Forecasting the Performance Measurement for Iraqi Oil Projects using Multiple Linear Regression

Nidal Adnan Jasim a**, Abdulrahman Adnan Ibrahim b, Wadhah Amer Hatem c

a Department of Civil Engineering, College of Engineering, University of Diyala, Diyala, Iraq.
b Department of Civil Engineering, College of Engineering, Tikrit University, Tikrit, Iraq.
c Middle Technical University, Baghdad, Iraq.

Keywords: Earned Value; Multiple Linear Regression (MLR); Schedule Performance Index (SPI); Cost Performance Index (CPI); To-Complete Cost Performance Indicator (TCPI); Predicting; Oil Projects

Abstract: Many oil and gas projects have been subjected to significant cost overruns and schedule delays, which is a major concern for the decision-makers in the oil industry. This paper aims to develop three mathematical models to estimate earned value indicators, the Schedule Performance Index (SPI), Cost Performance Index (CPI), and To-Complete Cost Performance Indicator (TCPI), to reduce the cost and time estimation error in Iraqi oil projects. The research methodology adopted artificial intelligence techniques using Multiple Linear Regression technology (MLR) to predict Earned Value (EV) Indexes to get standard local equations to measure the performance of Iraqi oil projects. The data is based on (83) monthly reports from 26 June 2015 to 25 August 2022 collected from the Karbala Refinery Project, selected as a case study. It is one of the Oil Projects Company (SCOP)- the Iraqi Ministry of Oil’s massive and modern projects, and it combines several projects into one project. The results showed numerous significant points, such as the average accuracy (AA%) for the CPI, SPI, and TCPI was 95.194%, 92.195%, and 83.706%, respectively, while the correlation coefficients (R) were 92.4%, 98.4%, and 93.7%. It was shown that there were relatively few differences between the theoretical and actual results. Therefore, the MLR technique was utilized in this paper to derive the prediction models for its more correct earned value predictions.
1. INTRODUCTION

The construction of oil and gas projects is critical because they help to run and make oil and gas production easier. However, these projects often face long-term risks that cause them to take longer, be more costly, and have a lower quality, which hurts their chances of success [1]. Both the technology and management of the oil and gas industries are complicated, so oil and gas projects are considered the most difficult. In addition to experience, project managers should follow a coherent reference framework based on constant monitoring and review of all official project stages from the beginning to the end of the project. Time, cost, and quality strategies are required to achieve efficient management in the oil and gas industry. Ultimately, this will force the need for techniques to reduce the chances of future project failures [2, 3]. Performance measurement is an important part of any project because it gives a basis for improving performance over time. Because the construction industry is so competitive and technology is changing so quickly, construction executives must continuously improve how their projects work. Most people agree that a project's success can be judged by how well it acts in terms of cost, time, and quality [4, 5]. Forecasting project performance is one of the hardest things to do when determining whether the project will be successful. A construction project cannot be well done without challenges and problems. To meet and overcome these challenges, an organization must clearly know how well it is doing. Good performance can help a construction project succeed. Efficiency in construction means that the job is accomplished on time and within budget. So, a project can be considered a set of unique, complicated, and linked tasks with the same goal or purpose and must be finished by a certain date, within budget, and according to requirements [6, 7]. The earned value management system (EVMS) is a good technique for project managers to track and control projects. It combines a project’s work scope, schedule, and cost elements, making reporting on its progress and cost status easier. The earned value management system combines time and cost management, both essential for managing projects [8]. Iraq’s oil projects have yet to accept modern methods to estimate the earned value. Therefore, this paper aims to show how to predict Earned Value (EV) Indexes using Multiple Linear Regression technology (MLR) models in oil projects, specifically refineries. This makes it easier for practitioners to use EV for scheduling and budget control in oil projects. This approach aims to assist decision-makers, i.e., managers, engineers, contractors, and planners, in making more accurate and trustworthy decisions. Additionally, decision-makers can comprehensively understand how the project will perform in the future to avoid unintended deviations from the original design. The objective of the current work is to create three alternative mathematical models to acquire local standard equations. The steps listed below can be used to accomplish these goals:
• Determining the artificial intelligence (AI) technique variables that affect the EV indices in Iraqi oil projects
• Building mathematical models that can be used to estimate the Schedule Performance Index (SPI), Cost Performance Index (CPI), and To-Complete Cost Performance Indicator (TCPI) in Iraqi oil projects before the execution phases
• Formulating equations to calculate the SPI, CPI, and TCPI for oil projects
• Testing the effectiveness and precision of the outcomes in mathematical models by verifying and validating their generated mathematical models

2. LITERATURE REVIEW
According to the following literature review, Machine Learning Regression Techniques (MLRT) have been effectively utilized in construction project management. For example, in an international study, Ottaviani & Marco [9] developed a linear model to improve the methodology for forecasting the project cost estimate at completion (EAC). The results indicated an important improvement over the traditional forecasting method, particularly for the Standard Deviation. Also, the model offered more accuracy and less variance. As for the Iraqi studies, Alfaham & Al Ajeeli [10] created a predictive model utilizing machine Learning Regression Techniques (MLRT) to assess the construction project quality for government buildings. To improve the quality of government buildings and decrease maintenance costs. The MLR model performed very well ($R^2 = 86.35\%$). Nassar & Erzaij [11] adopted (MLRT) to create anticipatory models for construction project crises by identifying and classifying the key factors that influence project goals and signal time and expense overruns and poor quality for projects before crises happen. Three equations were created from the MLR multiple linear regression findings to determine the percentage of overrun, i.e., time, cost, and quality, due to the construction project being influenced by crises. The above models’ respective correlation coefficients were 99.8%, 98.6%, and 96.5%. Jaber et al. [12] build a prediction model for earned value indexes using Machine Learning Regression Techniques (MLR) for tall building projects. The MLRT produced good estimation results regarding the correlation coefficient ($R$) obtained by MLR models for SPI, CPI, and TCPI, with $R$ values of 85.5%, 89.2%, and 86.3%, respectively.

3. METHODOLOGY
The following steps were considered to design and assess EV models:
1- Choosing a suitable Software.
2- Identifying the models MLR variables that influence the EV index in Iraqi Oil projects
3- Building mathematical models to predict earned value indexes, i.e., CPI, SPI, TCPI
4- Verification and validation of mathematical models

Fig. 1 displays the development of the MLR models’ methodology.

4. CASE STUDY BACKGROUND
The Karbala Refinery project was selected as a case study to achieve the research goal. Karbala Refinery is one of the massive projects whose schedule and planning budget have been studied, followed up on its implementation professionally using advanced computer programs by the implementing agency. The implementing agency used the Primavera program to develop the detailed structure of the various project activities, schedule them, distribute the responsibilities and resources needed to determine the initial budget through its various stages, and start implementing the project and preparing the reports on its progress. The Project Location is 25 km South of Karbala City, Iraq (100 km South of Baghdad City). The original contract value for the Karbala refinery project was ($6,023,000,000), and the original contract was 54 months. Oil was pumped to the Karbala refinery for the first time on 25 September 2022. The Karbala strategic refinery operation commenced on 20 October 2022. Fig. 2 shows the Construction Site Layout of the Karbala Refinery project.
5. MOTIVES AND REASONS FOR CHOOSING THE CASE STUDY

1. The Karbala refinery project is one of the Oil Projects Company (SCOP), the Iraqi Ministry of Oil’s massive and modern projects, and it combines several projects into one project.

2. The French company Technip specialized in building refineries was contracted as a consultant to the Oil Projects Company in the management, follow-up, and control of all project activities (EPC) in May 2013. Also, a contract was signed with a consortium of Korean companies (HDGSK) for a joint project to build an integrated refinery with high-quality and environmentally friendly specifications. The Design, Procurement, and Build (EPC) contract was signed in April 2014. As a result, accurate follow-up reports and information on costs and schedules required to implement the earned value management methodology have been provided.

3. Lack or scarcity of local and international research on the topic of applying earned value management and its relationship with Artificial Intelligent for oil projects, especially refineries

4. Despite the difficulties encountered in obtaining information about the Karbala refinery, as it required obtaining many approvals, the good documentation of all project information, in addition to the fact that the project is in the process of completion and its final stages, was a great incentive for the selection of the project.

6. PREPARATION OF DATA

Eighty-three reports on the Karbala refinery project were obtained from the Karbala Refinery Project Authority, the State Company for Oil Projects (SCOP), Iraqi Ministry of Oil. Seventy-three reports were used for building the (MLR) models, and ten were used for generalization. For each of the three models, i.e., CPI, SPI, TCPI, the data was separated into three categories: training, testing, and validation. The CPI model got 78% of the data in the training set, 11% in the test set, and 11% in the validation set. As a result, 57 reports were used for training, eight for validation, and eight to test this model. While the SPI model received 70% of the data from the training set, the test set received 5%, and the validation set received 25%. Consequently, 51 reports were used for training, 18 for validation, and 4 for testing. As for the TCPI model, the optimal division was 84% for the training dataset, 5% for the testing dataset, and 11% for the validation datasets. Consequently, 61 reports were used for training, 8 for verification, and 4 for testing. The precision of all these divisions was based on the lowest testing errors and highest Correlation Coefficients (r) value.

7. CHOOSING A SUITABLE STATISTICS SOFTWARE FOR MLR MODELS

Several applications can be used in statistical analysis, including Microsoft Excel, STATISTICA, MINITAB, and MATLAB. As for the present study, the statistical for the social sciences (SPSS) version 24 was utilized as the basic statistical analysis environment due to its ease of use, ability to develop high-quality
developing the three mathematical models using MLR to predict the earned value indexes.

10. **STATISTICAL ANALYSIS OF MLR MODELS**

The statistical analysis is summarized in Table 2. Several significant statistics must be considered, including the coefficients of correlation (R), determination (R²), adjusted R², and the prediction of the standard error of the estimate. The statistical analysis involved an output of EV indexes and input of BAC, ACWP, A%, BCWP, P%, and BCWS. Furthermore, the R-values for the CPI, SPI, and TCPI models were 87.1%, 97.4%, and 94.4%, respectively, indicating a high correlating rate. In addition, the R² refers to the variation rate in the input variable, which can be predicted from the output. The obtained R² values for the CPI, SPI, and TCPI were 75.9%, 94.9%, and 88.91%, respectively. The model was chosen based on the maximum (R²) and the smallest value of the standard error of the estimate.

**Table 2** Synopsis of Statistical Analysis of MLR Models.

<table>
<thead>
<tr>
<th>No.</th>
<th>Model R</th>
<th>R²</th>
<th>Adj. R²</th>
<th>Std. Er.</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CPI</td>
<td>0.871</td>
<td>0.759</td>
<td>0.747</td>
<td>a. Predictors: (Constant), ACWP, BCWS, BCWP</td>
</tr>
<tr>
<td>2</td>
<td>SPI</td>
<td>0.974</td>
<td>0.949</td>
<td>0.946</td>
<td>a. Predictors: (Constant), ACWP, BCWS, A%</td>
</tr>
<tr>
<td>3</td>
<td>SPI</td>
<td>0.973</td>
<td>0.948</td>
<td>0.946</td>
<td>b. Predictors: (Constant), BCWS, A%</td>
</tr>
<tr>
<td>4</td>
<td>TCPI</td>
<td>0.944</td>
<td>0.891</td>
<td>0.886</td>
<td>a. Predictors: (Constant), ACWP, BCWS, BCWP</td>
</tr>
<tr>
<td>5</td>
<td>TCPI</td>
<td>0.945</td>
<td>0.888</td>
<td>0.885</td>
<td>b. Predictors: (Constant), ACWP, BCWS, BCWP</td>
</tr>
</tbody>
</table>

11. **FIT TEST IN MLR MODELS**

Table 3, illustrates the resulting values for the goodness of fit test, which has been conducted to test the MLR equations besides the independent variables’ efficiency in the quantitative measuring scale on earned value indexes. The table below indicates that the fitness rates are suitable, considering the effects of independent variables on the earned value indexes.

12. **CREATING MACHINE LEARNING REGRESSION (MLR) MODELS**

Regression models are essentially used to find the linear combination variables with the ideal correlation with independent variables. Its equation could be used for all three models, i.e., SPI, CPI, and TCPI. The regression equation is expressed in Eq. (1):

\[ Y = \alpha + \beta_1F_1 + \beta_2F_2 + \beta_3F_3 + \beta_4F_4 + \beta_5F_5 + \beta_6F_6 \]
where

\[ Y \] is the dependent variable (earned value indexes; CPI, SPI, and TCPI)

\[ A \] is the regression constant.

\[ \beta_1 \text{ to } \beta_6 \] are coefficients of regression for the factors.

12.1. Cost Performance Index (CPI) Model

Table 4 presents the regression statistics for the CPI model and explains the estimation of the MLR, including the Standardized and Unstandardized Coefficients and the Standard Error, as well as the significance of independent and constant variables, which could be stated as in Eq (2):

\[
CPI = 1.381 - (1.643 \times 10^{-10} \times BCWS) + (1.138 \times 10^{-9} \times BCWP) - (1.233 \times 10^{-9} \times ACWP) \ 
\tag{2}
\]

Implementing the above equation can be clarified through a numerical example using the data applied in the MLR model training for CPI. The Planning Value (PV) is ($1,797,500,000), earned value (EV) is ($1,653,879,750), and AC is ($1,554,648,510). The predicted value obtained through the aforementioned equation is (CPI=1.051), which is relatively accurate compared to the actual value measured manually (CPI=1.064). These differences in values are relatively minor.

\[ R^2 = 0.995, \text{ Standard Error} = 0.000 \]

12.2. Schedule Performance Index (SPI) Model

Table 5. below presents the regression statistics for the SPI model, and could be stated as in Eq. (3):

\[
SPI = 0.474 - (2.934 \times 10^{-10} \times BCWS) + (1.326 \times A\%\) 
- (1.449 \times 10^{-10} \times ACWP) \ 
\tag{3}
\]

Implementing the above equation can be clarified through a numerical example using the data applied in the MLR model training for SPI. The Planning Value (PV) is ($1,797,500,000), the Actual Percentage (A\%) is 0.920\%, and AC is ($1,554,648,510). The predicted value obtained through the aforementioned equation equals (SPI=0.942), which is relatively accurate compared to the actual value measured manually (SPI=0.92). These differences in values are relatively minor.

\[ R^2 = 0.999, \text{ Standard Error} = 0.000 \]

13. Verification and Validation of the MLR Models

The models’ performance is evaluated statistically using several metrics, such as [13, 14]:

13.1. Mean Percentage Error (MPE)

It is one of the most significant measures of a proposed network’s accuracy. It is the mean of the percentage differences between the predicted and the observed values. It can be obtained using Eq (5):

\[
MPE = \frac{\sum|X-Y|}{n} * 100 \quad (5)
\]

13.2. Root Mean Squared Error (RMSE)

The second criterion is a popular measurement of error, characterized by the focus on larger errors more than smaller ones. It can be obtained using Eq (6):

\[
RMSE = \sqrt{\frac{\sum(Y-X)^2}{n}} \quad (6)
\]

13.3. Mean Absolute Percentage Error (MAPE)

Eq. (7) is used to calculate the mean absolute percentage error.

\[
MAPE = \frac{\sum|X-Y|}{n} * 100\% \quad (7)
\]

13.4. Average accuracy percentage (AA %)

The Average accuracy percentage (AA \%) states that the accuracy performance can be obtained by (100–MAPE) %. Therefore, the Average Accuracy (AA) could be defined through Eq. (8):

\[
(\text{AA}%) = 100\% - \text{MAPE} (8)
\]
13.5. The Coefficient of Correlation (R)

The Coefficient of Correlation (R) is a measure that determines the relative correlation and the goodness-of-fit between the predicted and observed data. Eq. (9) is used to express the coefficient of correlation.

\[ r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma (x - \bar{x})^2 \Sigma (y - \bar{y})^2}} \]  \hspace{1cm} (9)

13.6. The Coefficient of Determination (R²)

R² shows how well the model outputs match the target value.

where:

- x = actual value
- y = estimated value or predicted value
- n = total number of cases (for validation)

Table 7 shows that the best performances are observed when the CPI model was verified with a high correlation rate (R) of (92.4%). Fig. 3 presents the validation of the MLR model for CPI. Since the coefficient of determination (R²) equals (85.41%), it can be stated that the CPI model presents an excellent agreement with the actual measurements. The trained MLR models can be used to predict the earned value indexes of the ten-spare data unused in any subset yet. The generalization results of the CPI model with (R² = 82.03%) were excellent, as shown in Fig. 4. Table 8 shows that the SPI model performed well throughout the verification stage, with a high correlation (R) value of (98.4%) between the actual and estimated values. Fig. 5 presents the validation of the MLR model for SPI. Since the coefficient of determination (R²) equals (96.87%), it can be stated that the SPI model presents an excellent agreement with the actual measurements. Fig. 6 shows the generalization results for the SPI model with (R² = 94.13%), which are excellent. Table 9 shows that the best performances are observed when the TCPI model was verified with a high correlation rate (R) of (93.7%). Fig. 7 presents the validation of the MLR model for TCPI. Since the coefficient of determination (R²) equals (87.83%), it can be stated that the TCPI model presents an excellent agreement with the actual measurements. The results of generalization from the TCPI model with (R² = 87.56%) were excellent, as shown in Fig. 8. Table 10 draws a comparison between the three MLR models’ results. The MAPE and AA% obtained through the CPI model were 4.806% and 95.194%, respectively, whereas the SPI model obtained 7.11% and 92.89%, respectively. Finally, the values for these two parameters obtained through the TCPI model were 16.294% and 83.706%, respectively. The results indicated that these three MLR models agreed with the measured values.

![Fig. 3 Comparison of Predicted and Actual for the CPI Model.](image)

![Fig. 4 Generalization of CPI Model.](image)
The study described a Multiple Linear Regression approach for developing the three models and explained the data collection and statistical specifications of historical information. A prediction model was built using the MLR approach with a backward elimination technique to obtain accurate Earned Value Indexes through the SPSS software. The results showed that the MLR technology performed well in determining the AA% and R values for CPI, SPI, and TCPI. Their values were 95.194%, 92.195%, and 83.706% for AA% and 92.4%, 98.4%, and 93.7% for R for each CPI, SPI, and TCPI, respectively. Therefore, the MLR technique can be used as a part of the derivation process in the prediction model due to its accurate earned value predictions.

REFERENCES


[10] Alfahham AFH, Alajeeli HKB. Building a Predictive Model to Improve the Quality of Government Building Construction


