A Predictive Model for Scaffolding Man-hours in Heavy Industrial Construction Projects

W. Chu[a] , S.H. Han[a] , L. Zhen[b] , U. Hermann[c] ,and D. Hu[c]

[a]Department of Building, Civil and Environmental Engineering, Concordia University, Canada
[b]Department of Civil Engineering, University of New Brunswick, Canada
[c]PCL Industrial Management Inc., Edmonton, Canada

monian0627@gmail.com, sanghyeok.han@concordia.ca, zlei@ualberta.ca, RHHermann@pcl.com, dhu@pcl.com

Abstract –

In typical heavy industrial construction projects, scaffolding can account for 30% to 40% of the total direct man-hours. However, most industrial contractors estimate scaffolding man power based on a certain percentage of the direct work, which leads to cost increase and schedule delay due to inaccurate estimation. In order to aid industrial companies to plan and allocate the resources for scaffolding activities before construction, this paper proposes a methodology which combines the classification tree and multiple linear regression to estimate scaffolding manhours based on available project features. The evaluation matrix involves R Squared value (R^2), Adjusted R Squared value (Adj. R^2), mean absolute error (MAE), root mean squared error (RMSE), and relative absolute error (RAE). The proposed methodology has been tested on the historical scaffolding data in a heavy industrial project and the results showed its effectiveness.

Keywords –

Scaffolding Man-hours; Linear regression; Classification tree; Heavy Industrial Constructions

1 Introduction

Occupational safety and health services [1] defines scaffolding as any structure (suspended structure) which is built for temporary purposes, used for the support and/or protection of the construction workers by providing easy access to work areas horizontally and vertically, and also helps in material transferring. Due to these functions, heavy industrial construction projects usually involve various types and a large amount of scaffoldings to feed the need for different disciplines (e.g., civil, mechanical, and electrical), leading to increased project costs. In the construction site, the scaffolding should be installed, modified and/or dismantled in accordance with the requirements of various disciplines on their demand times in order to prevent project schedule delays. Due to the demand-based scaffolding operation, the construction domain has difficulty to plan scaffolding operation in the early phases of the project.

In practice, planning of scaffolding activities is completely subjective and differs from company to company [2]. The scaffolding tends to be planned and operated as an ad hoc way which leading to schedule delays and cost overrun due to the inefficient utilization of resources. As an effort to develop a scientific and practical planning method for the scaffolding activities, the previous research [2] claims that most construction companies regard scaffolding as a part of indirect expense and calculated as a percentage of the total man-hours of direct work . In this respect, previous study [3] has identified that scaffolding works accounts for up to 30%-40% of the total direct man-hours in the heavy industrial project.

Scaffolding has potential for significant productivity improvement with respect to project cost reduction in construction, especially industrial construction. However, planning and estimating of scaffolding works have received little attention in academia and practice. In its infancy, research on scaffolding mainly focused on structural performance [4]. With safety gaining prominence in scaffolding research, the factors contributing to scaffolding collapse have been studied more recently [5]. In terms of planning of scaffolding works as a temporary structure on-site, there are several research efforts that have investigated the application of artificial intelligence (AI) algorithms (e.g., fuzzy logic and genetic algorithm) and geometries of 3D models for temporary structure or facility planning [8]. However, these studies have not fully directed their efforts to developing methods or systems to improve efficiency of planning and estimating of scaffolding activities for productivity improvement based on the project time progress.

Thus, this paper proposes an integrated method that combines the classification tree with the multiple linear regression model in order to estimate the scaffolding man-hours efficiently and accurately in the heavy industrial construction projects. The proposed methodology mainly consists of five steps: (*i*) collecting data from construction site through cloud-based computer systems; (*ii*) identifying and cleaning potential outliers from the dataset; (*iii*) selecting and/or transforming the most important independent variables which affect the scaffolding man-hours through statistical diagnosis, (*iv*) developing classification tree based on selected text variables in the dataset, and at each tree node, the multiple linear regression performed to obtain a predictive model for man-hour estimation, and (*v*) evaluating the performance using the tenfold repeated cross-validation.

The proposed methodology has been implemented in Python 3.7 environment. It was tested with the historical scaffolding data in a heavy industrial project in Alberta, Canada, provided by an industrial collaborator.

2 Scaffolding Related Research

Based on the investigation of the previous studies done by authors in planning the scaffolding-related resources in construction domain, previous research has mainly focused on structural performance and safety management. Peng et al. [4] introduced a scaffolding design system in terms of the performance of steel and bamboo scaffolding. Yue et al. [5] studied the effect of wind load on scaffolding in order to promote safety in the design of integral-lift scaffolds. Based on optimization of the scaffolding schedule, which can help in coordinated safety management and control efforts, Hou et al. [6] introduced an operational framework by integrating mathematical models with virtual simulation to optimize scaffolding erections. According to the introduction of advanced technologies, Kim et al. [7] have actively integrated building information modelling (BIM), image processing or wireless sensors with optimization algorithms, to not only identify and mitigate the safety risks but also plan the scaffolding schedules to eliminate potential hazards. Furthermore, Cho et al. [10] have used machine learning algorithms (support vector machine) to assess real-time safety and unsafety status of the scaffolds based on different conditions of scaffolds (safe, overturning, overloading or uneven settlement), which adopts actual strain data of scaffolding members obtained by wireless sensors.

Construction Owners Association of Alberta (COAA) reports that scaffolding plan must provide the estimated scaffolding types, location, duration and quantity requirements including materials and labours [11]. Based on the result of planning and estimating the scaffolding activities, previous studies [12] have suggested that the effective management of the scaffolding in construction projects can improve productivity by: (i) preventing the delays of crews due to absence of scaffolding materials or even man-power; and (ii) understanding the resource requirements to avoid work space conflicts. However, in practice, industrial company plans and estimates the scaffold activities subjectively, which can be up to 40% of total direct man power of an industrial project, based on the regulations of company and engineer's experience which may cause excessive use of man-power, schedule delays and resource shortage.

As scaffolds and their supporting structures being a temporary work platform, there are several research efforts to investigate the applications of artificial intelligence (AI) algorithms and geometries of 3D models for temporary structure or facility planning [9]. However, these studies have not discovered the field of improving efficiency and accuracy of planning and estimating scaffolding man powers. Therefore, several studies [2-3] made an effort to analyse the factors affecting industrial scaffolding estimation based on historical data provided by a construction company. A simulation tool and linear regression models have been developed to predict a range of man-hour values for only scaffold erection on site in their works. However, since the lack of available scaffolding data, the analysis merely regarding scaffolding erection is insufficient. The authors also suggested that further analysis is required with more historical data from other industrial projects and other optimization algorithms in order to generalize a general methodology for estimation and planning of scaffolding works. As a recent study, Moon et al. [13] have investigated the effect on productivity of resource configurations measured during scaffolding operation as part of the construction of an actual liquefied natural gas (LNG) plant. Hou et al. [14] have proposed a feasible multi-object discrete firefly algorithm for optimizing scaffolding project resources and scheduling. However, this model needs to not only be improved in terms of accuracy for scheduling the scaffolding resources, but also be made generically applicable to other projects by incorporating various types of scaffolding constraint. In addition, these previous studies have had difficulty for further analysis or better models to plan and/or schedule the scaffolding activities due to the insufficient scaffolding data since it has not been attention and tends to be ignored in practice. As a result, the development of scientific and systematic methods is still required in the planning and estimating of scaffolding activities due to the lack of accuracy, efficiency, and applicability in the existing systems for various types of construction projects, especially heavy industrial projects.

3 Research Methodology

In order to predict the required man-hours for scaffolds accurately in the early planning phase of the project, this paper proposes an integrated method that combines the classification tree with the multiple linear regression model. As shown in Figure 1, there are in total five steps in the proposed methodology, which are data collection, data cleaning, input data determination, development of classification tree structure, and the final model evaluation with all the outputs from the previous steps.

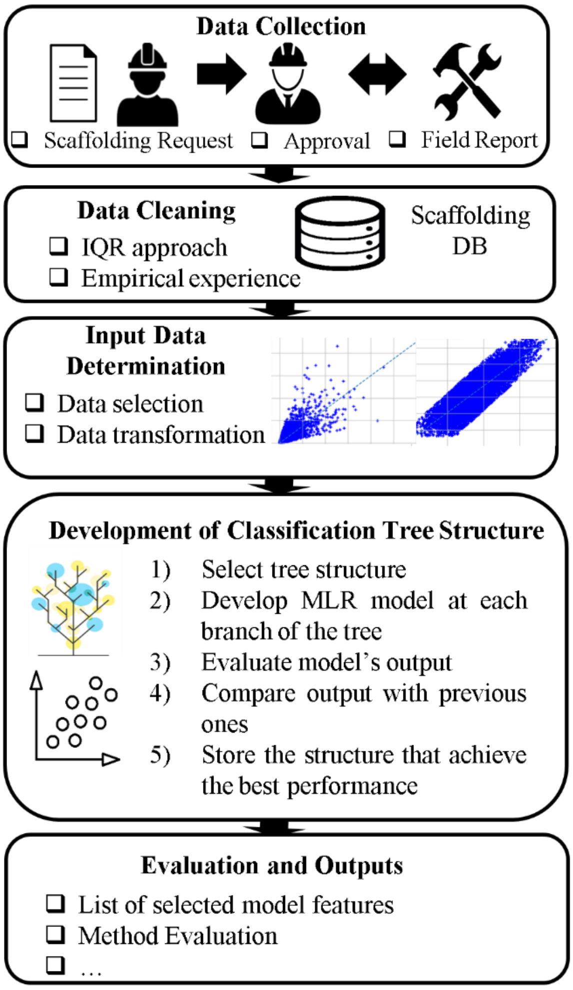

Figure 1. Flowchart of proposed methodology

3.1 Data Collection

As mentioned in Figure 1, the data collection process can be simplified as: scaffolding requests generated for approval along with detailed scope of work, and once the requests are approved and completed, the actual man-hours and work details are recorded. Individual scaffold component details are documented, and the weights summed up for each request by date. The information that can be tracked in the scaffolding activities may vary, but by nature, the key ones are work classification (i.e. erection, modification, and dismantle), scaffolding type (i.e. platform deck, tower, barricade, etc.), actual man-hours, total scaffolding weights. Meanwhile, other project related data that maybe likely to affect the manhour prediction should be extracted from other sources and consolidated for analysis, such as the average temperature during each scaffolding task, the elevation of the scaffold built according to the ground level, and the average aluminum percentage of the scaffolding.

3.2 Data Cleaning

The objective of data cleaning is to remove the outliers in the collected dataset. Outliers are extreme observations in the dataset that are not consistent with the trend of correlation in the data. In data collection process, the outliers may result from errors in data entry. There are generally, two ways of filtering outliers: (i) using statistical approaches to identify outliers mathematically regardless of the nature of data; and (ii) using user experience-based approaches to identify outliers based on users' logics. Often, the statistical approaches take less efforts than the user experience-based approaches due to the experience-based approach do not have a certain criterion but based on understanding, experience and trail-and-error tests. In this paper, an integrated data cleaning process has been adopted using both statistical and experience-based methods.

At first, the interquartile range (IQR), one of the statistical approaches, is used to identify the outliers. IQR is a measure of the location of middle 50% data in the dataset, and it is calculated by subtracting the first quartile (Q1) from the third quartile of the dataset (Q3). Potential outliers are defined as observations that fall below Q1 – 1.5* IQR or above Q3 + 1.5 * IQR. Moreover, under the guidance of the scaffolding experts, the following experience-based rules should be applied to further filter out the outliers: (*i*) value of man-hours is null or less than 5 hours; (*ii*) value of scaffolding weights is null or less than 20 lbs; and (*iii*) productivity (i.e. weight per man hour) is less than 6 lbs/hr, or greater than 125 lbs/hr. These experience-based rules may vary in different project scenario. However, the core concept is to remove data observations that are not physically feasible in scaffolding work, normally reflected by manhours, weight of scaffolding, and scaffolding weight divided by manhours (productivity).

3.3 Input Data Determination

In order to ensure the effectiveness of results from multiple linear regression model, there are five assumptions need to be fulfilled before the analysis [15]. The assumptions are (*i*) the relationship between dependent variables and independent variable should be approximately linear; (*ii*) the error term ε has zero mean; (*iii*) the error term ε has constant variance σ2; (*iv*) the independent variables are uncorrelated; and (*v*) the errors are normally distributed. Among these assumptions, the linearity between the dependent variables and independent variable and the independence among the independent variables are of the utmost importance. The zero mean of error term can be fulfilled by involving the intercept term in the regression equation and the normally distributed error generally is not a must-have check. The following model feature selection section and model feature transformation section are responsible for check independence and linearity among the dataset, separately.

3.3.1 Model Feature Selection

The model feature selection is one of the most important part in the field of machine learning since (i) the irrelevant input features can induce greater computational cost; (ii) the irrelevant input features may lead to overfitting, which in turn leads to poor results on the validation datasets. Feature selection methods can also adapt the dataset to better suit the selected machine learning algorithm, given that different algorithm may have various requirement for the features. In terms of multiple linear regression, a reliable set of features contains independent variables that are highly correlated to a dependent variable (i.e., scaffolding manhour), also called as a predicted variable, but uncorrelated with each other.

The condition that some of the independent variables are highly correlated is called collinearity [16]. Collinearity can lead to imprecise coefficient estimates during the development of the predictive model since it inflates the standard errors of the coefficients of collinear variables. In this respect, this paper uses correlation matrix which supports users to identify the collinearity problem. The correlation matrix for all the independent variables should be developed at each classification tree node. The Spearman method [17] has been adopted here to perform the correlation analysis since (*i*) it is a non-parametric procedure in which the observations are replaced by their ranks in the calculation of the correlation coefficient so that it can deal with data with outliers; (*ii*) it does not carry any assumptions about the distribution of the data (e.g. Pearson method requires both variables to be normally distributed) [18]. In the correlation matrix M, it is easy to identify that which two variables are highly correlated (> 0.5) [19] and which one of them should be removed to avoid collinearity.

While collinearity means the correlation between only two independent variables are high, multicollinearity can exist between one variable and linear combination of more than two variables [16]. As another indicator of model feature selection for linear regression, multicollinearity can cause regression coefficients to change dramatically in response to small changes in the model or the data. Thus, it may cause serious difficulty with the reliability of regression coefficients. In order to detect whether a regression model exists multicollinearity, Variance Inflation Factor (VIF) [20] of each independent variable need to be checked in the model. VIF is a traditional measure to detect the presence of multicollinearity in multi-linear regression model. It shows how much the variance of the estimator is inflated due to the linear relation between the regressors. Typically, a VIF, which is larger than ten, has been used as a rule of thumb to indicate serious multicollinearity.

3.3.2 Model Feature Transformation

Given a selected feature set, the quality of data can be enhanced by feature transformation. It is common that the real-world data may not show strong linearity between independent variables and predicted variable. However, there are several data transformation methods, such as logarithm, square root, reciprocal, cube root and square, can be applied to enhance the linear trend in data. There are also some guidelines for the selection of transformation method, such as if the standard deviation is proportional to the mean, the distribution can be positively skewed and logarithmic transformation can be performed, or if the variance is proportional to the mean, squared root transformation may be preferred etc. [21].

Among various transformation methods, logarithmic transformation is the most popular one. Using natural logs for variables on both sides of the linear regression equation can be called log-log model. Theoretically, any log transformation can be used in the transformation and all of them tend to generate similar results. However, using the natural log can be seen as the convention since the interpretation of the regression coefficients is obvious using the natural log. The coefficient in the natural log-log model represents the estimated percent change in the dependent variable for a percent change in the correspondent independent variable [22] .

3.4 Model Development

The classification tree structure determines the way how the overall dataset can be divided into several groups in which the regression model can be separately developed. The function of the classification tree is to cluster the similar observations together to obtain more accurate regression sub-models instead of messing all the data together and get only one model. The proper classification may largely alleviate the effort to achieve the required model efficiency and accuracy. To efficiently build the classification tree, the key categorical variables used to split the tree are of the utmost importance. There are several available categorical variables such as work classification, scaffold type and discipline. The repeated tenfold cross-validation can be used to compare the effectiveness of different classification tree structures developed using various combinations of categorical variables.

The classification tree needs to be utilized with multiple linear regression model. After developing the classification tree structure based on selected key categorical variables, multiple linear regression can be implemented at each tree end node, as long as there are sufficient number of records at the tree node. Previous researchers discovered that when the number of observations n >= 15 * k where k represents the number of independent variables, the model parameter estimation tends to achieve a prescribed level of accuracy [23]. On the other hand, if the number of records at certain nodes cannot meet this requirement, the upper level model should be used to complement the miss of model under certain classification category. Normally, the multiple linear regression model at each classification tree node can be denoted as Eq. (1).

$$y = \beta + \sum \alpha_i \, x_i \quad (1)$$

Where: y is the predicted variable (i.e. scaffolding manhour) for the *ith* record at current classification tree node; β is the intercept value; α_i is the coefficient for independent variable x_i.

3.5 Model evaluation

Repeated tenfold cross validation has been selected to evaluate the regression models. Kim has found that the repeated cross-validation estimator is recommended for general use regardless of the sample size [24]. Moreover, Witten et al. have conducted extensive tests with different machine learning techniques, and the results have shown that, by repeating the tenfold cross validation 10 times, the results can reliably estimate errors [25]. Tenfold cross validation means to split the whole dataset into ten stratified subsets of equal size ten times, and each subset can be used for testing once and the combination of the rest can be used for training. The final error estimates are averages across each of the fold. Repeat the tenfold cross validation ten times and the mean value would be the final validation result. The evaluation matrix has five parameters in all, including R Square (R^2), Adjusted R Square (Adj.R^2), Mean Absolute Error (MAE), Root Mean-Squared Error (RMSE) and Relative Absolute Error (RAE). All the calculation equations can be found in a book [25] .

4 Case study

The proposed methodology has been implemented in the Python 3.7 environment and the case study is based on a heavy industrial project with data provided by a construction company. The scaffolding related data has been collected onsite through cloud-based data systems, as well as company's internal systems (e.g. project control systems, payroll systems, etc). There are in total three categorical variables and twelve numerical variables have been collected. The categorical variables are work classification, scaffolding type, discipline of trade that scaffolding is built/modified for.

The numerical variables are: average temperature, apprenticeship ratio (the ratio of work carried out by apprentices), night-time ratio, overtime ratio, aluminium percentage of the scaffolding, percentage of completed project, scaffolding weight, workable area (available space for building scaffolds), average scaffolding employee time on site, elevation of the scaffolds, major pieces and minor pieces (number of large/small pieces of scaffolding materials). After data consolidation and cleaning, the dataset contains 12,087 valid observations in total. Figure. 2 compares the manhour distribution of the raw dataset and the dataset after data cleaning.

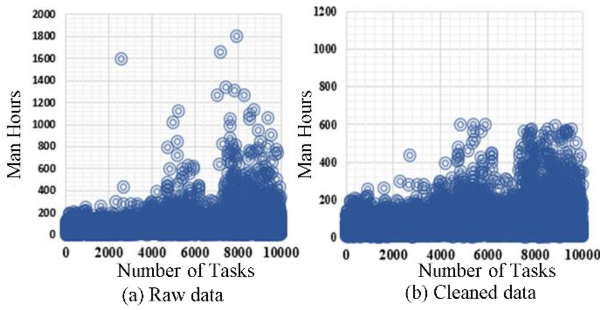

Figure 2. Manhour distribution comparison

After the data collection and data cleaning, the independence in the dataset should be checked. Table 1 illustrates a part of the sample correlation matrix of a classification tree branch (classification: Erection). The independent variables that have much lower correlation with others have been trimmed in the table due to the space limitation; only the most correlated variables were kept: workable area, scaffolding weight, major pieces, and minor pieces. It can be seen that almost all the values in Table 1 are larger than 0.5, which means the four variables are all highly correlated to each other. It should also be noted that all these variables are also highly correlated to the scaffolding manhours which is the predicted variable. According to the previous research, the collinearity can be simply addressed by keeping only one highly correlated independent variable but removing the others [26]. From the industrial view, these four variables describe the scaffolding work from a similar perspective. To determine which variable is of the most importance, the multi-linear regression models have been developed based on one of these four variables in turn with the rest un-collinear eight variables. The sample results using classification as key variable can be found in Table 2.

Table 1. Sample correlation matrix

	Workable Area	Weight	Major Pieces	Minor Pieces
Workable Area	1.00	0.66	0.63	0.49
Weight	0.66	1.00	0.87	0.62
Major Pieces	0.63	0.87	1.00	0.67
Minor Pieces	0.49	0.62	0.67	1.00
Manhour	0.51	0.60	0.61	0.59

Table 2. Result using different variables

Index	Workable Area	Weight	Major Pieces	Minor Pieces
R^2	0.54	0.73	0.69	0.58
Adj.R^2	0.54	0.73	0.68	0.58
MAE	70.69	51.51	52.18	61.82
RMSE	134.65	107.56	115.32	145.36
RAE	1.30	0.69	0.72	1.01

It can be clearly seen from Table 2 that R^2, Adj.R^2 are the highest while MAE, RMSE, and RAE are the lowest when selecting the scaffolding weight variable. Since the weight variable can give less error and better predicted results, it has been kept in the model feature set but the other three have been excluded to prevent the collinearity.

Moreover, Table 3 shows VIF of all the numerical independent variables in the model at the branch of Erection-Tower as an example. The result shows that there are four variables with VIF factor value larger than 10, which are night-time ratio, project complete percentage, employee time on site, and apprenticeship ratio. Thus, these four variables have been excluded from the analysis to address the multicollinearity problem. Table 4 has been created to detect multicollinearity one more time after the adjustments. It can be seen that the multicollinearity has been solved since all the VIF values are lower than 10 when regress the model using the selected variables.

Table 3. VIF of each numerical variable

VIF Factor	Features
3.61	Temperature
3.52	Aluminum percentage
1.58	Weights
2.69	Elevation meters
1.07	Night-time ratio
28.00	Overtime ratio
70.06	Project complete percentage
223.20	Employee time on site
76.84	Apprenticeship ratio

Table 4. VIF of modified variable

VIF Factor	Features
1.13	Temperature
1.22	Aluminum percentage
2.03	Weights
1.04	Elevation meters
2.04	Night-time ratio

The regression model has been built using the selected five variables which are temperature, aluminum percentage, weight, scaffolding elevation, and night-time ratio. However, some negative manhours are predicted by the model. This phenomenon prompts the discovery of the exact contribution of each explanatory variable. According to the summary of previous work [27], the Standardized Regression Coefficients (SRC) method is suitable to conduct the sensitivity analysis for the linear model. Table 5 shows the result of sensitivity analysis for the classification tree built upon work classification and scaffolding type. It should be noted that the average value of night-time is negative, which indicates that with the increase of night-time working, the required manhours decreases. Thus, the night-time ratio has been removed not only due to its minimum contribution to predict manhours, it is also counter-intuitive to industry experts.

Table 5. Average result of sensitivity analysis

	Temp	Weight	Elevation	Night time	Aluminum percentage
Avg	-0.05	0.83	0.05	-0.02	0.07

Thus, there are four independent variables have been selected at the end. Next, the linearity between the independent variables and predicted variable should be checked. From the sensitivity analysis, it is obvious that in the linear regression model, the scaffolding weight is far more important than other variables. As the most prominent variable, the relationship between weight and manhour has been discovered and some sample scattered figures have been plotted in Figure 3. It can be seen that the dots in the upper left figure are scattered. The original manhour do not show strong linear trend with original scaffold weights. However, after taking the natural logarithm of both manhour and weights, there exists clear linearity between log(weight) and log(manhour). Other data transformation ways have been tried out as well, such as reciprocal transformation (weights versus reciprocal of manhour) shown in the lower left figure, and single logarithmic transformation which means only transforming one side of variable (log(weight) versus manhour) as shown in the lower right figure. Nevertheless, the log-log model is obviously the best to generate linearity.

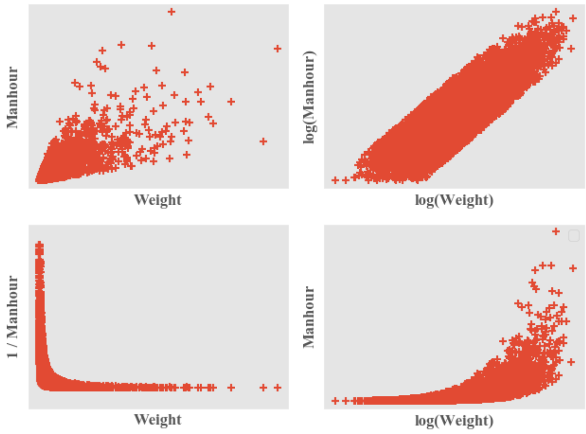

Figure 3. Comparison of data transformation

Furthermore, the log-log model should be applied to all the independent variables instead of only scaffolding weights. It is worth mentioning that since the log-log model can only be applied to positive variables, the temperature variable has been normalized in the range [0°, 10°]. In addition, to deal with the zeros in the dataset, a constant λ has been added to the log-log transformation equation as shown in Eq. (2).

$$\log(y) = \beta + \sum \alpha_i * \log(x_i + \lambda) \quad (2)$$

where λ can be one half of the smallest value in the dataset . In this study, λ = 0.001 has been adopted.

As for the classification tree structure, the trail-and-error method has been used to determine the best classification method. The repeated tenfold cross-validation results for various classification structures can be found in Table 6. Overall, the work classification works as the best classification tool since it gives the highest Adj. R2, way more than 0.7 which is the criterion to check whether a linear model is a good fit. Also, it produces less error than using other classification methods. After the regression models have been built at each tree node in the classification tree, the correspondent coefficient tables are stored for the future use.

Table 6. Cross-validation results

CT	R^2	Adj.R^2	MAE	RMSE	RAE%
WC	0.81	0.8	38.24	82.53	37.93
DIS	0.78	0.78	48.17	101.93	42.54
ST	0.76	0.76	47.49	87.94	41.09
WC+DIS	0.79	0.79	41.52	94.33	38.84
WC+ST	0.82	0.82	44.5	94.06	36.02
DIS+ST	0.79	0.78	48.04	99.36	40.5

5 Conclusion

As one of the largest temporary works, scaffolding is indispensable, but difficult to manage. Over decades, its requirement has been generally determined based on a percentage factor of the direct work in that project and the expert's opinion. It may result in an ineffective management of scaffolding-related resources and project cost. To address this practical issue, this paper proposes an integrated methodology to predict the required man-hours in the planning stage of the project. There are five main steps in the proposed methodology, which are the data collection, data cleaning, input data determination, development of classification tree structure, and model evaluation.

A case study has been implemented to validate the methodology. From the methodology, four independent variables have been selected at the end. Moreover, it has been found that when use the log-log transformation in the model, the linearity can be built between independent variables and dependent variable. After that, the results show that when using the work classification to build up the classification tree, the performance can be maximized. The best model produces R^2= 0.81, Adj. R^2=0.8, MAE=38.24, RMSE=82.53, and RAE=37.93%.

The current work has been proven to be effective to predict the manhours based on four independent variables which should be available in the early stage of scaffolding construction. However, with the increase of the amount of the available and reliable scaffolding data, the model can be further fine-tuned and trained. Moreover, the non-linear regression such as neural network is still a further research direction and needs to be discovered and compared with the proposed methodology in this research.

6 Acknowledgements

The authors would like to thank the support from our industry partner, PCL Industrial Management Inc. The research funding support from NSERC (Natural Sciences and Engineering Research Council) CRD grant (CRDPJ-536164-18) is also acknowledged.

References

[1] Occupational Safety and Health Services (OSHA) (1994), Safe Erection and use of Scaffolding, Department of Labour, Wellington, New Zealand.

[2] Kumar, C., S. M. AbouRizk, Y. Mohamed, H. Taghaddos, and U. Hermann (2013), Estimation and planning tool for industrial construction scaffolding, Proc., 30th ISARC, International Association for Automation and Robotics in Construction, Bratislava, Slovakia, 634–642.

[3] Wu, L., Y. Mohamed, H. Taghaddos, and R. Hermann (2014), Analyzing scaffolding needs for industrial construction sites using historical data, ASCE Construction Research Congress (CRC), 1596–1605.

[4] Peng, J. L., A. D. Pan, D. V. Rosowsky, W. F. Chen, T. Yen, and S. L. Chan (1996), High clearance scaffold system during construction II, Structural analysis and development of design guidelines, Engineering Structure, 18(3), 258–267.

[5] Yue, F., Y. Yuan, G. Li, K. Ye, Z. Chen, and Z. Wang (2005), Wind load on integral-lift scaffolds for tall building construction, Journal of Structural Engineering, 131(5).

[6] Hou, L., C. Wu, X. Wang, and J. Wang (2014), A framework design for optimizing scaffolding erection by applying mathematical models and virtual simulation, Proc., International Conference on Computing in Civil and Building Engineering, 323–330.

[7] Kim, K., Y. K. Cho, and Y. H. Kwak (2016), BIM-based optimization of scaffolding plans for safety, ASCE Construction Research Congress (CRC), 2709–2718.

[8] Sulankivi, K., T. Mäkelä, and M. Kiviniemi (2009), BIM-based site layout and safety planning, 1st Int. Conf. on Improving Construction and Use through Integrated Design Solution CIB, Espoo, Finland.

[9] Kim, J., M. Fischer, J. Kunz, and R. Levitt (2015), Semiautomated scaffolding planning: Development of the feature lexicon for computer application, Journal of Computing in Civil Engineering, 29(5).

[10] Cho, C., J.W. Park, K. Kim, and S. Sakhakarmi (2018), Machine Learning for Assessing Real-Time Safety Conditions of Scaffolds, Proc., 35th ISARC, International Symposium on Automation in Construction.

[11] Construction Owners Association of Alberta (COAA) (2013), Workforce Planning Subcommittee, Construction Work Packages Best Practice, Construction Owners Association of Alberta (COAA), Document Number: COP-WFP-SPD-16-2013-v1.

[12] Guo, S. J., (2002), Identification and resolution of work space conflicts in building construction, Journal of Construction Engineering and Management, 128(4), 287-295.

[13] Moon, S., J. Forlani, X. Wang, and V. Tam (2016), Productivity study of the scaffolding operations in liquefied natural gas plant construction: Ichthys project in Darwin, Northern Territory, Australia, Journal of Professional Issues in Engineering Education and Practice, 142(4).

[14] Hou, L., C. Zhao, C. Wu, S. Moon, and X. Wang (2017), Discrete firefly algorithm for scaffolding construction scheduling, Journal Computing in Civil Engineering 31(3).

[15] Lucko, G., Anderson-Cook, C. M., & Vorster, M. C. (2006). Statistical considerations for predicting residual value of heavy equipment. Journal of construction engineering and management, 132(7), 723-732.

[16] A. Alin, Multicollinearity, Wiley Interdiscip. Rev. Comput. Stat. 2 (2010) 370–374.

[17] Zar, J. H. (1972). Significance testing of the Spearman rank correlation coefficient. Journal of the American Statistical Association, 67(339), 578-580.

[18] D.G. Bonett, T.A. Wright, Sample size requirements for estimating Pearson, Kendall and Spearman correlations, Psychometrika. 65 (2000) 23–28. doi:10.1007/BF02294183.

[19] A. Hall, Mark, Correlation-based feature selection for machine learning, Diss. Univ. Waikato. (1999) 1–5. doi:10.1.1.149.3848

[20] Salmerón, R., García, C. B., & García, J. (2018). Variance Inflation Factor and Condition Number in. Journal of Statistical Computation and Simulation, 88(12), 2365-2384.

[21] Bland JM, Altman DG. Transforming data. BMJ 1996, 312:770

[22] Bland JM, Altman DG. The use of transformation when comparing two means. BMJ 1996, 312:1153.

[23] Stevens, J. P. 1995. Applied multivariate statistics for the social sciences. 3rd Ed., Lawrence Erlbaum Associates, Hillsdale, N.J

[24] J.H. Kim, Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap, Computer. Stat. Data Anal. 53 (2009) 3735–3745.

[25] Witten, I. H., Frank, E., Hall, M. A. & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques (4th Ed), Elsevier Science.

[26] S.B. Kotsiantis, D. Kanellopoulos, Data preprocessing for supervised leaning, Int. J. (2007) 1–7.

[27] Brevault, L., Balesdent, M., Bérend, N., & Le Riche, R. (2013). Comparison of different global sensitivity analysis methods for aerospace vehicle optimal design. 10th World Congress on Structural and Multidisciplinary Optimization.

[28] Osborne, J. (2010). Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research, and Evaluation, 15(1), 12.