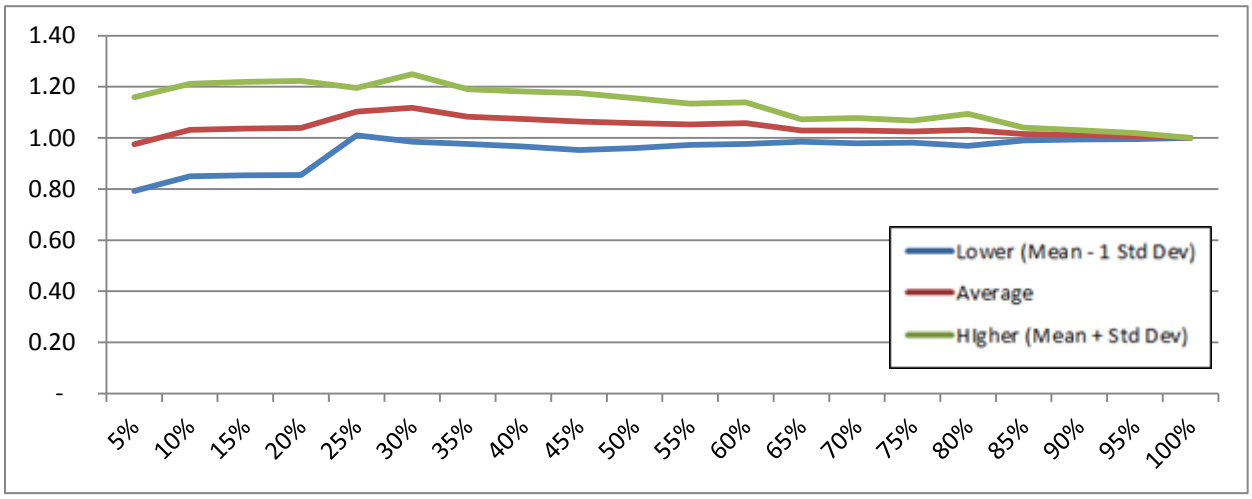

Figure C.1 – KPI2; Normalized Cost per Hour for Indirect Labor

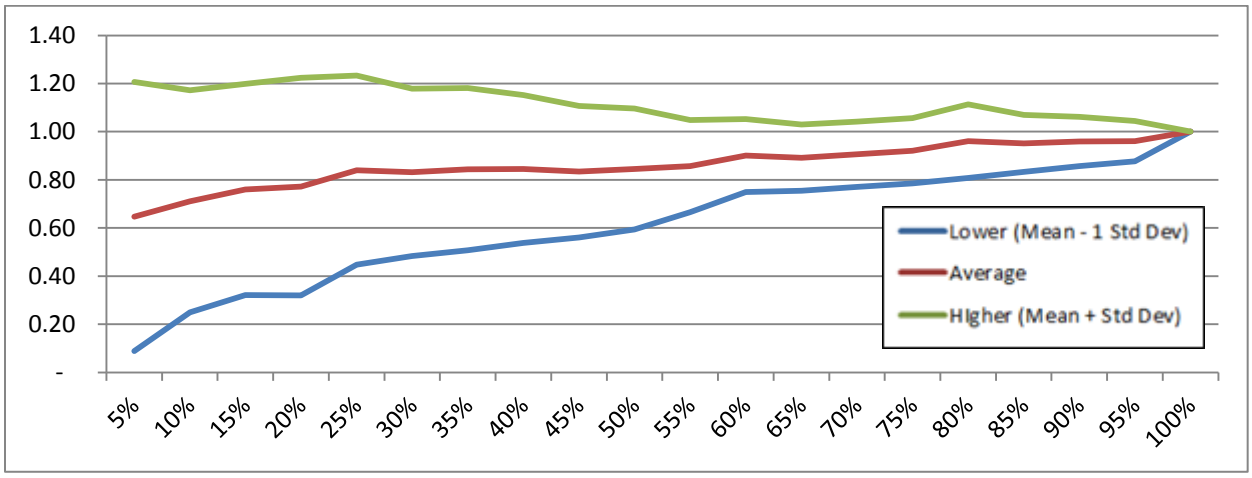

Figure C.2 – KPI 4; Normalized Cost per Hour for Equipment

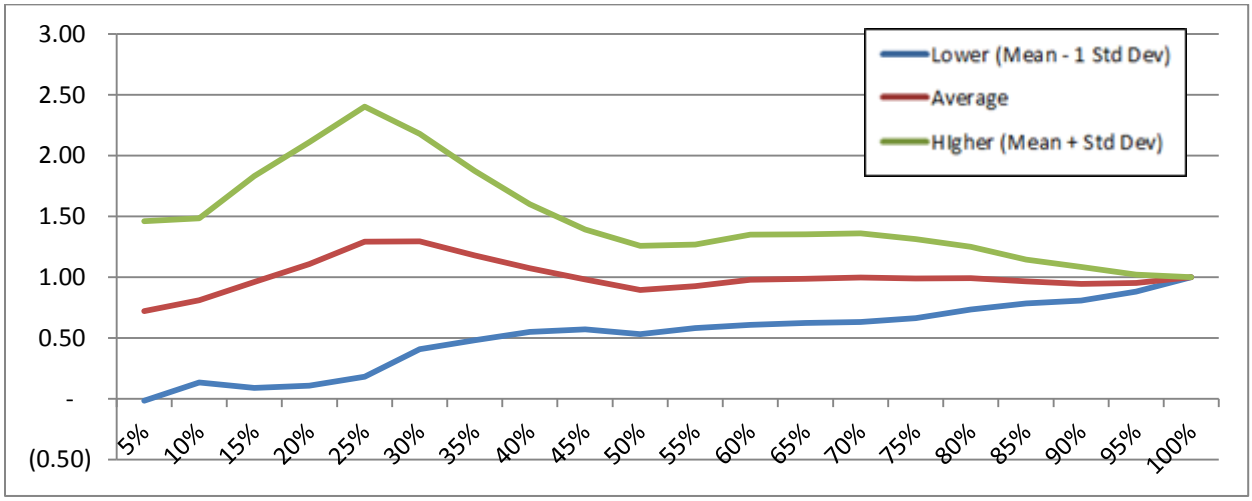

Figure C.3 – KPI 5; Normalized Cost per Hour for Subcontracts

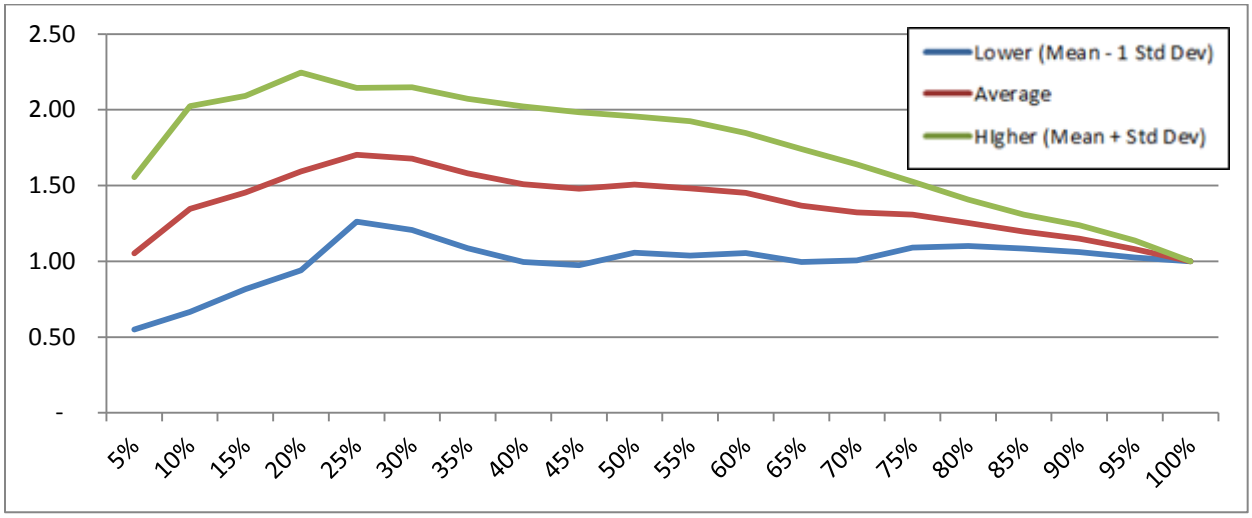

Figure C.4 – KPI 6; Normalized Cost per Hour for Material

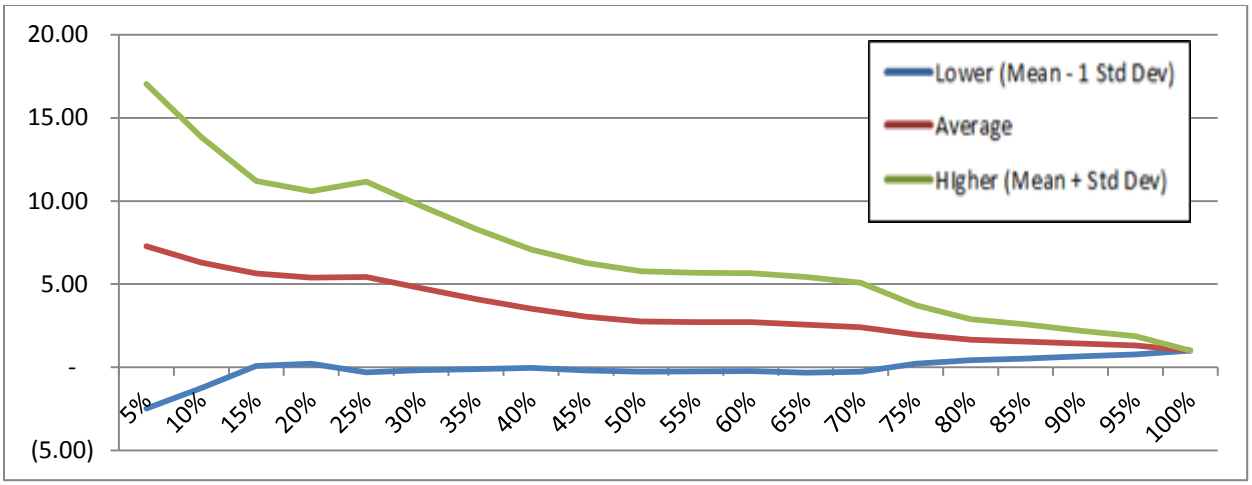

Figure C.5 – KPI 7; Normalized Cost per Hour for Small Tools & Consumables

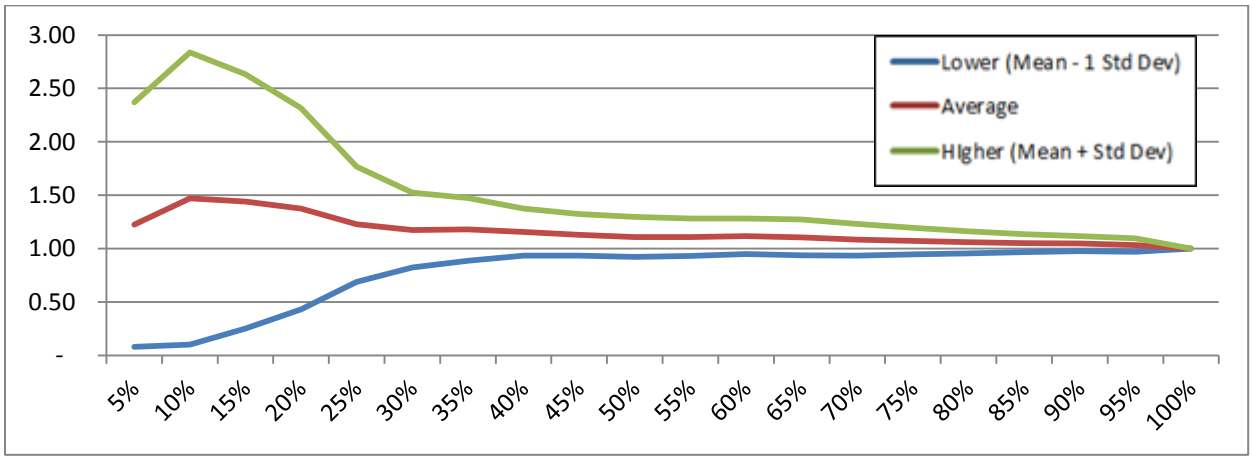

Figure C.6 – KPI 8; Normalized Cost per Hour for Travel

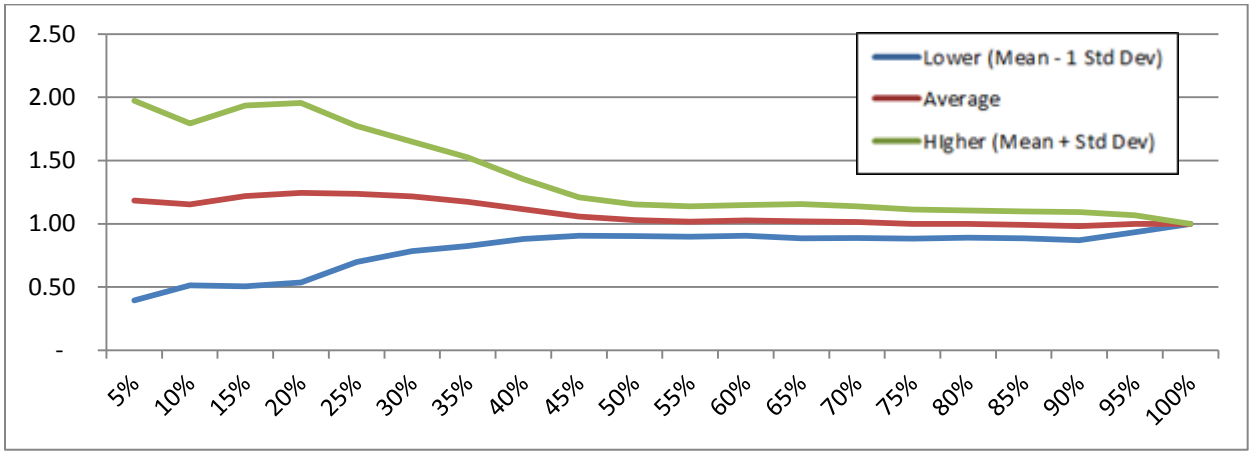

Figure C.7 – KPI 9; Normalized Cost per Hour for "Other"

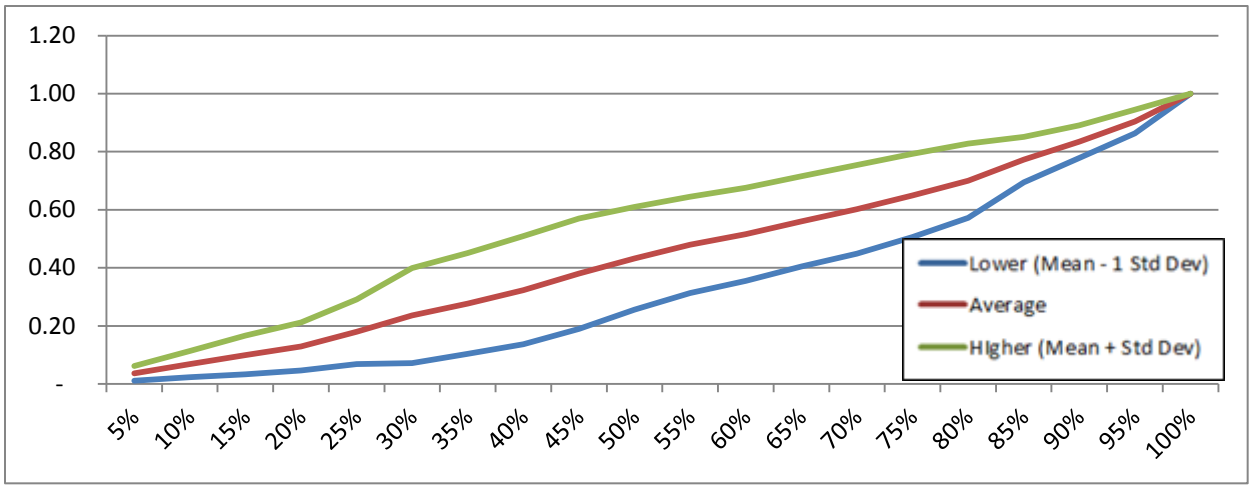

Figure C.8 –KPI 10; Normalized Percent of Direct Hours Spent

APPENDIX D – Confirming Normal Distribution of Historical KPI's

In order to confirm that Normal distributions were an appropriate distribution for modeling historical KPI's, two approaches were implemented.

The first method to confirm that historical KPI's were normally distributed was applying a Kolmogoroc – Smirnov (K-S) test. The K-S test compares differences between theoretical and observed cumulative probabilities. If the observed differences are greater than a theoretical critical difference (which is determined by the number of samples and desired confidence level), then the observed values do not match to the theoretical distribution. The Critical Value was retrieved from a table [4] for N=9 samples and a confidence of 95% and determined to be 0.432. In all cases differences between the observed and theoretical cumulative probabilities for each of the 10 KPI's, measured at 25%, 50% and 75% physically complete were determined to be less than the Critical Value (See Figure D.1). Therefore, the normal distribution appears to be acceptable. It should be noted that the small number of samples results in a high Critical Value and testing with additional samples is recommended to confirm this finding.

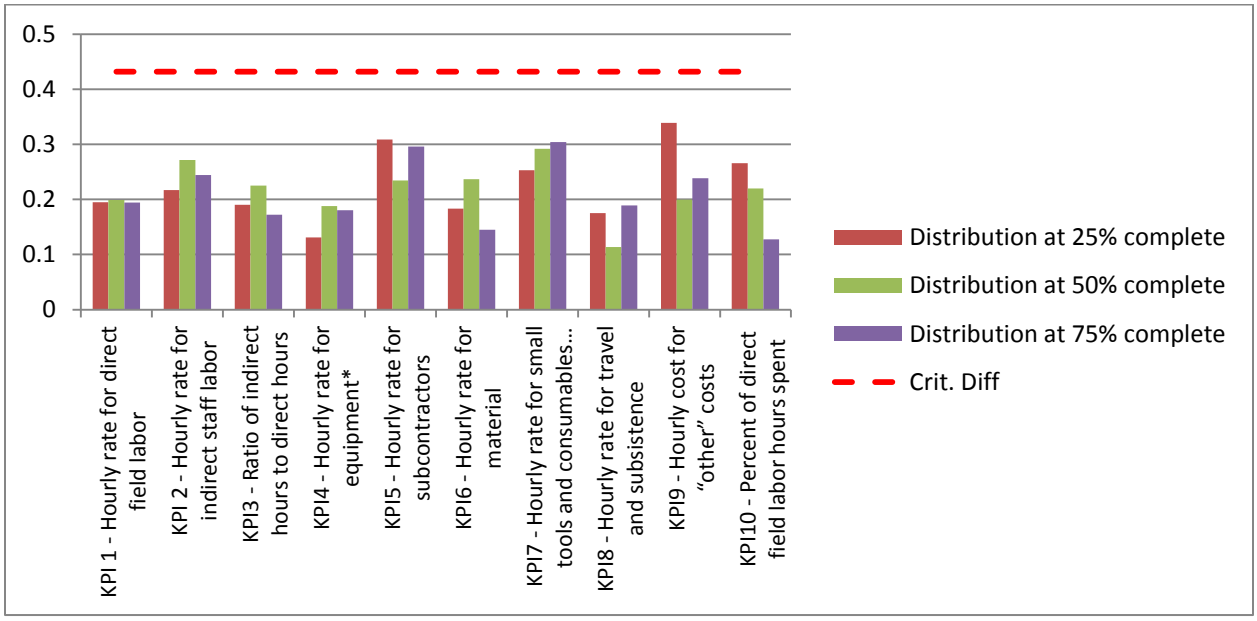

Figure D.1 - Kolmogorov–Smirnov (K-S) Test Scores

The second method to confirm that historical KPI's were normally distributed was a visual representation of the data known as a Quantile-Quantile (or Q-Q) plot. Q-Q plots show the *theoretical* value of each of the N samples on x-axis and the *observed* value of each of the N

samples on the y-axis. Both are converted to normalized standard distributions for convenience. For example, if the 3rd of 9 samples (cumulative probability of 28%) would be expected to occur at a Z-Score of -0.5896 based on a normal distribution curve (mean=0, standard deviation = 1) and the 3rd sample is *actually observed* at -0.7662, this would result as a point on the Q:Q plot of (-0.5896, -0.7662). This procedure is repeated for the other 8 sample points, resulting in the Q:Q plot for this particular historic KPI, at a particular percent complete for the job.

A perfect match between theoretical and observed values would result in a straight line with a slope of 1. Any variance from this highlights potential differences between theoretical values (based on an assumed normal distribution) and observed values.

Q-Q plots were completed for each KPI at intervals of 25, 50 and 75% complete and are shown below in Figures D.2 to D.11. The trendlines resulting from each KPI and percent complete combination are largely linear with slopes close to, but mostly less than, 1.0. This indicates a reasonable match to a normal distribution, however, more samples would be helpful to reinforce this assessment.

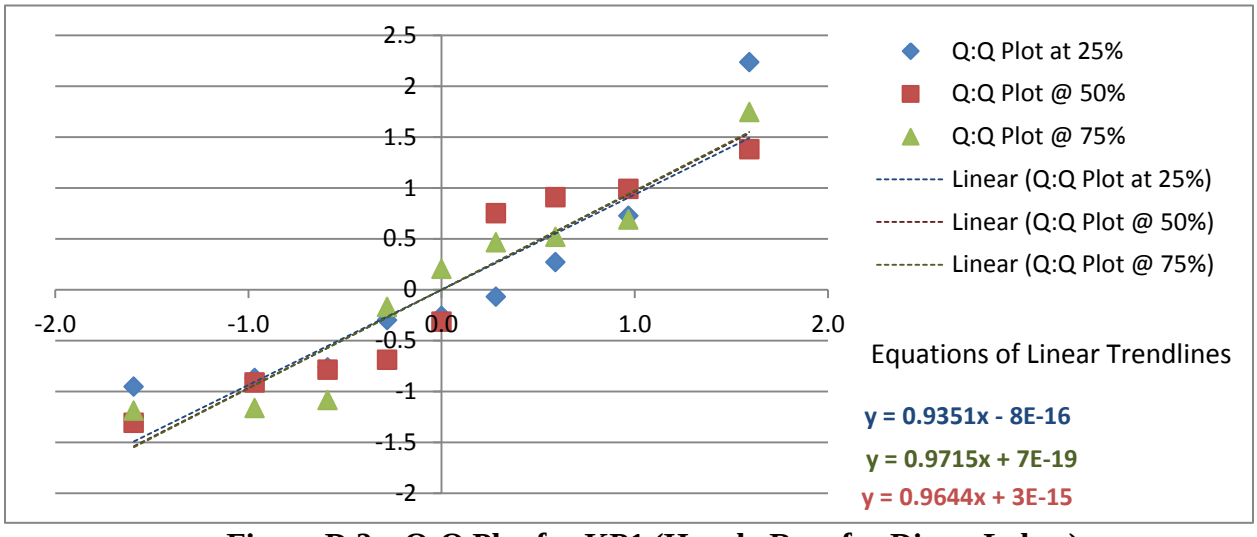

Figure D.2 – Q:Q Plot for KP1 (Hourly Rate for Direct Labor)

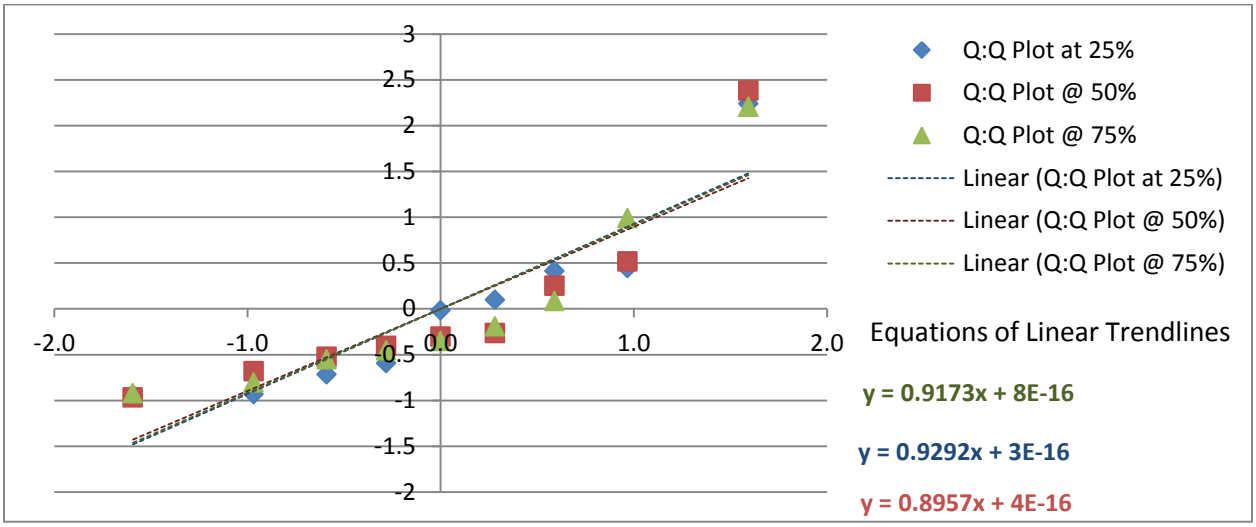

Figure D.3 – Q:Q Plot for KP2 (Hourly Rate for Indirect Labor)

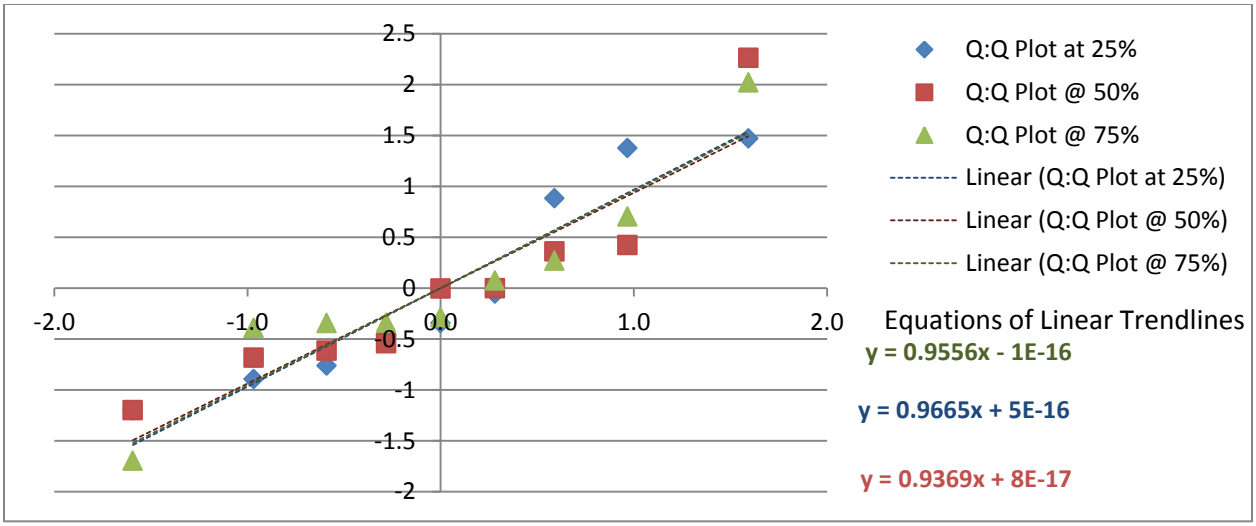

Figure D.4 – Q:Q Plot for KP3 (Ratio of Indirect to Direct Hours)

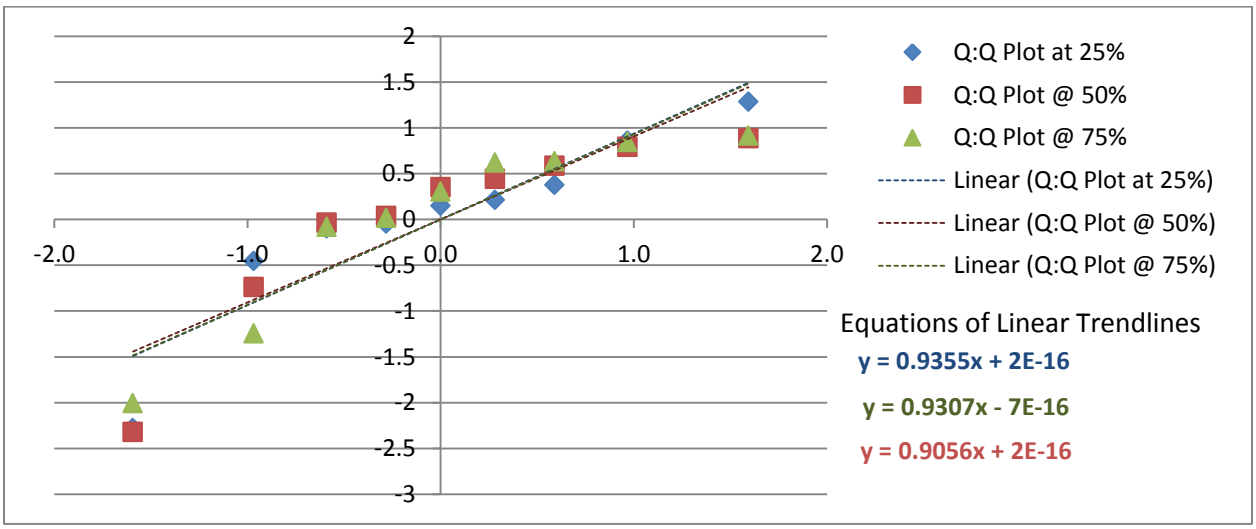

Figure D.5 – Q:Q Plot for KP4 (Hourly Rate for Equipment)

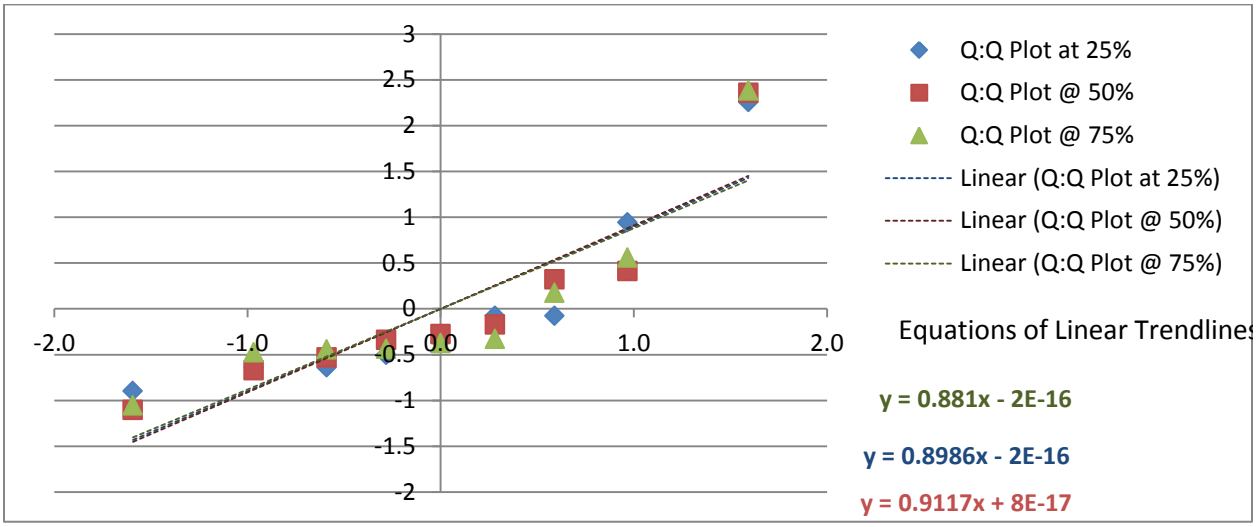

Figure D.6 – Q:Q Plot for KP5 (Hourly Rate for Subcontracts)

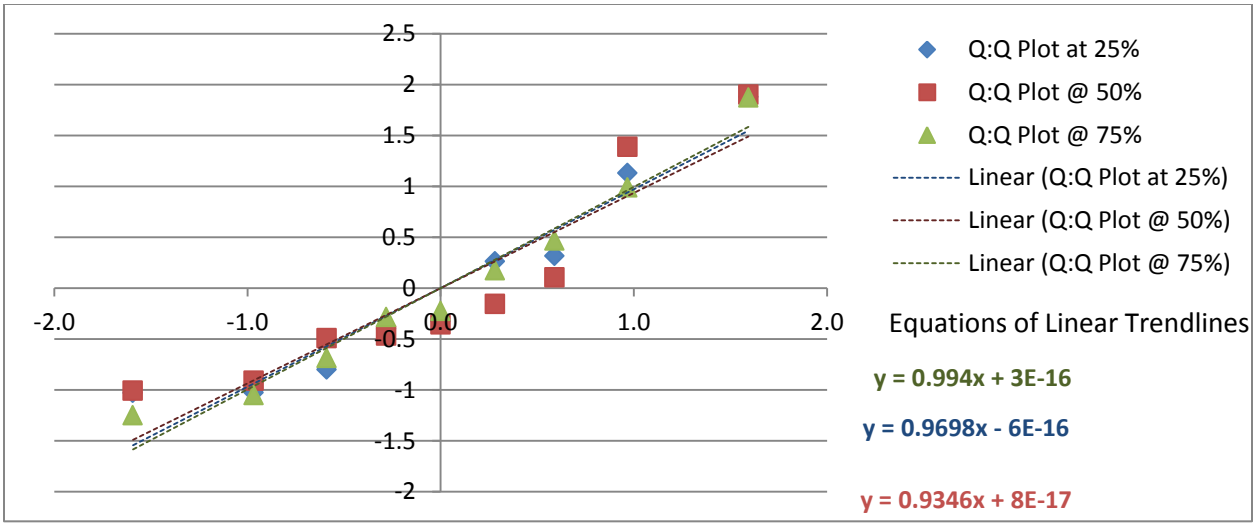

Figure D.7 – Q:Q Graph for KP6 (Hourly Rate for Material)

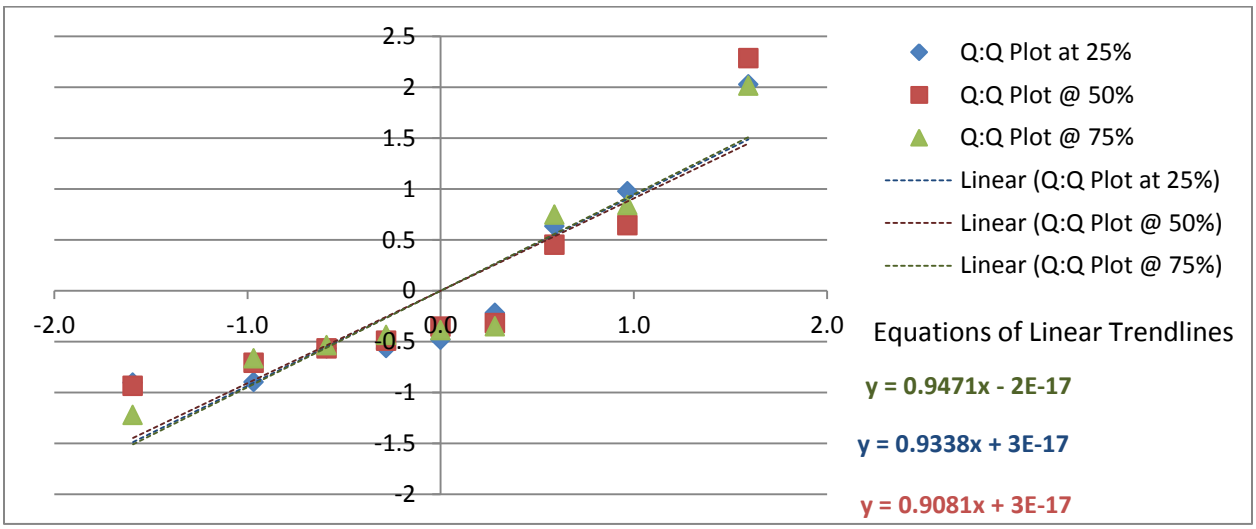

Figure D.8 – Q:Q Plot for KP7 (Hourly Rate for Small Tools & Consumables)

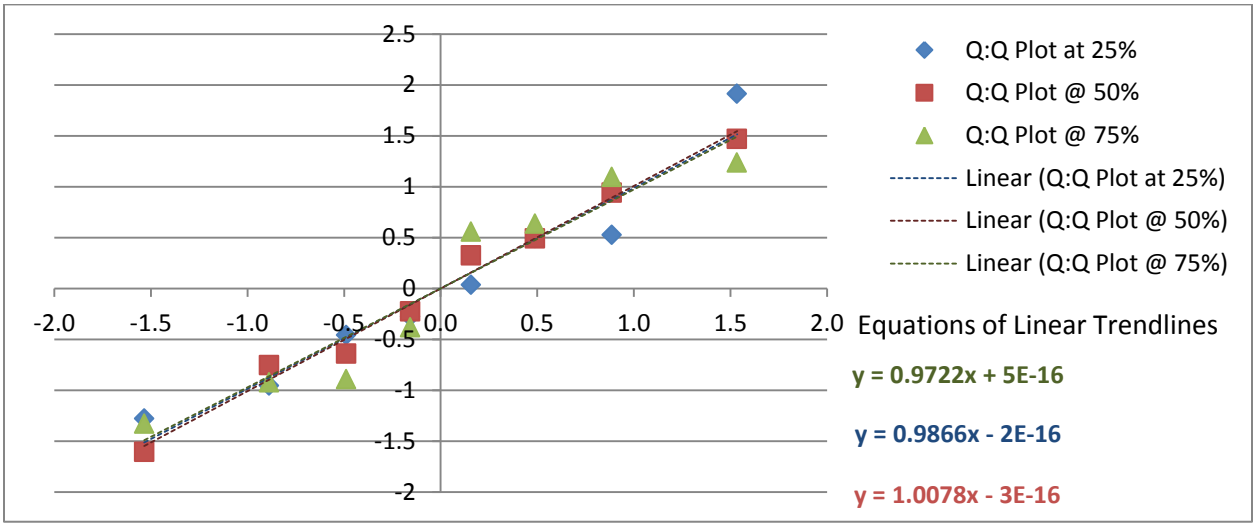

Figure D.9 – Q:Q Plot for KP8 (Hourly Rate for Travel & Subsistence)

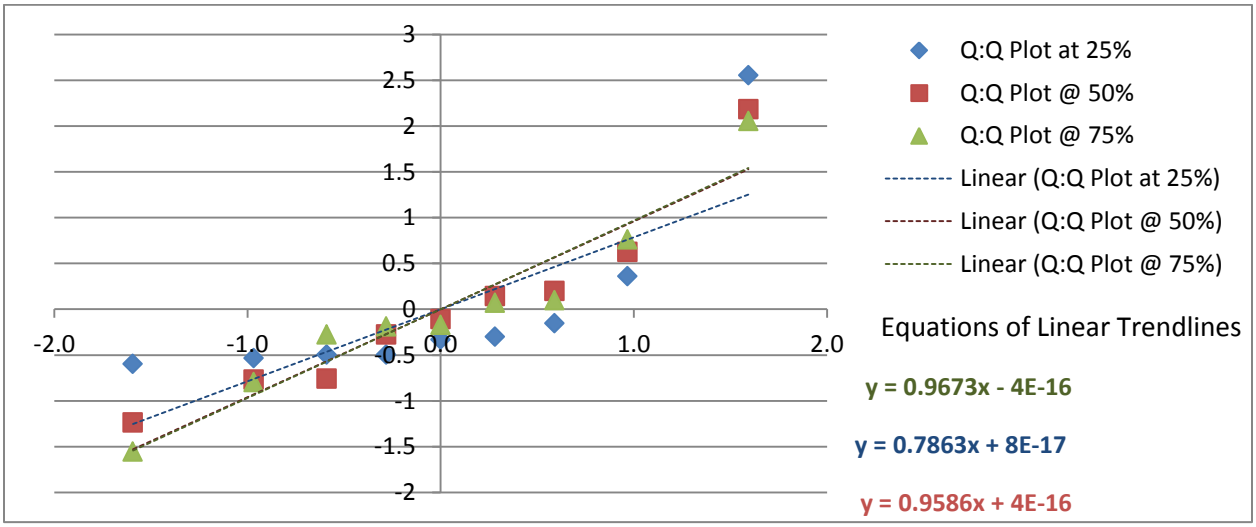

Figure D.10 – Q:Q Plot for KP9 (Hourly Rate for "Other" costs)

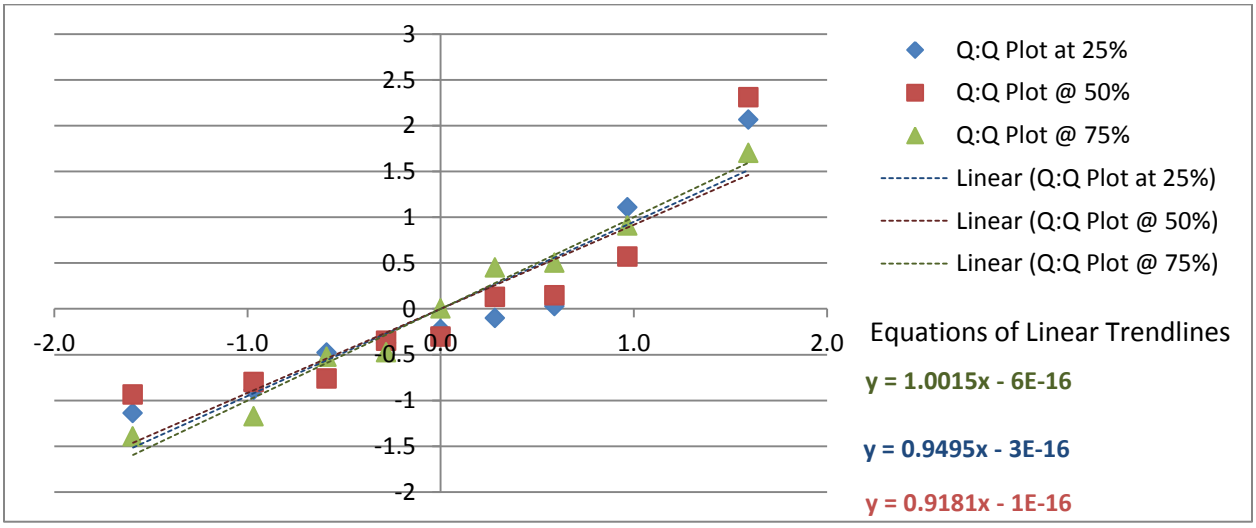

Figure D.11 – Q:Q Plot for KP10 (Percent of Direct Hours Spent)