Predicting in BIM Labour Cost with a hybrid approach
Simple Linear Regression and Random Forest

Jun Ze Wang and Jie Zhang
Engineering Management Department of Henan Technical College of Construction, Gongye Rd, Erqi District, Zhengzhou, Henan, China
Email address: 272261513@qq.com

Abstract: In most cases, BIM labour cost can only be calculated through the proportion of the gross floor area according to the practical projects, which always brings unpredictable risks to project managers since this simple method of linear regression contains a high risk of estimating error in construction phases. Therefore, this study will develop a new hybrid methodology which combines the Random Forest and Simple Linear Regression for eliminating the error of prediction on BIM labour cost in construction phases. A case study will be conducted to illustrate the prediction results through four completed projects, including two steel structure and two reinforced concrete projects.

1. Introduction:
BIM technology has been proved that it can provide enormous opportunities to enhance the project quality management, schedule management and cost management. However, the worthiness of BIM implementation has been regarded as a challenge in return-on-investment since evaluating the economic benefits of BIM is not an easy, simple task for projects manager. Most of the projects calculate BIM labour cost by gross floor area based on the type of the completed projects. However, the disadvantages of this method are evident, including the difficulty to detect and control error with high risks[1-5].

For solving this issue and facilitating the accuracy of prediction results, Random Forest is adopted to this research since it is widely used to deal with classification, unbalanced data and outliers [6]. In this study, the RF method would eliminate the deviation on selected projects, and cost breakdown structure (CBS) will be used to collect data from these projects.

2. Methodology:
The CBS will be established and separated into three phrases for BIM manager to collect labours’ working hours for the model to predict the BIM labour cost for the new project. The first phase is to record the project contents and corresponding labour hours by engineers. The second phase is to set up the cost database of a BIM group. In this level, the cost category includes different parts such as architecture, structure and Mechanical, Electrical and Plumbing (MEP) teams. Moreover, the operation of BIM team needs to be considered as well, including software and hardware investment and maintenance. This structure will be illustrated in figure 1. The last phase is to map the CBS cost activities, including building BIM model and revising BIM model. For example, to meet the practical requirements, the projects might be separated into structural BIM model, exterior BIM model and interior BIM model. A quotation will be given respectively, and cost-based activities need to be tailored[7-9]
Figure 1 BIM cost structure

Four residential building projects shown in table 1, are selected to collect relevant variables for machine learning. Accordingly, the output variable of the database is the gross floor area which based on the larger gross floor space needs higher human resources. Besides, the number of the basement, buildings were chosen as well to avoid deviation. In table 2, the first floor includes landscape area as well.
Establishing Random Forest Regression and Simple Linear Regression BIM model are chosen to predict the results. In order to evaluate these two models’ performance, Mean Absolute Error (MAE) are engaged and equations are written below:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

It is an effective way to test the validation database, including model stability. As shown in Table 3 indicates that Simple Linear Regression is more stable than the Random Forest model. However, only four projects were selected, which are not enough for the database establishment. Thus, the measurement results might change slightly when there are more projects data added into the model.[10-11]

| Project no. | Construction type | Gross floor area (m²) | Above-ground story | Basement story | Number of buildings |
|-------------|-------------------|-----------------------|-------------------|---------------|-------------------|
| 1           | Reinforced concrete | 18,920                | 18                | 3             | 1                 |
| 2           | Steel structure    | 38,793                | 39                | 5             | 1                 |
| 3           | Reinforced concrete | 13,185                | 12                | 2             | 1                 |
| 4           | Steel structure    | 29,792                | 27                | 5             | 1                 |

According to the above analysis, it is shown that the variance of the gross floor area may affect the performance of the models. Therefore, cluster analysis is engaged below to improve the accuracy of the model performance [12.] When the silhouette value ranges from -1-1, higher value figure indicates better clustering results. Therefore, the importance of predictors is shown in figure 2 to illustrate that the number of buildings and the first-floor area is more significant for clustering. After the cluster analysis, the RF model is not only better and more accurate for predicting BIM labour cost but also can converge faster than other models.

| Table 3 model comparison |
|--------------------------|
| cost activities         | Model | MSE  |
| Structural BIM model     | SLR   | 12.68|
| Working drawings         | SLR   | 2.57 |
|                          | RF    | 3.16 |
Based on the above analysis, Random Forest and Simple Linear Regression have its characters and suit in different circumstances. The model performance results show the RF, SLR and Hybrid approach respectively in table 4, based on the four projects, the SLR has more stable results compared to RF in steel structure, and RF has better results than SLR in rainforest concrete. However, the Hybrid has the most stable results in both structures according to the model performance results in table 4.

| Category           | Model | MSE  |
|--------------------|-------|------|
| Steel structure    | SLR   | 40.1 |
|                    | RF    | 41.8 |
|                    | Hybrid| 37.8 |
| Rainforest concrete| SLR   | 41.5 |
|                    | RF    | 40.6 |
|                    | Hybrid| 37.6 |

Therefore, a hybrid approach is proposed to acquire a more reliable result in the BIM labour cost prediction for project managers. In addition, CBS is proposed to adopt to divide collected data into two levels, organization level and project level. Meanwhile, a three-tier structure is essential for checking and revising the quality of data collection in BIM labour cost.

By applying the proposed hybrid approach, the prediction performance can be improved than a current practice which can reduce many risks for BIM projects later.

### 4. Conclusion:
This research has analysed the RF and SLR approaches in BIM labour cost prediction results. In addition, a new hybrid method combined RF and SLR are introduced and tested to acquire a better stable prediction performance in BIM labour cost. However, it still has its limitation. Due to limited funds and people, only four BIM projects are investigated to collect data for this research. It is not enough for data collection from the perspectives of structural types, scale and location of the projects. Thus, the study results may have its limitation. It is suggested that researchers should further intervene
in this limitation, expand the scope of investigation and establish a broader database to test the accuracy of the hybrid method proposed in this study.

Reference:
[1] B.K. Giel, R.R.A. Issa, Return on investment analysis of using building information modeling in construction, J. Comput. Civ. Eng. 27 (2013) 511–521, https://doi.org/10.1061/(asce)cp.1943-5487.0000164.

[2] B.K. Giel, R.R.A. Issa, BIM return on investment: a case study, Journal of Building Information Modeling 1 (2011) 24–27 https://www.brikbase.org/sites/default/files/Pages%20from%20jbim_spring11-5.giel_.pdf.

[3] D. Walasek, A. Barszcz, Analysis of the adoption rate of building information modeling [BIM] and its return on investment [ROI], Procedia Engineering 172 (2017) 1227–1234, https://doi.org/10.1016/j.proeng.2017.02.144.

[4] J. Neelamkavil, S. Ahamed, The Return on Investment from BIM-driven Projects in Construction, (2012), https://doi.org/10.4224/20374669.

[5] J. JR, BIM investment: Understanding value, return and models of assessment, AUBEA2013, (2014), pp. 1–19 http://hdl.handle.net/10453/28577.

[6] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32, https://doi.org/10.1023/A:1010933404324.

[7] A. Shtub, J.F. Bard, S. Globerson, Project Management: Processes, Methodologies, and Economics, Pearson Education Limited, Harlow, UK, 2013 ISBN: 978-1-292- 03940-4.

[8] M. Nemtsas, M. Nemtsas, TimesheetsLite, Timesheets MTS Software, (2010)https://www.timesheetsmts.com/timesheet-software-weblog/2010/07/timesheets-lite-version-234-re.html Accessed date: January 2019.

[9] C. H. Huang, S.-H. Hsieh, 17th International Conference on Construction Engineering and Project Management, Busan, Korea, 2020, https://doi.org/10. 13140/RG.2.1.3821.6403 (n.d.).

[10] T. Kononenko, M. Kukar, Machine Learning and Data Mining Introduction to Principles and Algorithms, Woodhead Publishing, Oxford, 2007 ISBN: 978-1-404275-21-3.

[11] I.H. Witten, E. Frank, M.A. Hall, C.J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, Amsterdam, 2017 ISBN: 978-1-904275- 21-3.

[12] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65, https://doi.org/10.1016/ 0377-0427(87)90125-7.