Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 196 (2022) 808–815

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

Multiple Linear Regression Model for Improved Project Cost Forecasting

Filippo Maria Ottaviani[a]*, Alberto De Marco[a]

[a]*Politecnico di Torino, Corso Castelfidardo 39, Torino 10129, Italy*

Abstract

Several studies have been conducted in the Project Management field further to improve the Earned Value Management (EVM) methodology to forecast the project cost estimate at completion (EAC). This work aims at developing a linear model to increase the accuracy of the standard EAC and minimize the variance of the error. The research is conducted on an EVM data set comprising 29 real-life projects for a total of 805 observations. Multiple linear regression analysis is performed to evaluate the number of regressors, the priority of the candidate EVM variables into the regression model, and to assess the diagnostics of the model fit. The new EAC formulation is benchmarked, the results show the model to provide higher accuracy and lower variance compared to the standard formulation.

Peer-review under responsibility of the scientific committee of the CENTERIS –International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021

Keywords: Earned Value Management; Estimate at Completion; Linear Regression.

* Corresponding author. Tel.: +39-333-546-5773.
E-mail address: filippo.ottaviani@polito.it

Peer-review under responsibility of the scientific committee of the CENTERIS –International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021
10.1016/j.procs.2021.12.079

1. Introduction

Accurately estimating the final cost of a project is essential to its success since it allows managers to shape their strategies and actions to address the occurring risks. The most common solution for this purpose is represented by adopting the Earned Value Management (EVM) methodology: an objective framework analyzing the project performance and evaluating its cost and time estimates at completion (EAC). The approach is based on the project costs - the Planned Value (PV), Earned Value (EV), and Actual Cost (AC) - which are used as a proxy to track the activities' progress and performance, therefore the EVM can be applied to any project.

Data collected is used as input to the project cost evaluation, consisting of the total expected cost at completion of the defined scope of work. During the project execution, the EAC is obtained by adding the AC incurred up to that point to the remaining expenses - the Estimate to Complete (ETC). The latter can be calculated using two different methods: the analytical approach consists of the bottom-up sum of the cost of the remaining activities, the statistical approach exploits the EVM indexes to obtain an objective estimate of future costs, assuming no further exogenous interventions. The formulas adopted are relatively easy to use but show inefficiencies as they only account for the actual works situation and assume the linearity of the project labor profile. This leads to the neglection of the indexes' trends over time and the assumption that no further shifts in performance or risk will occur in the future.

Several attempts have been made to solve such limitations and improve the EVM methodology by developing different EAC models or regression analyses. In most cases, the methods were tailored to the specific project context, and their performance was either tested on a small number of projects or differed significantly from case to case. Studies mainly focused on improving the accuracy of estimates at the expense of statistical implications or the deployment of such models in an industrial context.

This work explores the nature of the relationships between the variables associated with the EVM data to find the best model to predict the future value of the target variable EAC. For this reason, both EVM features and multiple linear regression analysis have been used to assess an exploratory model and evaluate its performance. The work explores the trade-offs between accuracy, variance, and difficulty in application.

The paper is structured as follows. In Section 1, the topics of EVM and EAC are introduced. Section 2 consists of an overview of the attempts to extend the traditional EVM framework. The research methodology is described in Section 3, while Section 4 contains the discussion of the benchmarking. In Section 5, conclusions are drawn, and an outlook for future work is provided.

2. Literature Review

The spectrum of studies dealing with EVM methodology is broad and consists of two main branches: the deterministic, index-based methods and the probabilistic, statistics-based ones.

As for the first category, the focus of the various attempts is to best dimension the performance factor used to compute the project ETC. [1] proved the accuracy of the EAC formulas, i.e., the PF, depends on several factors such as the system to be developed, the stage and phase of the projects. [2] compared five different EAC methods in terms of statistical accuracy and stability, showing that the most accurate ones assume the cost deviations to continue at a constant rate. The results also demonstrated that the closer a project is to completion, the smaller the magnitude of the forecasting error is, hence the need to weigh the EAC output according to the project stage. [3] also examined five different EAC formulas, differing in terms of PF, proving that each one is highly dependent on the characteristics of the project. In particular, the CPI-only based one could provide a lower bound to the EAC calculations. [4,5] both described the EVM limitations linked to use costs as a proxy due to the financial aspects failing to grasp most physical dynamics, such as mobilization of resource phase and lack of consideration of indirect costs.

Statistics-based methods were developed to solve the issues mentioned above. [6,7] applied the progress-based stochastic S-curve profile with the cost and time variances, considered in probabilistic terms rather than exact values, which proved to yield more accurate forecasts when used in high-risk or non-linear labor profile projects. [8] proposed a parameterized S-curve tool, obtained by deriving a modified logistics differential equation, for managing the cost of an ongoing project after being subject to adjustment to fit the PM conditions. Instead, [9,10] tested the Gompertz growth model performance and demonstrated its superiority over the traditional method when applied

during the early, middle and late project stages. Such methodology was further expanded by [11] and tested on 25-real life projects data, showing minor improvements for the curve fitting process. [12,13] tested the non-linear methodologies based on the Earned Schedule and Duration principles, showing better performance, on average, than traditional index-based formulae, especially in the early stages of project development when the practical benefits are most significant for project teams to take their corrective actions. [14] proposed the PNR labor profile as an alternative to the traditional EAC formulation and proved it to converge faster, together with less variability than standard Estimate-at-Completion (EAC) calculations. [15] constructed an evolutionary EAC model to generate project cost estimates that proved significantly more reliable than estimates achievable using currently prevailing formulae. [16] compared the performance of Support Vector Regression model with the standard EVM one showing the superiority of the former method. [17] introduced two additional multivariate control metrics that allow the dynamic monitoring of the project, suggesting the need for additional metrics to describe the EVM/ES management system correctly.

An initial attempt to adjust the EAC was made by [18]: the study compared nine types of regression models and identified the one which best suited a cost prediction model for reconstruction project. [19] underlined the need to adopt moving weights of cost categories in a project budget: such variables depend on the progress of works and time lags, by the contractual payment conditions and credit times. [20] developed a probabilistic model to predict the risk effects that used real historical data to estimate project cost and duration, such as the project budgeted cost and planned project duration. [21,22] provided a different approach to the EAC problem, developing a model to predict the specific AC and EV by proposing a simple formula of data transformation. [23] developed a framework for adaptively combining forecasts of project costs from the inside view and the outside view using Bayesian inference and the Bayesian model averaging technique. [24] also developed a Bayesian model to calculate the confidence intervals for the estimates of the EAC based on the integration of quantitative and qualitative data.

3. Research methodology

3.1. Data

The study was conducted on the EVM data related to 29 different real-world projects, selected from the dataset developed by [25,26]. No data cleaning was required. The variables that have been used are summarized as follows. *rEAC* stands for the project cost at completion normalized by the BAC while *fEAC* is the forecast that is computed through Eq.1.

$$fEAC = [AC + (1 - WP)/CPI]/BAC \tag{1}$$

TP stands for the normalized time period and consists of the ratio between the time unit at which the project review is carried out and the project planned duration. *WS* and *WP* stand, respectively, for the percentage of work scheduled and performed in time *TP*. *AC* stands for the actual costs in *TP* normalized by BAC, its values range from 0 to *rEAC*. *SPI* and *CPI* represent the schedule and cost performance indexes.

3.2. Model evaluation

All of the following procedures were realized in the SAS® Studio environment. The multiple linear model assumption was driven by the need to limit the application difficulty: unlike the more sophisticated deep learning algorithms, which focus on the time required and exact accuracy of the predictions, the white-box model allows for the introspection of its form, processes, and causal effects between the input variables.

The model development was divided into three steps. First, the generalized linear model selection procedure (GLMSELECT) was performed. Its output diagnostics were analyzed to evaluate the impact of the variables' sequential implementation on the fit criteria. Second, the correlations between the variables was assessed. The selected regressors were subjected to the linear regression analysis (REG) to compute the relative parameters. Because of the explicit correlations among the fundamental and derived quantities provided by the EVM data, both the possible underfitting and overfitting phenomena had to be limited. The first issue was solved by forcing the

regression model to implement the variable *fEAC* since it already provides a reliable basis upon which to base the predictions. The overfitting issue was addressed by combining the least absolute shrinkage and selection operator (LASSO) method with the *k*-fold cross-validation measure. The former consists of the criterion used to address the multicollinearity issue present in the EVM data through the regularization of the parameters; the latter is intended to solve the issue related to the low number of projects at disposition.

3.2.1. Generalized linear model selection procedure

Table 1 describes the evolution of the fit statistics during the LASSO search procedure. Overall, all the criteria improve as more variables enter the model except for the SBC (Schwarz information criterion) and CV PRESS (cross-validation predicted residual sum of squares) statistics which reach their optimum at Step 6, hence eliminating *SPI* from the set of proposed regressors. The criteria already converge during the earlier steps as their marginal improvements tend to decrease and get negligible. In addition to this, the *F* value for Step 4 is relatively higher than the previous and successive steps. Although the null hypothesis cannot be rejected as its *p*-value is inferior to the α level of .05, such issue provides a hint at the possible shift in the model variance the introduction of a fourth regressor would lead to. For this reason, the number of regressors is limited to three.

Table 1. Summary of the generalized linear model selection procedure

Step	Effect Entered	Model R-Square	Adjusted R-Square	AIC	AICC	BIC	CP	SBC	ASE	CV PRESS	*F* Value	Pr > *F*
1	*fEAC*	.4302	.4295	660.8	660.8	-147.1	35383.2	-140.5	0.832	60.7	606.3	<.0001
2	*CPI*	.9837	.9836	-2192.4	-2192.4	-2997.4	238.0	-2989.1	0.023	19.4	27156.3	<.0001
3	*WS*	.9841	.9840	-2211.2	-2211.1	-3016.5	213.6	-3003.1	0.023	18.7	20.9	<.0001
4	*AC*	.9843	.9842	-2220.7	-2220.6	-3026.3	201.2	-3007.9	0.022	17.0	11.5	0.0007
5	*WP*	.9864	.9863	-2335.5	-2335.4	-3140.2	68.4	-3118.1	0.019	15.3	124.9	<.0001
6	*TP*	.9874	.9873	-2392.0	-2391.8	-3195.9	9.9	-3169.8	0.018	15.1	60.2	<.0001
7	*SPI*	.9875	.9873	-2394.9	-2394.7	-3198.7	7.0	-3168.1	0.018	15.1	4.9	0.0273

3.2.2. Correlation analysis

The selected regressors to be implemented into the model are *fEAC*, *CPI* and only one variable between *WS*, *AC* and *WP*. The Pearson ρ coefficients and respective *p*-values are reported in Table 2. For both *fEAC* and *CPI*, all the related ρ coefficients are statistically significant. The strongest correlation that subsists between the former and the other variables is the one with *CPI* (ρ=-.78040). Overall, the least correlated variable is *WP* as it is both weakly correlated to *fEAC* (ρ=-.12127) and *CPI* (ρ=-.11756). Therefore, *WP* is selected as the third regressor to enter the model. The correlations between the selected variables do not pose a threat to the model performance.

Table 2. Pearson correlation table

	rEAC	*WS*	*WP*	*AC*	*CPI*	*fEAC*
rEAC	1.00000					
WS	-0.07965 0.0238	1.00000				
WP	-0.09851 0.0052	0.95157 <.0001	1.00000			
AC	0.07611 0.0308	0.91269 <.0001	0.94794 <.0001	1.00000		
CPI	0.10906 0.0019	-0.12275 0.0005	-0.11756 0.0008	-0.27042 <.0001	1.00000	
fEAC	0.32584 <.0001	0.16061 <.0001	0.12127 0.0006	0.35776 <.0001	-0.78040 <.0001	1.00000

3.2.3. Multiple linear regression analysis

The definitive model to be analysed is described by Eq. 2.

$$\widehat{rEAC} = \beta_0 fEAC + \beta_1 CPI + \beta_2 WP + \varepsilon \quad (2)$$

The ANOVA results are shown in Tables 3 and 4. The F statistic for the overall model is highly significant (F=16674.4, p<.0001). The Adjusted R-Square is very close to unity (Adj R-Sq=.9842), as expected from the implementation of the *fEAC* variable. According to a rule of thumb, the Coefficient of Variation can be considered moderate (Coeff Var ~ 12.76664 ≤ 15).

Table 3. Analysis of variance – model, error and uncorrected total related measures

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Model	3	1158.04974	386.01658	16674.4	<.0001
Error	802	18.56656	0.02315		
Uncorrected Total	805	1176.61630			

Table 4. Analysis of variance – model criteria

Criterion	Value
Root MSE	0.15215
Coeff Var	12.76664
Adj R-Sq	0.9842

Table 5 contains multiple statistics related to the model regressors. The t statistics and the corresponding p-values confirm all the parameters to be significant. Concerning the multicollinearity issue, no Tolerance statistic falling below .10 and VIF statistic being greater than 10 (or close to 5, according to which rule of thumb is adopted) confirm the threat posed by the variance to the model accuracy to be low, yet not negligible.

Table 5. Parameter estimates table

Variable	Label	DF	Parameter Estimate	Standard Error	*t* Value	Pr > |*t*|	Tolerance	Variance Inflation	95% Confidence Limits	
fEAC	*fEAC*	1	0.70612	0.01103	64.00	<.0001	0.18296	5.46579	0.68446	0.72778
CPI	*CPI*	1	0.45998	0.01080	42.59	<.0001	0.23245	4.30209	0.43878	0.48118
WP	*WP*	1	-0.07931	0.01575	-5.03	<.0001	0.32904	3.03918	-0.11023	-0.04838

In line with the considerations made for Table 2, the results in Table 6 suggest a relationship subsisting between the variables *fEAC* and *CPI* as the Proportion of Variation value for both regressors in row 3 exceeds the reference value of .5. Although the Condition Index does not exceed 10 (according to a rule of thumb), the multicollinearity issue does not affect the model performance.

Table 6. Collinearity diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation		
			fEAC	*CPI*	*WP*
1	2.62895	1.00000	0.02429	0.02951	0.03969
2	0.25305	3.22319	0.02933	0.29876	0.82461
3	0.11799	4.72019	0.94638	0.67173	0.13569

4. Results

According to the results in Table 5, the fitted model is described by Eq. 3.

$$\widehat{rEAC} = .70612 fEAC + .45998 CPI - 0.07931 WP \quad (3)$$

The MAPE and Standard Deviation statistics of *fEAC* and Eq.3 have been calculated to benchmark the new formula. The results are summarized in Table 7: the difference between the two MAPEs is limited while the Standard Deviation of the fitted model is considerably lower (9 percentage points circa).

Table 7. Benchmarking of the fitted model performance

Model	MAPE	Standard Deviation
fEAC	.1576	.1990
Eq. 3	.1391	.1102

The adjustment made to *fEAC* through Eq. 3 succeeds in improving the EAC forecasting accuracy and reducing the error variance. Instead, the residuals analysis was carried out to evaluate the fitted model bias.

Fig.1 shows the quality of the predictive model by comparing, for each observation, the real vs the predicted response. The magnitude of the errors is heavily influenced by the *fEAC* estimations being carried out during the early stages of the projects, during which the quality and reliability of the EVM data are limited. A remark could be made by comparing the cases in which the model overestimates or underestimates the target variable, respectively. In the first case, the average error is far more limited, with the maximum negative residual being close to .8. In the latter case, the residuals are less in number but greater in value.

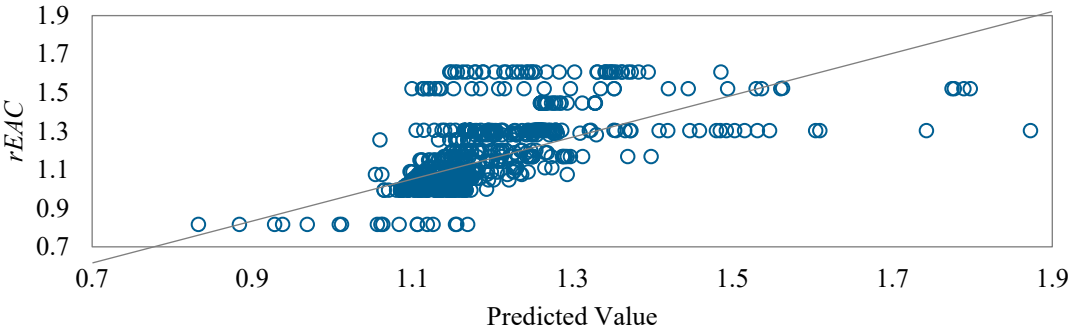

Figure 1 Observed by predicted value plot: model bias due to projects settings

Fig.2 shows the plots of the residuals against the three regressors *fEAC*, *CPI* and *WP*. Concerning Fig.2a, the same considerations from Fig.1 apply: when *fEAC* is inferior to 1, the residuals tend to be negative; the inverse applies when *fEAC* is greater than 1. Fig.2b shows the residuals trend to switch from increasing, right after the *CPI* value of 1, to decreasing at $CPI \sim 1.25$: the reason behind such behaviour could be represented by the temporary shift of the *CPI* that is only temporary and does not remain constant overtime. The last plot in Fig.2c shows the residuals converging to 0 as $WP \rightarrow 1$: this is confirmed by the real project dynamics by which the more work is performed, the closer the project is to its completion therefore the less risk is likely to occur, and less the predictions are subject to deviate from the actual values.

To conclude, the fitted model is biased, but the reasons behind that are not due to the methodology. First, the model tends to overestimate the *rEAC* even when the project is on budget. This either proves that most projects experience cost overruns, or the EVM data set is biased itself. Second, the excessive shifts of the *CPI* from the value of 1 lead to an excessive inflation of the *rEAC*, due to both the *CPI* and *fEAC* regressors. Although, the negative

value of the parameter related to *WP* is such to reduce the bias by shifting the *rEAC* value downwards, hence reducing the inflation.

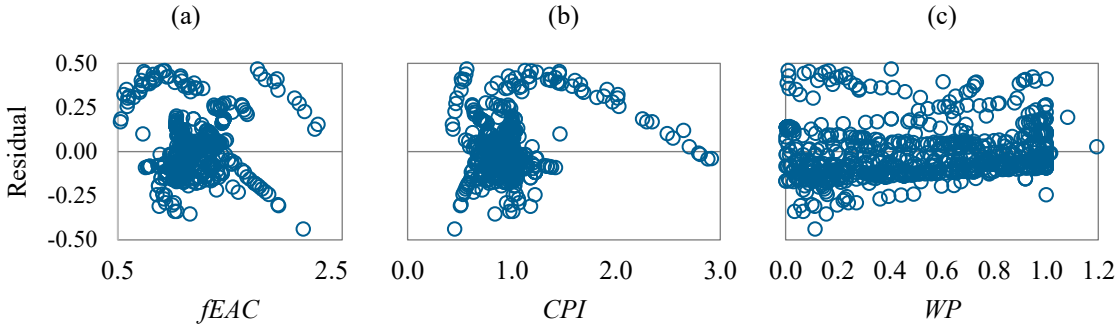

Figure 2 (a) Residual vs *fEAC*; (b) Residual vs *CPI*; (c) Residual vs *WP* plots

5. Conclusions

The objective of this study was to develop a model to refine the project cost estimate at completion. The work was subject to the constraint by which the improvement of the prediction accuracy should not compromise the error variance. For this reason, the regression analysis was chosen to adjust the original EAC output.

The linear model was computed through a 3-steps statistical analysis during which the regressors were first chosen, the absence of correlation among them was verified, and the model properties were analyzed; its output is represented by Eq. 3. The first regressor, *fEAC*, consists of the forecast to be adjusted, and its implementation prevents the underfitting phenomenon. The second regressor, *CPI*, suggests the cost performance index may not be used only to forecast the remaining work, as in the *fEAC* formulation, but the whole project costs. Both *fEAC* and *CPI* coefficients being positive means a weighted sum of the two variables must be made. The third regressor, *WP*, shows a negative coefficient: the closer the project gets to its completion, the more the *fEAC* and *CPI* contributions to the output are reduced.

The model performance was benchmarked against the index-based EAC method in terms of variance and accuracy. The results show a notable improvement over the standard forecasting method, especially for the Standard Deviation. This confirms the need to adjust the simple EAC computation as it alone would fail to capture the correlation between the EVM variables' evolution over time. It is also confirmed the need for additional EVM variables to account for the trends in the cost and time performance indexes.

Like any other algorithm, the model performance is highly dependant on the reliability and availability of EVM data the analyses are based upon. For this reason, the estimate of the three parameters may be biased. Such an issue could either be solved by processing more real-life project data or introducing additional constraints to the optimization procedure used to evaluate the regressors' parameters.

Furthermore, instead of the linear regression analysis which was chosen because of its easy applicability in managerial and industrial contexts, future research may explore more sophisticated models such as non-linear regression and machine-learning algorithms, correctly tailored for the PM domain of application.

References

[1] Christensen, David (1993) "The Estimate at Completion Problem: A Review of Three Studies." *Project Management Journal*. **24** (**1**): 37–42.

[2] Zwikael, Ofer and Globerson, Shlomo (2000) Evaluation of Models for Forecasting the Final Cost of a Project.

[3] Lipke, Walt (2004) "Independent estimates at completion - Another method." *CrossTalk*. (**10**): 26–30.

[4] Cândido, Luis F., Paula, José, and Neto, Barros (2014) Critical analysis on earned value management (EVM) technique in building construction.

[5] Fleming, Quentin W. and Koppelman, Joel M. (1997) "Earned value project management." *Cost Engineering (Morgantown, West Virginia)*. **39** (**2**): 13–15.
[6] Barraza, Gabriel A., Back, W. Edward, and Mata, Fernando (2000) "Probabilistic Monitoring of Project Performance Using SS-Curves." *Journal of Construction Engineering and Management*. **126** (**2**): 142–148.
[7] Barraza, Gabriel A., Back, W. Edward, and Mata, Fernando (2004) "Probabilistic Forecasting of Project Performance Using Stochastic S Curves." *Journal of Construction Engineering and Management*. **130** (**1**).
[8] Cioffi, Denis F. (2005) "A tool for managing projects: An analytic parameterization of the S-curve." *International Journal of Project Management*. **23** (**3**): 215–222.
[9] Narbaev, Timur and De Marco, Alberto (2014) "An Earned Schedule-based regression model to improve cost estimate at completion." *International Journal of Project Management*. **32** (**6**): 1007–1018.
[10] Narbaev, Timur and De Marco, Alberto (2014) "Combination of Growth Model and Earned Schedule to Forecast Project Cost at Completion." *Journal of Construction Engineering and Management*. **140** (**1**): 04013038.
[11] Huynh, Quyet Thang, Le, The Anh, Nguyen, Thanh Hung, Nguyen, Nhat Hai, and Nguyen, Duc Hieu (2020) A Method for Improvement the Parameter Estimation of Non-linear Regression in Growth Model to Predict Project Cost at Completion. in: Proceedings - 2020 RIVF International Conference on Computing and Communication Technologies, RIVF 2020, Institute of Electrical and Electronics Engineers Inc.
[12] Warburton, Roger D.H., De Marco, Alberto, and Sciuto, Francesco (2017) "Earned schedule formulation using non-linear cost estimates at completion." *Journal of Modern Project Management*. **5** (**1**): 75–81.
[13] Warburton, Roger D.H. and Cioffi, Denis F. (2016) "Estimating a project's earned and final duration." *International Journal of Project Management*. **34** (**8**): 1493–1504.
[14] Warburton, Roger D.H. (2011) "A time-dependent earned value model for software projects." *International Journal of Project Management*. **29** (**8**): 1082–1090.
[15] Cheng, Min Yuan, Peng, Hsien Sheng, Wu, Yu Wei, and Chen, Te Lin (2010) "Estimate at completion for construction projects using evolutionary support vector machine inference model." *Automation in Construction*. **19** (**5**): 619–629.
[16] Wauters, Mathieu and Vanhoucke, Mario (2014) "Support Vector Machine Regression for project control forecasting." *Automation in Construction*. **47**: 92–106.
[17] Colin, Jeroen, Martens, Annelies, Vanhoucke, Mario, and Wauters, Mathieu (2015) "A multivariate approach for top-down project control using earned value management." *Decision Support Systems*. **79**: 65–76.
[18] Chen, Wei Tong and Huang, Ying Hua (2006) "Approximately predicting the cost and duration of school reconstruction projects in Taiwan." *Construction Management and Economics*. **24** (**12**): 1231–1239.
[19] Park, Hyung K., Han, Seung H., and Russell, Jeffrey S. (2005) "Cash Flow Forecasting Model for General Contractors Using Moving Weights of Cost Categories." *Journal of Management in Engineering*. **21** (**4**): 164–172.
[20] Abu Hammad, Ayman A., Ali, Souma M. Alhaj, Sweis, Ghaleb J., and Sweis, Rateb J. (2010) "Statistical Analysis on the Cost and Duration of Public Building Projects." *Journal of Management in Engineering*. **26** (**2**): 105–112.
[21] Chen, Hong Long (2014) "Improving Forecasting Accuracy of Project Earned Value Metrics: Linear Modeling Approach." *Journal of Management in Engineering*. **30** (**2**): 135–145.
[22] Chen, Hong Long, Chen, Wei Tong, and Lin, Ying Lien (2016) "Earned value project management: Improving the predictive power of planned value." *International Journal of Project Management*. **34** (**1**): 22–29.
[23] Kim, Byung-Cheol and Reinschmidt, Kenneth F. (2011) "Combination of Project Cost Forecasts in Earned Value Management." *Journal of Construction Engineering and Management*. **137** (**11**): 958–966.
[24] Caron, Franco, Ruggeri, Fabrizio, and Pierini, Beatrice (2016) "A Bayesian approach to improving estimate to complete." *International Journal of Project Management*. **34** (**8**): 1687–1702.
[25] Batselier, Jordy and Vanhoucke, Mario (2015) "Construction and evaluation framework for a real-life project database." *International Journal of Project Management*. **33** (**3**): 697–710.
[26] Vanhoucke, Mario, Coelho, José, and Batselier, Jordy (2016) "An Overview of Project Data for Integrated Project Management and Control." *The Journal of Modern Project Management*. **3** (**3**).