Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 1649086, 11 pages
https://doi.org/10.1155/2019/1649086

*Research Article*

# Research on Prediction Method of Reasonable Cost Level of Transmission Line Project Based on PCA-LSSVM-KDE

**Zhao Xue-hua,[1] Miao Xu-juan,[1] Zhang Zhen-gang,[2] and Hao Zheng[3]**

[1]*Economy and Technology Research Institute, State Grid Xin Jiang Electric Power Corporation, Wulumuqi 830011, China*
[2]*State Grid Handan Electric Power Supply Company, Handan 056035, China*
[3]*School of Economics and Management, North China Electric Power University, Beijing 102206, China*

Correspondence should be addressed to Hao Zheng; 18811358655@163.com

Received 6 March 2019; Accepted 18 July 2019; Published 1 August 2019

Academic Editor: Javier Martinez Torres



In order to reduce the investment risk, the evaluation standard of transmission line project investment planning becomes higher, which puts forward higher requirements for the reasonable level prediction of transmission line project cost. This paper combines principal component analysis (PCA) with the least squares support vector machine (LSSVM) model and establishes a point prediction model for transmission line project cost. Based on the analysis of the error of the point prediction model, the kernel density estimation (KDE) method is innovatively introduced to estimate the prediction error, and the probability density function of the error is obtained. Then, according to different confidence levels, the corresponding cost intervals are obtained, which means that the reasonable level of transmission line project cost is obtained. The results show that the coverage rate of the cost prediction interval under 85% confidence level is 88.57%. This conclusion shows that the model has high reliability and can provide a reliable basis for the evaluation of transmission line project investment planning.

## 1. Introduction

With the rapid development of national economy, the demand for power energy is increasing. Transmission line project is an important part of power grid construction, and rational evaluation of its investment planning is an important part of cost control. At present, cost control line [1] and general cost [2] are mostly used as evaluation criteria in power industry, and a specific value is given, which makes the evaluation results less compatible. Therefore, in order to make a reasonable evaluation of investment planning and determine the reasonable cost level of transmission line projects, reasonable cost intervals should be given on the basis of specific cost control lines.

There are many factors affecting the transmission line projects' cost, which have the characteristics of randomness and instability. However, the general point prediction results cannot represent the variability of the transmission line project cost, and the information provided by them is often insufficient to meet the requirements of investment decision-making, which brings risks to the decision-making work. If the deterministic point prediction results can be given, and the fluctuation range of cost can be described at the same time, it will be helpful for power enterprises to make more reasonable investment decisions and make more reasonable investment planning evaluation. Therefore, the purpose of this paper is to study the interval prediction method of transmission line project cost, get the fluctuation interval of the cost at a certain confidence level, and obtain the prediction results in the sense of probability. The results of probabilistic prediction can provide more valuable information to decision makers, help them better grasp the changes of data, and also help power enterprises to make investment planning, risk analysis, and reliability evaluation.

With the development of machine learning and intelligent algorithm, the research of project cost point prediction has developed rapidly, and the prediction accuracy has greatly been improved. Ji and Abourizk constructed a special absorption Markov chain, which takes into account the uncertainty of rework caused by quality, and

stochastically modeled the manufacturing process of building products so as to estimate and control the rework cost caused by quality [3]. Lesniak and Juszczyk established a regression model based on the artificial neural network to estimate field management cost quickly and reliably [4]. Bhargava et al. introduced a risk-based polynomial model and Monte Carlo simulation to predict that the project will follow a specific cost increase path in its development phase and will produce a given level of cost deviation severity [5]. Lesniak and Zima proposed a case-based reasoning method for estimating the construction cost of sports venues and CBR method based on historical data and sustainable development criteria was used to estimate the initial cost of construction projects [6]. Juszczyk et al. put forward a method of predicting the construction cost of stadium based on neural network and evaluated its prediction quality and accuracy [7].

The support vector machine (SVM), proposed by Vapnik [8], is an effective method based on statistical learning theory. The algorithm does not use the principle of minimum empirical risk to minimize the training error but is based on the principle of structural risk minimization to minimize the upper limit of generalization error so that the global optimal solution can be obtained theoretically [9, 10]. Least squares support vector machine (LSSVM) can effectively simplify computational complexity and improve operational efficiency by changing inequality constraints to equality constraints [11, 12], which makes it have great advantages in multifactor prediction. Liang et al. proposed a hybrid model based on the wavelet transform (WT) and LSSVM and optimized it with improved cuckoo search (CS) so as to achieve accurate load forecasting [13]. Kang et al. proposed a hybrid ensemble empirical mode decomposition (EEMD) and LSSVM methods to improve the accuracy of short-term wind speed prediction [14].

The research on the prediction method of engineering cost has developed rapidly, and many scholars have carried out in-depth research on the prediction of electric power engineering cost. Kong et al. established a cost prediction model of transmission and transformation project based on SVM which used SVM to solve the regression equation, and then the cost was predicted by the model [15]. Wang took the comprehensive cost index of transmission and transformation as the basis of project investment decision-making and probed into the establishment of a model for project investment cost prediction by using the Markov chain [16]. Lu et al. took full account of the characteristics of subitem cost and adopted different methods to forecast separately and then superimposed to get the total cost [17]. Wang et al. established the EEMD-BP model and used BP algorithm to forecast the trend components, and the final prediction results were obtained by considering the prediction values of the trend components and the fluctuation intervals [18]. Yi et al. evaluated the global sensitivity of input variables and proposed a neural network prediction method of transmission line project cost based on feedforward and postpropagation multilayer perception structure [19]. Wang et al. established the REGR-WNN prediction model and compared it with REGR and WNN models separately, and the prediction accuracy of this method is higher [20].

Some scholars have done some research in the area of interval prediction. Grounds et al. believed that prediction intervals show a series of values with specific probability and have the potential to improve decision-making compared with point prediction, so interval prediction may have important benefits [21]. Samal and Tripathy used a nonparametric kernel density method to express wind speed measurements that were not suitable for parameter distribution and used the chi-square test and Kolmogorov–Smirnov goodness-of-fit test to evaluate their applicability [22]. Fan et al. used nuclear density estimation to express the maximum allowable temperature in a period of time on the basis of estimating the instantaneous state real-time thermal rating (RTTR) when predicting the RTTR of overhead lines [23]. Amara et al. discussed the influence of temperature on the nonlinear relationship of power demand and proposed an adaptive conditional density estimation (ACDE) method based on kernel density estimation (KDE) to improve the accuracy of load forecasting [24].

## 2. Research Method

*2.1. Principal Component Analysis (PCA).* Principal component analysis (PCA) is a method of transforming a set of variables that may be correlated into a series of linear irrelevant variables by orthogonal transformation. The transformed variables are called principal components, which can eliminate the multiple collinearities between data. PCA can be divided into the following steps:

(1) The original index data are a $p$-dimensional random vector $x = (x_1, x_2, \ldots, x_p)^{\mathrm{T}}$. Standard transformation of $n$ samples $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^{\mathrm{T}}$ is carried out: $Z_{ij} = (x_{ij} - \overline{x}_j)/s_j$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$ where $\overline{x}_j = (\sum_{i=1}^{n} x_{ij})/n$ and $s_j^2 = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2/(n-1)$, and the normalized matrix $Z$ is obtained.

(2) The normalized matrix is calculated, and the sample correlation coefficient matrix $R = (Z^{\mathrm{T}}Z)/(n-1)$ is obtained.

(3) The characteristic equation $|R - y_p| = 0$ of $R$ is solved, and $p$ characteristic roots are obtained, and then the number of principal components is determined so as to ensure that the cumulative contribution rate of principal components can exceed 85%. For each $y_p$, the unit eigenvector $b$ can be obtained by solving the system of equations $Rb = y_p b$.

(4) The standardized index variables are transformed into the main component $U_{ij} = Z_i^{\mathrm{T}} b_j$, $j = 1, 2, \ldots, m$.

(5) PCA extracts $p$ totally new and unrelated variables by concentrating $p$ observation variables as follows:

Before PCA,

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \left(x_1, x_2, \ldots, x_p\right). \quad (1)$$

After PCA,

$$\begin{cases} F_1 = a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p, \\ F_2 = a_{21}x_1 + a_{22}x_2 + \cdots + a_{2p}x_p, \\ \vdots \\ F_p = a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pp}x_p. \end{cases} \quad (2)$$

*2.2. Least Squares Support Vector Machine (LSSVM).* Although the traditional support vector machine (SVM) is good enough to avoid falling into the shortcomings of the local optimal solution, it will prolong the running time of the computer if the capacity of the data set is large because the SVM uses quadratic programming in the process of solving. Therefore, Suykens proposed the least squares support vector machine (LSSVM); that is, LSSVM is the improvement and perfection of the support vector machine [25]. LSSVM is characterized by solving linear equations rather than quadratic programming problems, which can simplify the calculation process and reduce the solving time effectively. As a result, its application has become more and more widespread. Regression algorithms for LSSVM are described as follows:

Given the training set $D = \{(x_i, y_i)\}_{i=1}^{n}$, $x_i \in X \subset R^m$, $y_i \in Y \subset R$, $i = 1, 2, \ldots, n$, where $x_i$ represents the $m$-dimensional input vector and $y_i$ represents the output vector corresponding to $x_i$. The regression model is constructed from the nonlinear mapping function as follows:

$$f(x) = \omega^{\mathrm{T}}\varphi(x) + b, \quad (3)$$

where $\omega$ represents the weight vector and $b$ represents the offset. According to the model complexity and fitting error, the objective function of the LSSVM algorithm is as follows:

$$J(\omega, \xi) = \frac{1}{2}\omega^{\mathrm{T}}\omega + \frac{\gamma}{2}\sum_{i=1}^{n}\xi_i^2. \quad (4)$$

The constraint condition is

$$y_i = \omega^{\mathrm{T}}\varphi(x_i) + b + \xi_i, \quad i = 1, 2, \ldots, n, \quad (5)$$

where $\xi_i$ represents regression error and $\gamma$ represents the penalty coefficient. The function of $s$ is to adjust the error, and the larger the value of $\gamma$, the smaller the corresponding regression error will be. For the parameter $\gamma$ in the model to be optimized, the Lagrange function is

$$L(\omega, b, \alpha, \xi) = J(\omega, \xi) - \sum_{i=1}^{n}\alpha_i\left(\omega^{\mathrm{T}}\varphi(x_i) + b + \xi_i - y_i\right), \quad (6)$$

where $s$ denotes the Lagrange multiplier. According to the quadratic programming KKT condition, the following results are obtained:

$$\begin{aligned} \omega &= \sum_{i=1}^{n}\alpha_i\varphi(x_i), \\ \sum_{i=1}^{n}\alpha_i &= 0, \\ \alpha_i &= \gamma\xi_i, \\ \omega^{\mathrm{T}}\varphi(x_i) + b + \xi_i - y_i &= 0. \end{aligned} \quad (7)$$

After eliminating $\omega$ and $\xi_i$, the final matrix linear equations are obtained from the Mercer condition $K = \varphi(x_k)^{\mathrm{T}}\varphi(x_j) = K(x_k, x_j)$, $k, j = 1, 2, \ldots, n$ as follows:

$$\begin{bmatrix} 0 & 1_n^{\mathrm{T}} \\ 1_n & K + \gamma^{-1}I_n \end{bmatrix}\begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \quad (8)$$

where $1_n = [1, 1, \ldots, 1]^{\mathrm{T}}$ and $I_n$ is an $n * n$ unit matrix, and the LSSVM regression function is obtained by solving the final equations as follows:

$$f(x) = \sum_{i=1}^{l}\alpha_i K(x, x_i) + b. \quad (9)$$

In this paper, the radial basis function is chosen as the kernel function:

$$K(x, x_i) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right). \quad (10)$$

*2.3. Interval Prediction Theory Based on Kernel Density Estimation (KDE).* In point prediction of transmission line project cost, the predicted value $y_i$ is only the approximate value of the real value $\hat{y}_i$, and the probability that the predicted value is exactly equal to the real value is very small. In order to ensure the reliability of investment decision-making, it is necessary to estimate an approximate range of the predicted value and how much credibility (or confidence level) the range covers the real value. This range of variation is generally expressed by intervals, known as confidence intervals. When the number of samples and confidence level remain unchanged, the length of the confidence interval is inversely proportional to the accuracy of interval estimation. Confidence intervals generally have a two-sided confidence interval $(a, b)$ and one-sided confidence interval $(-\infty, \mathrm{a})$ and $(b, +\infty)$. The confidence interval calculated in this paper is a two-sided confidence interval, which is defined as follows.

Let $X = (x_1, x_2, \ldots, x_n)$ be a sample of the population $F(X)$, and the random interval $[\theta_{\mathrm{L}}, \theta_{\mathrm{U}}]$ of statistics $\theta_{\mathrm{L}}(X)$ and $\theta_{\mathrm{U}}(X)$ is called an interval estimate of $\theta_i$. For a given confidence level $1 - \alpha (0 < \alpha < 1)$, satisfy the following equation:

$$P\left(\theta_{\mathrm{L}} < \theta_i < \theta_{\mathrm{U}}\right) = 1 - \alpha, \quad (11)$$

where $\theta_L$ and $\theta_U$ are the lower and upper confidence limits of the error values, respectively. Interval $[\theta_L, \theta_U]$ is called the confidence interval under the confidence level $1-\alpha$. This paper uses the equal tail confidence interval:

$$P(\theta_i < \theta_L) = \frac{\alpha}{2},$$
$$P(\theta_i > \theta_U) = \frac{\alpha}{2}. \tag{12}$$

The probabilistic density function estimation problem is to estimate its probability density function through samples. There are usually parametric, semiparametric, and nonparametric estimation methods. Nonparametric density estimation only takes the data of the sample itself as the basis of probability density function estimation. It does not need to make assumptions about the form of sample distribution beforehand and can deal with arbitrary density distribution. The commonly used nonparametric density estimation methods include histogram density estimation method and kernel density estimation method. Although the concept of the histogram density estimation method is simple and easy to use, the result is discontinuous; that is, the density estimation value at the boundary of the region will drop to 0 suddenly, and the efficiency is low. Kernel density estimate (KDE) is proposed by Parzen, also known as Parzen window estimation. And it is a very effective nonparametric density estimation method [26]. Its general expression is as follows:

$$f(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{13}$$

where $n$ is the total number of samples, $h$ is the bandwidth or smoothing parameter, $X_i$ is the given sample, and $K(\cdot)$ is the kernel function, satisfying the following conditions:

$$K(t) \geq 0,$$
$$\int_{-\infty}^{+\infty} K(t)dt = 1,$$
$$\sup K(t) < +\infty,$$
$$\int_{-\infty}^{+\infty} K^2(t)dt < +\infty,$$
$$\lim_{x\longrightarrow\infty} K(t)t = 0. \tag{14}$$

KDE can be regarded as the integration of forms centered on each observation sample point. Its performance depends on the selection of the kernel function and window width. If the selection of window width is too large, some characteristics of distribution will be concealed and excessive averaging will make the estimator deviate greatly; if the selection of window width is too small, the whole estimation, especially the tail, will be disturbed greatly, thus increasing the variance trend.

This paper uses relative error to define the deviation between the predicted value and the actual value of construction cost:

$$e_i = \frac{y_i - \widehat{y}_i}{\widehat{y}_i}. \tag{15}$$

The error probability density function can be obtained by nonparametric kernel density estimation, and then the cumulative probability distribution function of relative error (as a random variable of error) can be obtained by integral shown as follows:

$$F(\varepsilon) = \int_{-\infty}^{\varepsilon} f(e)de. \tag{16}$$

According to the cumulative probability distribution function of the error and the point prediction value of the sample, the confidence interval with confidence level $1-\alpha$ can be obtained as follows:

$$\left[\frac{y_i}{1+\widetilde{F}(\alpha_2)}, \frac{y_i}{1+\widetilde{F}(\alpha_1)}\right], \tag{17}$$

where $\alpha_1 = \alpha/2$, $\alpha_2 = 1-(\alpha/2)$, and $\widetilde{F}(\cdot)$ is the inverse function of $F(\varepsilon)$.

The main research ideas and methods of this paper are shown in Figure 1. Firstly, this paper considers many impact indexes of transmission line project cost and extracts and screens the indexes. Secondly, PCA is used to reduce the dimension of the original index, and the corresponding principal component is used as the input variable of the prediction model. Then, a point prediction model for the transmission line project is constructed based on LSSVM. Finally, the probability density function of the prediction model error is obtained by KDE, and then the cost interval at a certain confidence level is obtained.

## 3. Selection of Cost Indexes and Data Source

*3.1. Analysis of Influencing Factors of Cost and Screening of Indexes.* The purpose of this paper is to study the interval prediction method of transmission line project cost and to guide the investment decision of transmission line project. The engineering characteristics of transmission line projects play a decisive role in project cost. Therefore, in order to predict the transmission line cost, it is necessary to comprehensively analyze the factors affecting the cost, and select the engineering characteristics that have a greater impact on the transmission line cost as the indexes affecting the transmission line cost, combining with the practical experience of transmission line project.

Transmission line project can be divided into six major units, namely, foundation project, tower project, erecting engineering, grounding project, annex project, and auxiliary project. Therefore, when identifying the influencing indexes of transmission line project cost, this paper firstly analyzes the six major units of the project and identifies their main influencing indexes. For the cost of each unit project, combined with the engineering practice experience, the engineering characteristics related to the site where the project is located, large quantities of works and high price of materials and other indexes which have a greater impact on the cost are selected as the analysis object. After classifying and summarizing, the influencing indexes of the overall cost are obtained, and the specific identification process is shown in Figure 2.

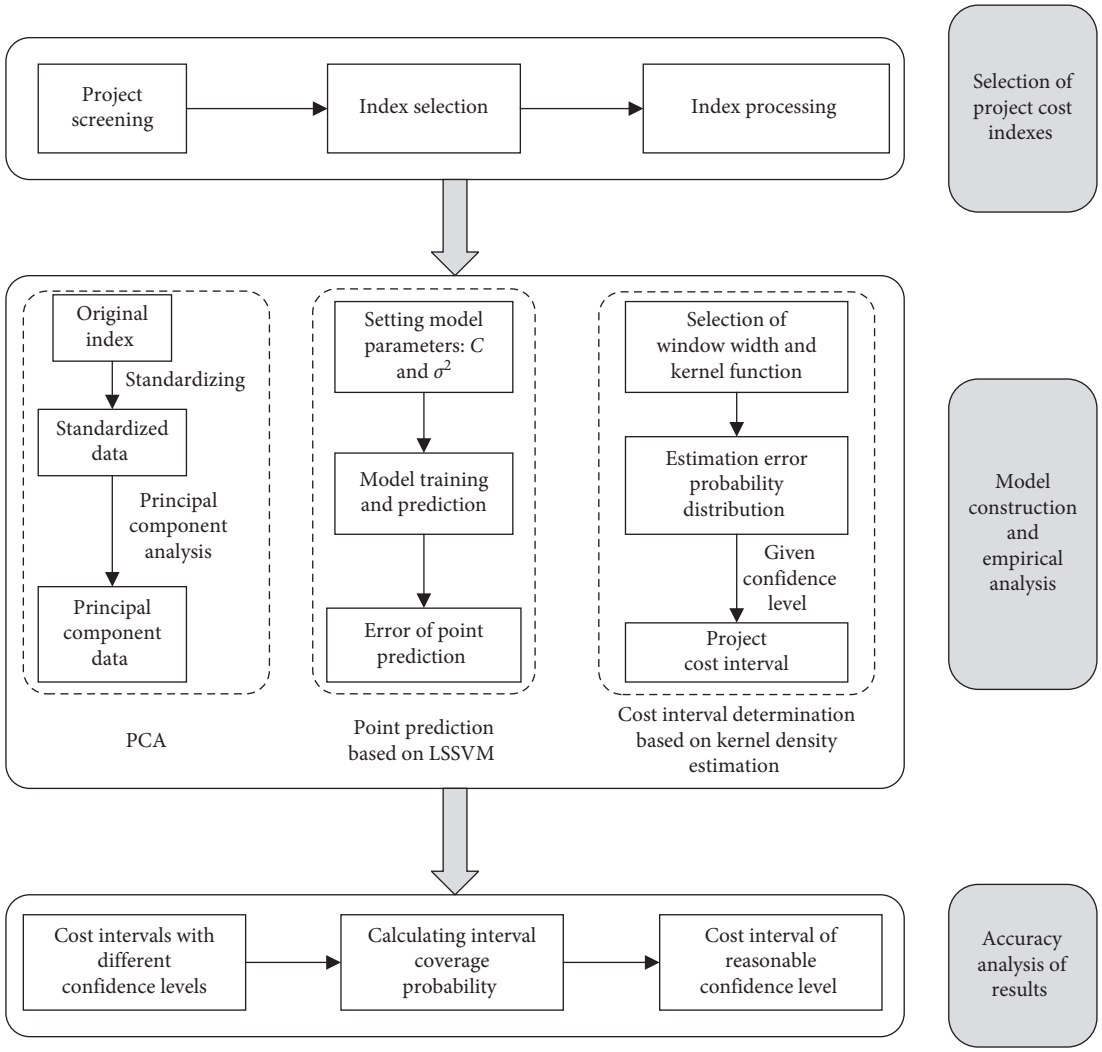


Figure 1: Flow chart of the prediction model.

In Figure 2, some indexes have an impact on different unit project cost, so the cost impact indexes in Figure 2 are summarized and divided into natural indexes, technical indexes, and economic indexes as shown in Table 1.

*3.2. Sample Data Source.* This paper collects the actual data of transmission line project for model empirical analysis. According to the settlement data of transmission line projects completed and put into operation in Xinjiang by State Grid Corporation in 2017, 140 representative projects under 110 kV voltage level are selected. The original data samples in this paper are obtained through index processing [27], as shown in Table 2.

## 4. Construction of Cost Interval and Empirical Calculation

*4.1. Dimension Reduction of Cost Impact Indexes.* There may be multiple collinearities among variables, so this paper makes principal component analysis of sample data to exclude the influence of correlation among variables and reduce the number of variables at the same time. Before principal component analysis, the KMO test and Bartlett spherical test were used to study the correlation between variables. The KMO value of 13 indexes is 0.679, and the significance level of the Bartlett spherical test is far less than 0.01. The test results show that the principal component analysis is feasible.

After principal component analysis, aggregated indexes that can express most of the information instead of the original indicators are selected. These aggregated indexes have almost the same function as the original indexes and can meet the needs of analysis. In addition, the selection of these aggregated indexes reduces the number of input indexes of the prediction model, reduces the complexity of the prediction model, and improves the running speed of the model. This paper argues that the aggregated indexes with cumulative variance over 85% can express the vast majority of the overall information. After SPSS dimensionality reduction, Table 3 gives the percentage and cumulative percentage of the total information explained by the

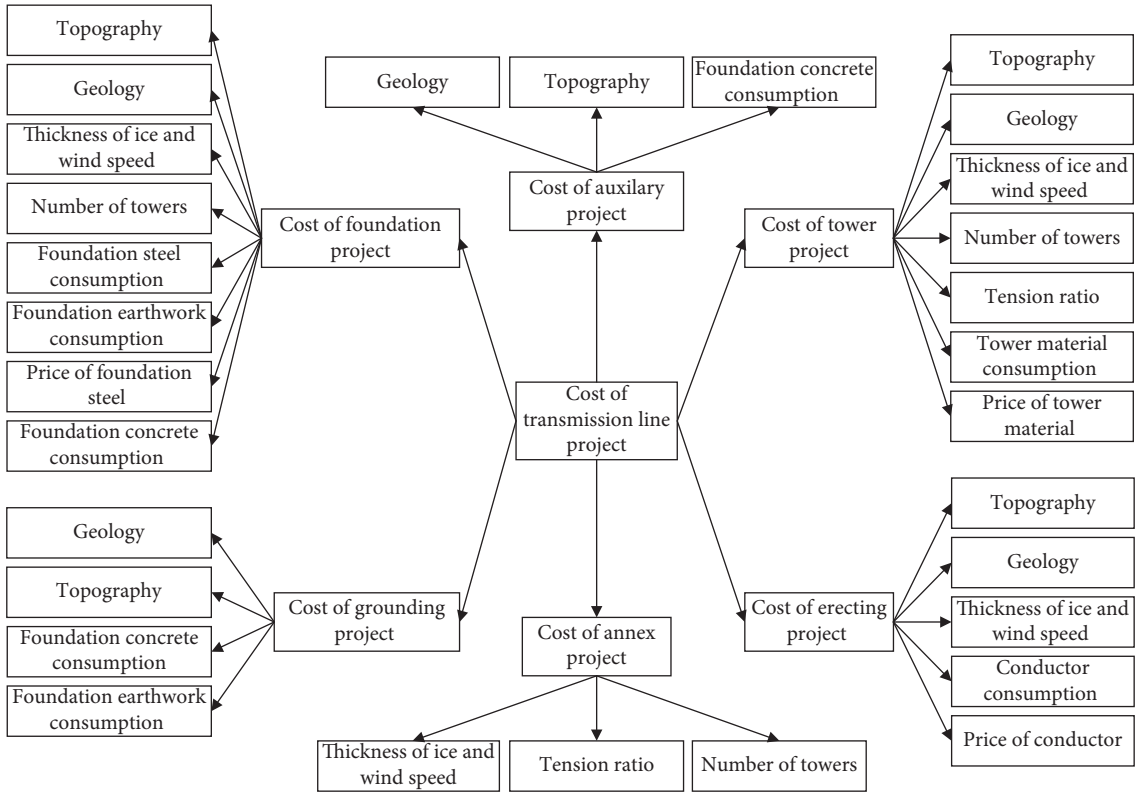


Figure 2: Selection of cost influencing indexes.

Table 1: Influencing indexes of transmission line project cost.

| Index type | Influencing indexes | Number |
|---|---|---|
| Natural indexes | Topography | $X1$ |
| | Geology | $X2$ |
| | The thickness of ice and wind speed | $X3$ |
| Technical indexes | Conductor consumption | $X4$ |
| | Tension ratio | $X5$ |
| | Number of towers | $X6$ |
| | Tower material consumption | $X7$ |
| | Foundation earthwork consumption | $X8$ |
| | Foundation concrete consumption | $X9$ |
| | Foundation steel consumption | $X10$ |
| Economic indexes | Price of conductor | $X11$ |
| | Price of tower material | $X12$ |
| | Price of foundation steel | $X13$ |

Table 2: Sample of original data of transmission line project cost.

| Sample | $X1$ | $X2$ | $X3$ | $X4$ | $X5$ | $X6$ | $X7$ | $X8$ | $X9$ | $X10$ | $X11$ | $X12$ | $X13$ | Unit length cost (10 kUSD/km) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.30 | 4.00 | 10.00 | 3.85 | 0.65 | 7.16 | 62.82 | 0.13 | 4.06 | 2.86 | 0.62 | 1.29 | 0.59 | 7.58 |
| 2 | 1.00 | 5.00 | 12.99 | 0.41 | 6.46 | 108.77 | 1243.08 | 1.00 | 9.64 | 0.60 | 0.62 | 1.25 | 0.38 | 7.17 |
| 3 | 1.00 | 6.00 | 7.00 | 6.25 | 5.26 | 92.14 | 808.04 | 1.00 | 10.17 | 2.01 | 0.62 | 1.29 | 0.57 | 5.69 |
| 4 | 1.05 | 3.24 | 7.00 | 3.66 | 0.23 | 34.04 | 74.46 | 0.09 | 1.38 | 2.91 | 0.63 | 1.30 | 0.28 | 5.35 |
| 5 | 1.25 | 3.50 | 4.50 | 2.27 | 0.99 | 20.93 | 231.72 | 0.46 | 13.29 | 2.96 | 0.66 | 1.67 | 0.40 | 8.89 |
| 6 | 1.00 | 2.00 | 4.50 | 2.35 | 0.71 | 30.52 | 117.24 | 0.23 | 8.07 | 2.90 | 0.58 | 1.32 | 0.56 | 10.56 |
| 7 | 1.00 | 3.60 | 7.84 | 3.09 | 0.64 | 11.85 | 43.25 | 0.18 | 4.89 | 2.86 | 0.62 | 1.29 | 0.51 | 5.86 |
| 8 | 1.00 | 4.00 | 4.46 | 2.26 | 1.82 | 41.44 | 154.37 | 0.27 | 21.55 | 2.94 | 0.61 | 1.20 | 0.28 | 8.65 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 140 | 1.00 | 2.10 | 3.00 | 2.47 | 2.48 | 34.14 | 149.96 | 0.58 | 15.69 | 2.91 | 0.66 | 1.26 | 0.26 | 7.40 |

Table 3: Total variance explained.

| Component | Initial eigenvalues | | |
|---|---|---|---|
| | Total | % of variance | Cumulative (%) |
| 1 | 3.441 | 26.469 | 26.469 |
| 2 | 2.155 | 16.573 | 43.042 |
| 3 | 1.695 | 13.040 | 56.081 |
| 4 | 1.293 | 9.945 | 66.026 |
| 5 | 1.065 | 8.196 | 74.222 |
| 6 | 0.965 | 7.422 | 81.644 |
| 7 | 0.787 | 6.055 | 87.699 |
| 8 | 0.535 | 4.116 | 91.816 |
| 9 | 0.357 | 2.749 | 94.565 |
| 10 | 0.275 | 2.118 | 96.683 |
| 11 | 0.222 | 1.705 | 98.389 |
| 12 | 0.137 | 1.050 | 99.439 |
| 13 | 0.073 | 0.561 | 100.000 |

corresponding principal components. It can be concluded that 87.699% of the total information can be explained by the first seven principal components, which is over 85%. Therefore, it is considered that extracting the seven principal components can better explain the information contained in the original variables.

In order to establish the expression between the principal component and 13 cost impact indexes, the component score coefficient matrix is obtained by SPSS calculation as shown in Table 4.

According to this matrix, the expression between seven principal components and cost impact indexes can be obtained. Taking the first principal component as an example,

$$\begin{aligned} U1 = {} & -0.0564X1 + 0.0636X2 - 0.0094X3 - 0.0.0263X4 \\ & + 0.2702X5 + 0.2662X6 + 0.2525X7 + 0.2316X8 \\ & + 0.1115X9 - 0.0824X10 + 0.0143X11 - 0.0198X12 \\ & - 0.0112X13. \end{aligned} \tag{18}$$

*4.2. Establishment and Training of LSSVM Point Prediction Model.* The penalty coefficient of the model parameter is set as $\gamma = 50$, and the kernel function parameter is set as $\sigma^2 = 0.4$ to establish the LSSVM cost point prediction model. The input set of the point prediction model is composed of seven principal components which are processed by PCA, and the unit length cost is selected as the output variable. The LSSVM point prediction model is trained with the first 70 samples, and the latter 70 samples are used to verify the point prediction model, and the corresponding prediction values are obtained. The real values are sorted from small to large, and the predicted values and real values, as well as the corresponding relative errors, are obtained as shown in Figure 3.

The point prediction model is the basis of construction cost interval. This paper uses mean absolute percentage error (MAPE), root-mean-square error (RMSE), mean prediction error (MRE), and determination coefficient ($R^2$) to evaluate and analyze the point prediction model. The calculation formulas of each index are as follows:

$$\begin{aligned} \text{MAPE} &= \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \widehat{y}_i}{\widehat{y}_i}\right|, \\ \text{RMSE} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}, \\ \text{MRE} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\left|y_i - \widehat{y}_i\right|}{\widehat{y}_{\max}}, \\ R^2 &= 1 - \frac{(1/n)\sum_{i=1}^{n}\left(y_i - \widehat{y}_i\right)^2}{(1/n)\sum_{i=1}^{n}\left(y_i - \widehat{y}_{\text{mean}}\right)^2}, \end{aligned} \tag{19}$$

where $y_i$ represents the predicted value of project cost, $\widehat{y}_i$ represents the actual value of project cost, $n$ represents the number of samples, and $\widehat{y}_{\max}$ and $\widehat{y}_{\text{mean}}$ represent the maximum and average value of the actual value of project cost. The MAPE value, RMSE value, MRE value, and $R^2$ value of the LSSVM model are 8.59%, 4.99, 4.94%, and 88.67%. It can be concluded that the LSSVM point prediction model has high prediction accuracy and can further establish the cost interval.

*4.3. Establishment of Cost Interval.* On the basis of point prediction, the probability density of prediction error is obtained by nonparametric kernel density estimation according to the relative error of point prediction. Then the probability distribution polarity curve of the prediction error is fitted by cubic spline interpolation, and the $(\alpha/2)$ and $1 - (\alpha/2)$ quantile are found. In this paper, the optimal window width is given automatically by using MATLAB, and the kernel function is the Gauss kernel function:

$$K(t) = \frac{1}{\sqrt{2\pi}}e^{t^2/2}. \tag{20}$$

The estimated error probability curve is shown in Figure 4.

Figure 4 shows that the kernel density curve is in good agreement with the frequency histogram, retaining the internal characteristics, and in the tail, it is in good agreement with the normal distribution, with less interference. Cubic spline interpolation is used to fit the probability distribution of relative error of prediction, and the corresponding quantile $(\alpha/2)$ and $1 - (\alpha/2)$ are found. The confidence intervals of the predicted values are calculated when the confidence level is 95%, 90%, 85%, and 80%, as shown in Table 5 and Figure 5.

This paper uses prediction interval coverage probability (PICP) to evaluate the reliability of the prediction interval. The formula is as follows:

Table 4: Component score coefficient table.

| | $U1$ | $U2$ | $U3$ | $U4$ | $U5$ | $U6$ | $U7$ |
|---|---|---|---|---|---|---|---|
| $X1$ | −0.0564 | 0.3094 | 0.2365 | 0.1536 | −0.2281 | 0.0275 | −0.3417 |
| $X2$ | 0.0636 | 0.3136 | 0.1300 | 0.1818 | −0.0485 | −0.3319 | 0.0203 |
| $X3$ | −0.0094 | 0.1011 | −0.2514 | 0.5814 | 0.2054 | 0.1198 | 0.6599 |
| $X4$ | −0.0263 | −0.1219 | 0.4681 | 0.4893 | −0.0597 | 0.2384 | −0.0754 |
| $X5$ | 0.2702 | 0.0184 | −0.0648 | −0.0640 | 0.0762 | −0.0284 | 0.1344 |
| $X6$ | 0.2662 | −0.0336 | 0.0089 | 0.0300 | −0.0251 | 0.0379 | −0.0868 |
| $X7$ | 0.2525 | 0.0701 | −0.0916 | 0.0931 | 0.0672 | 0.0549 | −0.1686 |
| $X8$ | 0.2316 | −0.0451 | 0.0576 | −0.0521 | −0.0594 | 0.0719 | 0.1753 |
| $X9$ | 0.1115 | −0.2788 | 0.3882 | 0.0888 | −0.0602 | 0.0311 | −0.0484 |
| $X10$ | −0.0824 | −0.3102 | 0.0243 | 0.0310 | 0.2293 | −0.2194 | 0.2107 |
| $X11$ | 0.0143 | 0.1091 | 0.3816 | −0.1316 | 0.4250 | −0.6262 | 0.2349 |
| $X12$ | −0.0198 | 0.1560 | 0.3159 | −0.3781 | −0.0701 | 0.5271 | 0.6500 |
| $X13$ | −0.0112 | 0.0895 | 0.0432 | −0.0299 | 0.7493 | 0.4521 | −0.4179 |

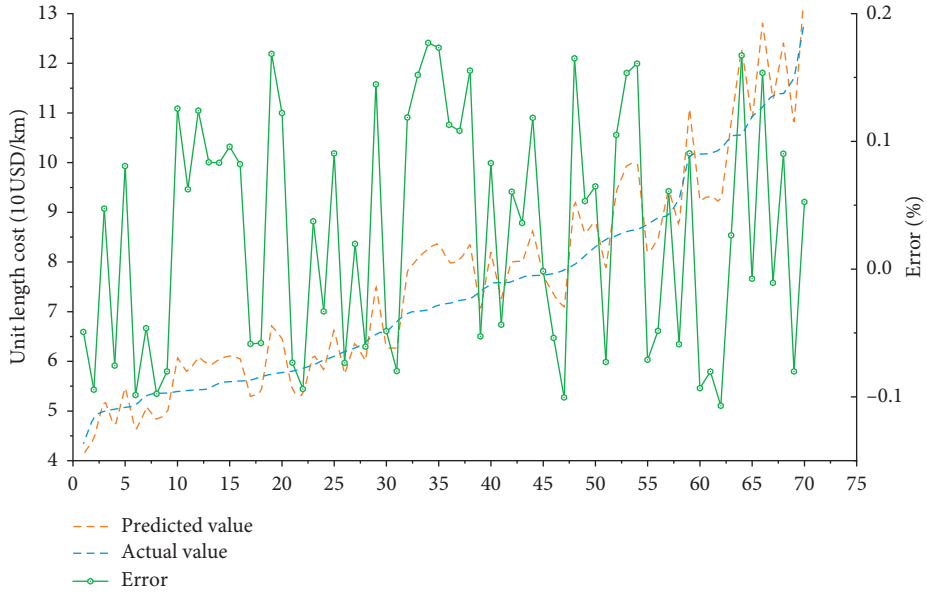


Figure 3: Comparison and verification of results.

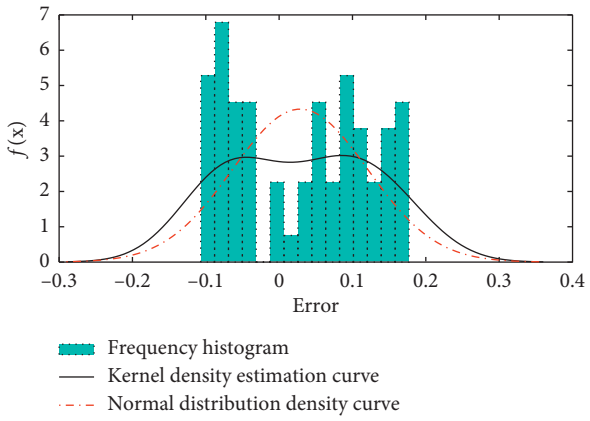


Figure 4: Curve of error distribution for kernel density estimation.

$$\mathrm{PICP} = \frac{1}{N}\sum_{i=1}^{N}\rho_i, \tag{21}$$

where $N$ is the total number of samples and $\rho_i$ is the Boolean quantity. If the actual value falls within the prediction range, $\rho_i = 1$, otherwise $\rho_i = 0$. When PICP = 1, it means that all the actual values fall within the predicted range, and the reliability is the highest. If only for the sake of purely pursuing reliability, the endpoint of the prediction interval can be scaled to the boundary value, but the prediction interval thus obtained loses its practical significance. Therefore, in order to get an effective prediction interval, PICP should be as close as possible to the preset confidence level in the actual interval prediction. If PICP is far less than the confidence level, the predicted interval is invalid and needs to be reconstructed. PICP at each confidence level is shown in Table 6.

Table 5: Cost intervals with different confidence levels.

| Sample | Actual value | Point prediction value | Cost prediction interval | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | 80% confidence level | | 85% confidence level | | 90% confidence level | | 95% confidence level | |
| 1 | 4.34 | 4.13 | 3.60 | 4.53 | 3.56 | 4.58 | 3.50 | 4.71 | 3.41 | 4.87 |
| 2 | 4.90 | 4.44 | 3.87 | 4.87 | 3.83 | 4.93 | 3.76 | 5.06 | 3.67 | 5.23 |
| 3 | 5.01 | 5.24 | 4.57 | 5.76 | 4.52 | 5.82 | 4.44 | 5.98 | 4.34 | 6.18 |
| 4 | 5.04 | 4.65 | 4.06 | 5.11 | 4.01 | 5.17 | 3.94 | 5.31 | 3.85 | 5.49 |
| 5 | 5.08 | 5.49 | 4.78 | 6.02 | 4.73 | 6.09 | 4.65 | 6.25 | 4.54 | 6.47 |
| 6 | 5.10 | 4.59 | 4.01 | 5.04 | 3.96 | 5.10 | 3.89 | 5.24 | 3.80 | 5.42 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 70 | 12.84 | 13.52 | 11.79 | 14.84 | 11.65 | 15.01 | 11.45 | 15.41 | 11.18 | 15.94 |

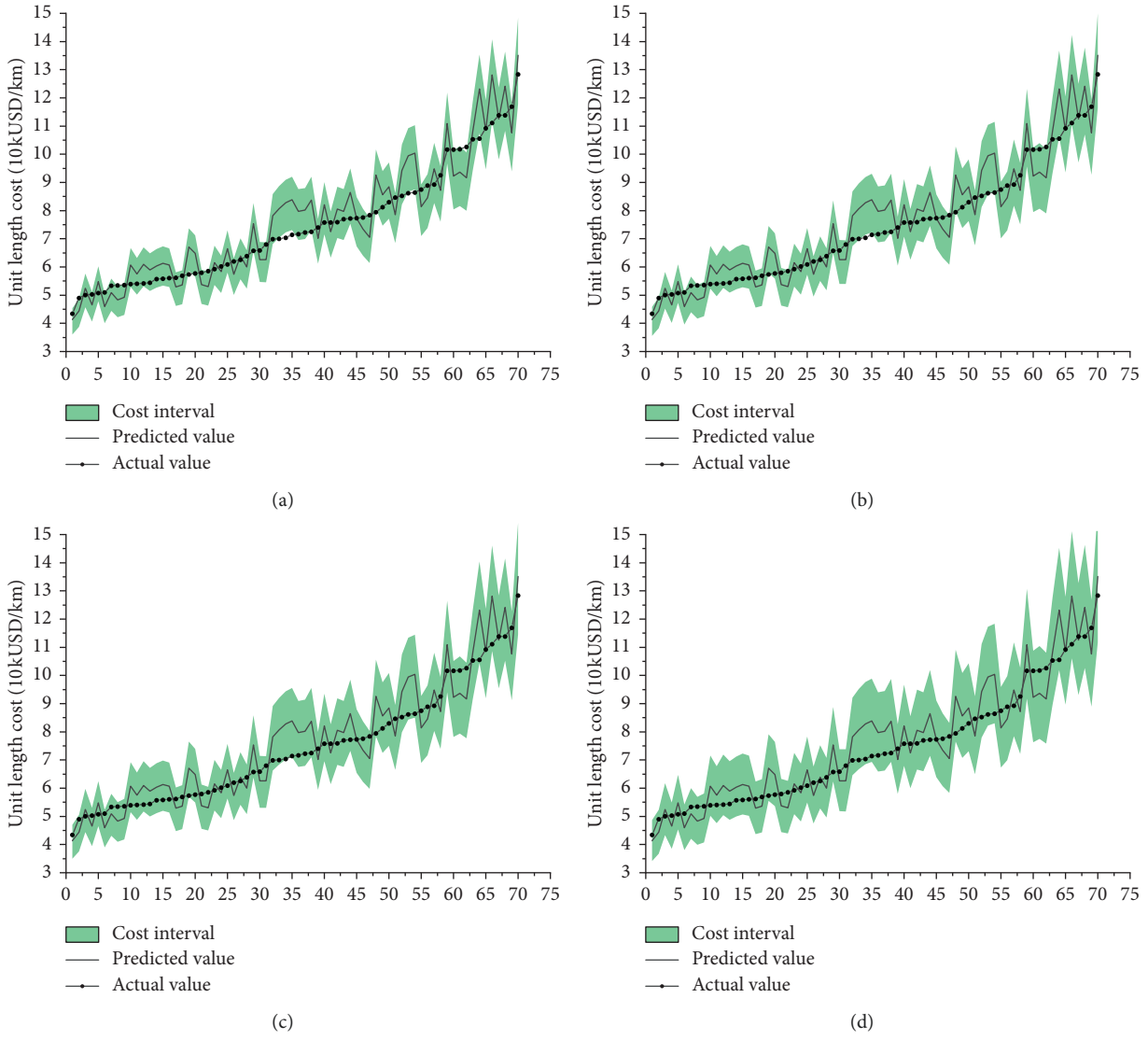


Figure 5: Cost intervals at different confidence levels: (a) 80%, (b) 85%, (c) 90%, and (d) 95%.

Table 6 shows that PICP at 85% confidence level is the closest to the confidence level. It shows that the interval prediction model has high reliability at 85% confidence level, which ensures the accuracy of prediction and does not lose practical significance because of the high value. Therefore, the cost interval under 85% confidence level should be selected.

Table 6: PICP for cost intervals with different confidence levels.

| Confidence level (%) | PICP (%) |
|---|---|
| 80 | 75.71 |
| 85 | 88.57 |
| 90 | 100 |
| 95 | 100 |

## 5. Conclusion

The investment planning of transmission line project is of great significance to improve the investment benefit of the power grid project. Strengthening the evaluation of investment planning is an important means to determine investment rationally. Therefore, it is necessary to change the cost prediction from point prediction to interval prediction in order to improve the compatibility and reliability of investment planning evaluation. This paper proposes a reasonable cost-level prediction model based on PCA, LSSVM, and KDE. The PCA is used to screen and reduce the dimension of transmission line project cost data, and the principal component which can basically describe the factors affecting the cost is obtained as the input set of the prediction model. Using the theoretically mature LSSVM point prediction model, the nonlinear mapping between transmission line engineering characteristics and the cost is determined, and the model is trained and predicted. The error of the point prediction model is analyzed, and the probability density function of error is estimated by the KDE method. The corresponding cost interval is obtained according to different confidence levels. Finally, the accuracy and reliability of the interval prediction model are verified by calculating the PICP of different confidence level cost intervals. The results show that the PICP of the cost interval prediction model based on PCA-LSSVM-KDE is 88.57% at 85% confidence level, which not only guarantees sufficient accuracy but also has strong practical significance.

In summary, the reasonable level of the cost prediction model based on PCA-LSSVM-KDE proposed in this paper has good compatibility and reliability for transmission line project cost prediction. This model can have strong practical significance and good application effect in transmission line project cost investment planning and evaluation.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare that there are no conflicts of interests regarding the publication of this paper.

## Acknowledgments

This study was supported by the National Natural Science Foundation of China (NSFC) (71501071), Beijing Social Science Fund (16YJC064), and Fundamental Research Funds for the Central Universities (2017MS059 and 2018ZD14).

## References

[1] Z. Z. Han, "Study on transmission line cost control line based on Monte Carlo simulation," *China Power Enterprise Management*, vol. 15, pp. 92–96, 2016.

[2] W. Liu, B. He, and Y. P. Wang, "Deepening analysis of universal cost of overhead transmission line project," *China Power Enterprise Management*, vol. 3, pp. 22–24, 2016.

[3] W. Ji and S. M. Abourizk, "Data-driven simulation model for quality-induced rework cost estimation and control using absorbing Markov chains," *Journal of Construction Engineering and Management*, vol. 144, no. 8, article 04018078, 2018.

[4] A. Lesniak and M. Juszczyk, "Prediction of site overhead costs with the use of artificial neural network based model," *Archives of Civil and Mechanical Engineering*, vol. 18, no. 3, pp. 973–982, 2018.

[5] A. Bhargava, S. Labi, S. Chen, T. U. Saeed, and K. C. Sinha, "Predicting cost escalation pathways and deviation severities of infrastructure projects using risk-based econometric models and Monte Carlo simulation," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 8, pp. 620–640, 2017.

[6] A. Lesniak and K. Zima, "Cost calculation of construction projects including sustainability factors using the case based reasoning (CBR) method," *Sustainability*, vol. 10, no. 5, p. 1608, 2018.

[7] M. Juszczyk, A. Lesniak, and K. Zima, "ANN based approach for estimation of construction costs of sports fields," *Complexity*, vol. 2018, Article ID 7952434, 11 pages, 2018.

[8] V. N. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, NY, USA, 1995.

[9] C. Aldrich and L. Auret, "Statistical learning theory and kernel-based methods," in *Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods*, pp. 117–181, Springer, London, UK, 2013.

[10] L. N. Jiang, *Research on the Predict of the Construction Cost Based on Support Vector Machine*, Hebei University of Engineering, Handan, China, 2009.

[11] R. Dong, J. Xu, and B. Lin, "ROI-based study on impact factors of distributed PV projects by LSSVM-PSO," *Energy*, vol. 124, pp. 336–349, 2017.

[12] R. G. Gorjaei, R. Songolzadeh, M. Torkaman, M. Safari, and G. Zargar, "A novel PSO-LSSVM model for predicting liquid rate of two phase flow through wellhead chokes," *Journal of Natural Gas Science and Engineering*, vol. 24, pp. 228–237, 2015.

[13] Y. Liang, D. Niu, M. Ye, W.-C. Hong et al., "Short-term load forecasting based on wavelet transform and least squares support vector machine optimized by improved cuckoo search," *Energies*, vol. 9, no. 10, p. 827, 2016.

[14] A. Q. Kang, Q. Tan, X. Yuan, X. Lei, and Y. Yuan, "Short-term wind speed prediction using EEMD-LSSVM model," *Advances in Meteorology*, vol. 2017, Article ID 6856139, 22 pages, 2017.

[15] J. Kong, X. Y. Cao, and F. Xiao, "Research on cost forecasting model of power transmission and transformation project based on support vector machine," *Modern Electronics Technique*, vol. 41, no. 4, pp. 127–130, 2018.

[16] D. Wang, "Application of Markov chain in comprehensive cost index prediction of power transmission and transformation engineering," *Engineering Cost Management*, vol. 4, pp. 41–44, 2014.

[17] Y. Lu, D. X. Niu, J. P. Qiu, and W. Liu, "Prediction technology of power transmission and transformation project cost based on the decomposition-integration," *Mathematical Problems in Engineering*, vol. 2015, Article ID 651878, 11 pages, 2015.
[18] X. H. Wang, W. N. Wen, Y. C. Lu, and D. Xu, "EEMD-BP based on the cost of power transmission and transformation project uncertainty factor forecast," *China Power Enterprise Management*, vol. 6, pp. 79–84, 2016.
[19] T. Yi, Z. Yan, H. Lv et al., "Unit investment prediction of transmission line projects: a neural network approach," *Journal of the Balkan Tribological Association*, vol. 22, no. 2A, pp. 1659–1668, 2016.
[20] X. Wang, L. An, and Y. Zhang, "Cost prediction of power transmission and transformation project based on the combined variable weight model of REGR-WNN," *China Power Enterprise Management*, vol. 13, pp. 93–96, 2016.
[21] M. A. Grounds, S. Joslyn, and K. Otsuka, "Probabilistic interval forecasts: an individual differences approach to understanding forecast communication," *Advances in Meteorology*, vol. 2017, Article ID 3932565, 18 pages, 2017.
[22] R. K. Samal and M. Tripathy, "Estimating wind speed probability distribution based on measured data at Burla in Odisha, India," *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, vol. 41, no. 8, pp. 918–930, 2019.
[23] F. Fan, K. Bell, and D. Infield, "Transient-state real-time thermal rating forecasting for overhead lines by an enhanced analytical method," *Electric Power Systems Research*, vol. 167, pp. 213–221, 2019.
[24] F. Amara, K. Agbossou, Y. Dubé, S. Kelouwani, A. Cardenas, and J. Bouchard, "Household electricity demand forecasting using adaptive conditional density estimation," *Energy and Buildings*, vol. 156, pp. 271–280, 2017.
[25] J. A. K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least Squares Support Vector Machines*, World Scientific, Singapore, 2002.
[26] Z.-W. Li and P. He, "Data-based optimal bandwidth for kernel density estimation of statistical samples," *Communications in Theoretical Physics*, vol. 70, no. 6, pp. 728–734, 2018.
[27] T. Yi, H. Zheng, Y. Tian, and J.-P. Liu, "Intelligent prediction of transmission line project cost based on least squares support vector machine optimized by particle swarm optimization," *Mathematical Problems in Engineering*, vol. 2018, Article ID 5458696, 11 pages, 2018.