Analysis on The Influence Factors of Transmission Line Project Cost Based on Random Forest Model

Li Ma\textsuperscript{1, \*}, Ming Zhou\textsuperscript{1}, Shengwei Lu\textsuperscript{1}, Qixin Wang\textsuperscript{1}, Shuang Liao\textsuperscript{1}, Yue Ming\textsuperscript{1}, Sicong Wang\textsuperscript{2}

\textsuperscript{1}State Grid HuBeI Electric Power Company Limited Economic
\textsuperscript{2}State Grid HuBeI Electric Power Company Limited Economic Research Institute

*Corresponding author: mal12@hb.sgcc.com.cn

Abstract. There are many factors influencing the cost of transmission line project. The complex construction environment, different construction level and other factors could lead to large differences in the cost of comparable transmission line projects under the same voltage level. Based on the division of the transmission line projects, this paper analyzes and identifies the main factors affecting the cost of transmission line project. The random forest model is used to select 15 of the influencing factors. And then, this paper verifies the main factors selected, providing references for prediction.

Keywords: Transmission line, random forest, project cost.

1. Introduction
In the construction of power transmission and transformation projects, the management and control of the project cost plays an important role for the entire project management. Analyzing the cost level, structural fluctuation, influencing factors and development trends, and realizing effective control of cost, are of great significance to rationally verify transmission costs, increase grid investment efficiency, and improve project investment management. However, due to many factors affecting project cost, the complex construction environment, and different construction level, etc., there are large differences in the cost of comparable power transmission and transformation projects under the same voltage level. In this way, for a certain project, the number of samples of comparable transmission and transformation project cost is relatively less. And it is difficult for auditors to review and estimate the project cost based on experiences and conventional statistical estimation models. Therefore, according to the situation with many influencing factors, how to find the main technical indicators concerning cost estimation of power transmission and transformation projects via attribute selecting method in data mining is practically important for the cost management.

2. Literature Review
Among the many factors that affect the cost of power transmission and transformation projects, those factors whose small changes will cause significant changes in investment costs are called the main factors \cite{1}. The mining of the main factors needs to be based on the correlation between quantity, price, and cost. It is necessary to deeply explore the structure and fluctuation law of the cost through the analysis of the overall cost and sub-items cost of the power transmission and transformation project.
costs [2]. Zhiyuan Hu [3], Shengguang Miao [4] and others combined the characteristics of overhead line engineering cost, and theoretically analyzed the sensitive factors of it. Analyzed the influence degree on the project cost of these sensitive factors under different terrain, different transport distances and different geological conditions.

The study found that the factors that are most closely related to the cost of power grid projects include: equipment prices, main material prices, and construction area acquisition costs, etc.,[5] Jinlan Hu [6] and others comprehensively used principal component analysis and multiple regression model to systematically research the investment influencing factors of power engineering. On this basis, they constructed an investment prediction and analysis model of power project based on BP neural network model and optimized by particle swarm algorithm, and carried out relevant applications. Principal component analysis is a widely used method in factor analysis, and many scholars apply it to the analysis of factors affecting power engineering cost [7-9].

3. Model and Methodology
Domestic scholars have done a lot of research on the analysis of sensitive factors in the cost of transmission line projects. They applied a method combining qualitative and quantitative, mainly including analytic hierarchy process, principal component analysis, factor analysis, multiple regression model and random forest model, etc. In particular, random forest model has a better influence on the factor selection of transmission line engineering. This paper takes the line engineering in the transmission and transformation project as the research object, applying the random forest model to select the factors that affect the line engineering cost, and using the predicted results to compare with the actual results.

The Random Forest (RF) belongs to an algorithm of machine learning in data mining. The algorithm is coming up with by Breiman, based on CART algorithm, referring to the idea of random decision forest. In the process of classification tree generated by machine learning, it applies the random combination of row variables and column variables to generate many random classification trees. And then these trees are aggregated into a forest to form a random forest. Random forest improves the prediction accuracy without significantly increasing the amount of calculation. It is an algorithm especially designed for the classification and estimation of high dimensional and small sample size data, so that it can explain the effect of attributes on dependent variables among thousands of attributes. Hence, it is widely applied in the estimation of high dimensional and small size sample.

Random forest generally uses gini values as the criteria for segmenting nodes. In weighted random forests, the essence of weights is to assign larger weights to small categories and smaller weights to large categories, that is, to give smaller categories big punishment. The weight has two functions. The first function is to get the weighted calculation of gini value in the process of segmenting nodes selection. The expression is as follows:

\[
i(N) = \frac{\sum_{i=1}^{c} (n_i W_i)^2}{\sum_{i=1}^{c} n_i W_i}
\]

\[
\Delta i = i(N_L) - i(N_R)
\]

In this expression above, N represents the unseparated node. NL and NR represent the separated left and right nodes respectively. Wi is the classification weight of class C samples. ni is the number of classification samples in the node, and \(\Delta i\) is the reduction in impurity. The larger the value, the better the separation effect of the separation point.

The second function is at the end node while the classification weight is used to determine its classification label. The expression is as follows:
On the basis of characteristic importance, the steps of characteristics selection are as follows:

1. Calculate the importance of each characteristic and sort it in descending order;
2. Determine the proportion to be eliminated, and eliminate the corresponding proportion of characteristics based on their importance so as to obtain a new feature set;
3. Repeat the process above with the new feature set until there are \( m \) characteristics are left (\( m \) is the value set in advance);
4. According to each feature set obtained above and out-of-bag error rate corresponding to the feature set, select the feature set with the lowest out-of-bag error rate.

4. Analysis

4.1. Theoretical Analysis

Combined with the division of overhead power transmission line projects in the "Budget Preparation and Calculation Regulations for Power Grid Project Construction (2013 Edition)" issued by the National Energy Administration, the cost of the overhead power transmission line project itself consists of six-unit project, including basic project, tower project, grounding project, stringing project, accessory project and auxiliary project. Through reading relevant literature and analyzing case studies, it is found that the main factors affecting these projects are as follows:

1. Terrain. Different terrain will cause different costs. Common terrains for UHV overhead lines include flat land, hills, rivers, sludge, mountains and deserts. The more complex the terrain, the greater the difficulty of construction, and then the higher the project costs.
2. Wind speed. Under different wind speeds, the load of the transmission line tower will be different relatively, which will cause the differences in the weight of the tower, thereby affecting the project costs.
3. Icing. The size of icing has a greater impact on the design of tower weight. The increasing load caused by icing has higher requirements for the design of tower, and then the tower weight will also increase correspondingly. In addition, the conductor selection will also be impacted when the icing is heavy.
4. Geology. Under the conditions composed of ordinary soil, hard soil, loose sand, rock, muddy water, quicksand, dry sand, puddles, etc., different basic types are adopted.
5. Transport distance. The construction of overhead line projects requires a lot of building materials and facilities, and the cost of transporting these items to the construction site is directly related to the transport distance. In some places with more dangerous terrain, the use of transportation equipment is also subject to certain restrictions. At this time, it needs to be operated manually, and the cost of labor is much higher than the cost of mechanical equipment.
6. Wire type. Wire selection is mainly determined by transmission capacity, which has been determined in early planning, so the cost in the design stage is mainly affected by the price of wire material.
7. Stringing mode. The stringing mode plays a decisive role on the cost of UHV overhead transmission line projects. There are three different types: wire type, tower type and basic type. As for wire type, different wire types have different transmission capacity. Generally, the larger the transmission capacity, the higher the wire price. In fact, the transmission capacity will not change easily due to the fact that this has been determined in early planning. Therefore, it is not meaningful to study it. For the tower type, because the steel consumption of iron tower accounts for a large proportion of the tower materials required for the entire project, reaching 26% to 39%, the larger the steel consumption, the heavier the tower weight, and then the project costs are higher.
8. Basic type. Different basic types will affect the amount of earth and stone work, concrete, and steel. The construction and material cost directly affect the cost of foundation projects.

In summary, there are many factors affecting the cost of UHV overhead line projects, but these factors are not unrelated. For example, wind speed, icing, and ground wire weight will affect the steel
consumption of tower. The tower weight has an impact on fundamental power, which in turn affects the amount of foundation works. In this way, there is a correlation among factors such as wind speed, icing, ground wire weight, steel consumption of tower, amount of concrete and steel and so on. Therefore, studying the main influencing factors of overhead line project cost is of great significance to manage and control the entire project.

4.2. Empirical Analysis

4.2.1. Data collection and processing. The original data in this article stems from 45 groups of 110kV transmission line projects completed by a research institute in 2016-2018. We collect 16 indicators, including transmission line project cost(y), line length(x1), terrain(x2), tower cardinality(x3), tensile ratio(x4), concrete volume(x5), amount of steel(x6), amount of earthwork and stonework(x7), insulators(x8), turret cardinality(x9), amount of turret material(x10), turret material unit price(x11), amount of wire(x12), wire unit price(x13), OPGW length(x14), and OPGW unit price(x15) etc., Among them, the project cost is the dependent variable, and the others are independent variables.

As the terrain is a descriptive factor among the 15 influencing factors identified, it needs to be quantified before analysis. The method is as follows:

The terrain of the transmission line route includes flat land, hills, rivers, sludge, and mountains. They are divided into 5 levels in order and represented by 1, 2, 3, 4, and 5 respectively. Specifically, flat land is the first level and mountain are the fifth level. If a transmission line route contains multiple terrains, and then it is necessary to perform weighted average on the terrain covered according to the proportion of different terrains. The original data is shown below:

| Sample | (X1), km | (X2), unit | (X3), % | (X4), % | (X5), m³ | (X6), t | (X7), m³ | (X8), unit |
|--------|----------|------------|--------|---------|---------|--------|---------|----------|
| Line 1 | 12.00    | 2.3        | 58     | 26.00   | 2204.49 | 142.00 | 2832.04 | 450      |
| Line 2 | 19.90    | 2.1        | 69     | 18.84   | 2434.72 | 188.00 | 9231.70 | 347      |
| Line 3 | 2.23     | 1.8        | 6      | 83.00   | 389.66  | 27.08  | 474.24  | 936      |
| Line 4 | 3.30     | 1.8        | 9      | 44.00   | 393.67  | 22.64  | 1253.49 | 286      |
| Line 5 | 21.30    | 3.5        | 63     | 32.00   | 1451.75 | 52.00  | 2615.38 | 482      |
| ...    | ...      | ...        | ...    | ...     | ...     | ...    | ...     | ...      |
| Line 45| 11.90    | 1.9        | 33     | 46.80   | 1275.58 | 61.58  | 7956.34 | 172      |

| Sample | (X9), unit | (X10), t | (X11), yuan/t | (X12), t | (X13), yuan/t | (X14), km | (X15), yuan/km | (Y), Ten thousand yuan |
|--------|------------|----------|---------------|----------|---------------|-----------|-----------------|------------------------|
| Line 1 | 161        | 160.85   | 6926          | 35.00    | 21650         | 12.00     | 17000           | 410                    |
| Line 2 | 301        | 300.95   | 7543          | 96.00    | 21200         | 8.50      | 20000           | 990                    |
| Line 3 | 69         | 68.58    | 6810          | 7.74     | 13105         | 7.54      | 17038           | 1172                   |
| Line 4 | 82         | 82.49    | 6810          | 11.37    | 13105         | 2.30      | 17038           | 501                    |
| Line 5 | 366        | 365.77   | 7200          | 65.00    | 20000         | 9.50      | 20000           | 226                    |
| ...    | ...        | ...      | ...           | ...      | ...           | ...       | ...             | ...                    |
| Line 45| 24         | 270.21   | 5875          | 47.64    | 12585         | 10.30     | 17102           | 522                    |
4.2.2. **Factor selecting based on Random Forest.** Use the random forest model above to filter attributes. Select the project cost as the decision variable, and the others as the attribute variables to form a preliminary data set. Use the randomForest package in R to analyze, and the gini index of each attribute can be obtained. The results are shown in Figure 1.

![Image of importance](image)

**Figure. 1** Variable importance score

In Figure 1, IncNodePurity represents the purity of the node, which has the same meaning as the gini index. The larger the value, the higher the importance of the attribute. Among the influencing factors, there are 10 indicators whose IncNodePurity index exceeds $4 \times 10^5$, including tensile ratio, line length, OPGW length, tower cardinality, terrain, amount of turret material, turret material unit price, wire unit price, amount of earthwork and stonework, and concrete volume.

4.2.3. **Factor verification based on Random Forest.** Using 10 main influencing factors selected above, 10 transmission line projects sample data are selected. Take the project cost as the decisive variable, and the prediction results and error rates obtained are shown in Figure 2.

![Image of error rates](image)

**Figure. 2** Error rates between predicted value and actual value
It can be seen from Figure 2 that the line of predicted results obtained by those 10 selected main indicators is very close to the line of actual values corresponding to all indicators. Fortunately, the error rates are controlled within 10%. The results show that the main influencing factors of cost selected by random forest can explain all indicators to a large extent, indicating that these 10 main indicators can more accurately predict the cost of transmission line projects.

5. Conclusions
This paper uses the random forest model to study selection problem of factors affecting the cost of transmission line projects in the case of high dimensional and small size samples. In view of the multi-factor and high-dimensional situation, the random forest method is used to screen out the main factors, and the indicators are reduced in dimension. At the same time, the main influencing factors selected including tensile ratio, line length, OPGW length, tower cardinality, terrain, amount of steel, turret material unit price, wire unit price, earthwork volume, concrete volume, etc., can predict the cost of transmission line projects effectively. With the enrichment and improvement of transmission line project data, the accuracy of the analysis of factors affecting the cost and the cost forecasting model will be further improved. The application of the model can provide some references for scientific analysis of the reasonable level of cost, and improvement of investment control ability.

References
[1] Jun Cao, Yawen Wang, Xiankai Zeng. Statistical Analysis of Sichuan Power Grid Project Cost during the "Twelfth Five-Year Plan"[J]. Telecom World, 2013(23):174-176.
[2] Analysis Report on Transmission and Transformation Project Cost of State Grid Corporation of China (2011 Edition) [R]. Beijing: State Grid Beijing Economic and Technological Research Institute, 2011.
[3] Zhiyuan Hu. Analysis on Sensitive Factors of UHV Overhead Transmission Line Project Cost [J]. Theoretical Research in Urban Construction, 2017(19): 9.
[4] Shengguang Miao, Kai Wen. Analysis on Sensitive Factors of UHV Overhead Transmission Line Project Cost [J]. Chinese Consulting Engineers, 2016(5):62-65.
[5] Wenxiu Xu, Libin Zhang, Ke Lv and others. Research on Cost Management and Control of Transmission and Transformation Projects Based on the Concept of Whole Process [J]. China Science and Technology Expo, 2015(8): 203-203.
[6] Jinlan Hu, Xiaobin Gao, Chongming Liu and others. Empirical Study on Influencing Factors of Power Grid Project Investment Based on BP Neural Network [J].Engineering Economy, 2017, 27(10): 15-19.
[7] Yongmei Ji, Dongping Wu, Ying Peng and others. Analysis on Factors Affecting the Cost of Substation Project Based on SPSS Principal Component Analysis Method [J]. Management and Administration, 2014(2): 125-130.
[8] Jinlan Hu, Xiaobin Gao. Identification of Factors Affecting the Cost of Transmission and Transformation Projects and Data Modeling Analysis [J]. Journal of North China Electric Power University (Social Sciences), 2016(5):55-59.
[9] Junye Xiao, Xiaoling Hao, Xiaoqian Tang and others. Factors Influencing Power Engineering Cost Based on Principal Component Analysis [J]. Management and Administration, 2014(3): 128-133.