Mathematical Biosciences and Engineering

MBE, 20(8): 15044–15074.
DOI: 10.3934/mbe.2023674
Received: 25 March 2023
Revised: 04 July 2023
Accepted: 10 July 2023
Published: 14 July 2023

http://www.aimspress.com/journal/MBE

Research article

Construction cost prediction system based on Random Forest optimized by the Bird Swarm Algorithm

Zhishan Zheng[1], Lin Zhou[2], Han Wu[3] and Lihong Zhou[4,*]

[1] School of Big data and Computer, Jiangxi University of Engineering, Xinyu 338000, China
[2] School of Civil Engineering, Jiangxi University of Engineering, Xinyu 338000, China
[3] School of Infrastructure Engineering, Nanchang University, Nanchang 330047, China
[4] School of Architecture and Environmental Engineering, Nanchang Institute of Science and Technology, Nanchang 330047, China

* **Correspondence:** Email: lihongzhou2021@163.com; Tel: +8615179130995.

Abstract: Predicting construction costs often involves disadvantages, such as low prediction accuracy, poor promotion value and unfavorable efficiency, owing to the complex composition of construction projects, a large number of personnel, long working periods and high levels of uncertainty. To address these concerns, a prediction index system and a prediction model were developed. First, the factors influencing construction cost were first identified, a prediction index system including 14 secondary indexes was constructed and the methods of obtaining data were presented elaborately. A prediction model based on the Random Forest (RF) algorithm was then constructed. Bird Swarm Algorithm (BSA) was used to optimize RF parameters and thereby avoid the effect of the random selection of RF parameters on prediction accuracy. Finally, the engineering data of a construction company in Xinyu, China were selected as a case study. The case study showed that the maximum relative error of the proposed model was only 1.24%, which met the requirements of engineering practice. For the selected cases, the minimum prediction index system that met the requirement of prediction accuracy included 11 secondary indexes. Compared with classical metaheuristic optimization algorithms (Particle Swarm Optimization, Genetic Algorithms, Tabu Search, Simulated Annealing, Ant Colony Optimization, Differential Evolution and Artificial Fish School), BSA could more quickly determine the optimal combination of calculation parameters, on average. Compared with the classical and latest forecasting methods (Back Propagation Neural Network, Support Vector Machines, Stacked Auto-Encoders and Extreme Learning Machine), the proposed model exhibited higher forecasting accuracy and efficiency. The prediction model proposed in this study could better support the prediction of construction cost, and the prediction results provided a basis for optimizing the cost management of construction projects.

Keywords: building engineering; construction cost prediction; Random Forest; Bird Swarm Algorithm

1. Introduction

As a pillar industry of the country, particularly in developing countries, the construction industry plays a major role in domestic economic growth. Moreover, competition in the construction market has intensified. For construction enterprises, the cost management of construction projects is the center of all business and management activities, and forecasting construction cost is an important part of the entire cost management system [1]. Therefore, the accurate prediction of the cost of construction projects bears profound significance in reducing the construction cost and improving the efficiency of enterprises.

Currently, research results related to cost forecasting have been widely available, but two prominent problems remain. 1) Most research results are tantamount to developing the construction cost prediction index system for construction projects at the project level. This type of index system is highly detailed, and its scope of application is narrow, failing to provide reference and insight into the cost management of construction enterprises. 2) Current research methods are mainly static qualitative predictions, such as expert meetings, procedural investigation and subjective probability methods. These approaches are largely influenced by subjective factors, and the effective handling of the nonlinear characteristics of small samples in cost forecasting presents a challenge [2].

With the development of artificial intelligence, the application of machine learning algorithms in the study of cost prediction has become a trend [3]; however, inadequacies in related research results still arise. A decision tree (DT) requires considerable data preprocessing, and the prediction results easily fall into a local optimum [4]. A Support Vector Machine (SVM), a typical nonlinear modeling tool, has a highly complex computational function and its computational performance in solving multi-classification problems is minimal [5]. The Artificial Neural Network (ANN), the most commonly used nonlinear modeling tool, possesses excellent nonlinear modeling ability but has several disadvantages related to learning, local minimum and slow convergence speed [6].

Proposed by Breiman in 2001, random forest (RF) [7] is a combined classification intelligent algorithm based on statistical learning theory. The basic ideas for the application of RF in nonlinear modeling are as follows. 1) A plurality of weak classifiers is combined to form a strong classifier. 2) These weak classifiers play a complementary role. 3) By reducing the influence of a single classifier error, the classification accuracy and stability are improved. Compared with other classical nonlinear modeling techniques, RF exhibits robust data mining ability, high prediction accuracy and good tolerance for outliers and noise. In addition, RF is not easy to appear in the fitting scene. At present, RF has been successfully applied in forecasting precious metal prices [8], the gross domestic product (GDP) [9] and power load [10].

In the RF model, the appropriate selection of the number of DTs and split features can efficiently reduce the complexity of the RF and improve its computational performance to form an enhanced integrated classifier. Classical metaheuristic optimization algorithms, such as the Particle Swarm Optimization (PSO) [11] and Genetic Algorithms (GA) [12], are broadly used to solve the optimal calculation parameters of the RF. However, the capabilities of these two algorithms for global retrieval are not satisfactory.

Bird Swarm Algorithm (BSA) is a novel biological heuristic algorithm proposed by Meng et al. [13]

in 2015. BSA simulates the biological behaviors of birds in nature, such as foraging, vigilance and migration. Reporting to the classical metaheuristic optimization algorithm, this algorithm has the characteristics of decentralized search, maintaining population diversity and avoiding falling into a local optimum. BSA has been successfully applied in the optimal parameter calculation of the Back Propagation Neural Network (BPNN) [14], torsional capacity evaluation of RC beams [15] and the optimization of the vehicle powertrain [16]. To the best of our knowledge, there is no research result on the optimization of RF calculation parameters based on BSA.

Therefore, to address the inadequacies, such as low prediction accuracy and efficiency, we developed a novel model for predicting the construction cost of a building. This study presents the following major contributions. 1) Most index systems of construction cost prediction in related studies were built at the project level, which had some disadvantages, such as a narrow scope of application. From the perspective of construction enterprises, the index system of construction cost prediction was constructed for the first time. The index system exhibited good adaptability and could be widely applied in construction cost management for construction enterprises. 2) A construction cost prediction model based on RF optimization by BSA was proposed. This model showed strong data mining ability and high prediction accuracy, effectively enhancing the prediction accuracy and efficiency of the construction cost.

The remainder of this manuscript is arranged as follows. Section 2 summarizes the research results associated with this study; Section 3 analyzes the factors influencing architectural project construction cost, and constructs the related prediction index system; Section 4 introduces the construction cost prediction method based on the RF optimized by BSA; and Section 5 presents a case study to verify the science and effectiveness of the proposed model. Section 6 compares the computational performance of different algorithms to emphasize the advantages of the model proposed in this study. Section 7 summarizes the research results and limitations of this study.

2. Related work

Kim et al. [6] proposed a prediction model of construction cost, based on the Regression Comprehensive Moving Average Model (RCMAM) and ANN. The complexity and prediction workload of this hybrid model were larger than those of RCMAM or ANN. Whether the prediction accuracy was improved due to the increase in model complexity or the advanced nature of the hybrid model itself was difficult to assess. Thus, the evaluation that this hybrid model proposed by Kim exhibited enhanced calculation accuracy might be biased. With research on the concrete engineering cost in Egypt, Elfaham [17] emphasized the influence of inflation on the prediction results. The reasonable method of eliminating this effect was to translate the costs generated at different time points into the same time position. In the study by Elfaham, ANN was designed to build the prediction model but was not compared with other nonlinear modeling methods. Cao and Ashuri [18] constructed a highway construction cost prediction model based on Long Short-Term Memory. However, their study only considered the quantitative factors, such as the price of building materials and the wages of construction workers; it lacked research on the qualitative factors, such as the level of construction management and the ecological environment of construction. In accordance with the project management practice in construction engineering, these qualitative factors also exerted a significant effect on the construction cost. Using Complex Network Analysis, Mao and Xiao [19] developed a novel construction cost prediction model. The influencing factors of the construction cost were

regarded as network nodes, and the cost was expected by analyzing the relationship of each network node. However, CNA is a typical sociological research technique, which is easily influenced by the subjective judgment of managers.

Pierioch and Risse [8] used RF to construct the price prediction model of precious metals. Reports on classical methods, such as Multiple Linear Regression (MLR), indicate that the results based on RF showed higher prediction accuracy. In addition, multivariate and unifoliate prediction results were compared, and the multivariate prediction was found to be more accurate than univalent prediction. Using the economic data of Japan from 2001 to 2018, Yoon [9] constructed an RF-based forecasting method for Japan's GDP. However, the effect of RF initial calculation parameters on the prediction results was not analyzed. Fast and accurate forecasting of short-term power load has consistently been a difficult area in power management research. Accordingly, Dang et al. [10] developed a stochastic RF model to effectively quantify the uncertainty of power load forecasting. Compared with three other classical power load forecasting techniques, this model, based on stochastic RF, was proved to exhibit higher forecasting speed and forecasting accuracy.

Imitating the foraging, alert and flight behaviors of birds, Meng et al. [13] proposed BSA, which more effectively avoided falling into the local optimal solution by solving 18 classic test problems. Zhang et al. [14] successfully found BPNN for Quadrature Amplitude Modulation Signal Recognition in 5G communication systems by BSA. However, their study failed to compare the computational performance of BSA and the classical optimization algorithm. Using ANN as an example, Kaya [20] analyzed the optimization performance of 16 metaheuristic algorithms. Numerous tests showed that calculation by BSA showed the highest speed and stability. Wu et al. [16] solved the typical multi-objective optimization problem with a vehicle powertrain by using BSA. Compared with other optimization algorithms, BSA was more suitable for multi-objective optimization.

3. Prediction index system for building project construction cost

3.1. Analysis of factors influencing the building project construction cost

The project construction cost is the sum of all costs incurred by construction enterprises to successfully complete construction projects. Construction projects are one-off and distinct, rendering the construction costs of different projects significantly different. The building project construction cost generally includes the material, labor, machinery, financial and management costs.

Regulations on the Construction Cost Management of Capital Construction Projects in China, together with previous research results [4,17–19], indicate that the influencing factors of project construction cost are determined from the following aspects.

1) Building scale (X_1)

Building scale refers to the size of the volume, pattern, or scope of building projects. The scale of the building is directly proportional to the materials, machinery and personnel needed for the project; thus, it is directly proportional to the construction cost. The scale of the building generally includes influencing factors, such as the type of structure, total height, area of the standard floor and basement area.

2) Project Management (X_2)

Project management is an important factor considered to determine whether the project cost control is economical. Contrary to other research results [4], project management is separately listed to more clearly analyze its main effect on the construction cost. Project management consists of four

types: type of contract, the difficulty of resource scheduling, the proportion of managers and the difficulty of quality and safety management.

3) Site conditions (X_3)

On-site conditions determine the difficulty of the construction and thus also markedly influence the consumption of construction costs. This primary index includes the quality of life of employees and the quality of the construction environment, all of which influence the smooth progress of a project. In this study, site conditions were mainly divided into the distance from the location of the material supply, frequency of disasters and rationality of site layout.

4) Fluctuation of price (X_4)

The cost of project construction could be simply regarded as quantity multiplied by the unit price. The first three indicators mostly affect the project construction cost of buildings by influencing the quantity of work. Fluctuation in price was selected as an indicator representing the influence of price change on the project construction cost. Fluctuation in price mostly included fluctuations in material, labor and machinery costs.

3.2. Proposed prediction index system

On the basis of the principles of comprehensiveness, science, timeliness, applicability and comparability, the prediction index system was constructed. Details are listed in Table 1.

Table 1. Prediction index system of the building project construction cost.

Primary index	Secondary index	Unit	References
Building scale: X_1	Type of structure: X_{11}	-	[4,21]
	Total height: X_{12}	m	[17,18]
	Area of standard floor: X_{13}	m^2	[6,18,22]
	Area of basement: X_{14}	m^2	[2,23]
Project Management: X_2	Type of contract: X_{21}	-	[3,24]
	Difficulty of resource scheduling: X_{22}	-	[6,17]
	Proportion of managers: X_{23}	%	[2,18]
	Difficulty of quality and safety management: X_{24}	-	[4,19]
Site conditions: X_3	Distance from the material supply place: X_{31}	Km	[21,23]
	Frequency of disasters: X_{32}	Times/Year	[18,25]
	Rationality of site layout: X_{33}	-	[2,3,5]
Fluctuation of price: X_4	Material price index: X_{41}	-	[18,26]
	Rationality of site layout: X_{42}	-	[26,27]
	Machinery price index: X_{43}	-	[19,26]

In order to facilitate readers, this paper briefly summarized the types and data characteristics of indicators in Table 1. There were two kinds of indicators in Table 1, qualitative indicators and quantitative indicators. Qualitative indicators (X_{22}, X_{24} and X_{42}) were obtained by questionnaires or expert interviews, while quantitative indicators (all indicators excepted X_{22}, X_{24} and X_{42}) were obtained by quantitative methods based on statistics. For the data of X_{11} and X_{21}, preset and

corresponding integers were taken according to the types of projects. The data sources of X_{22}, X_{24}, X_{33} were questionnaire survey results. See the following Section 3.3 for the qualitative explanation of different data results. X_{32} was the average number of natural disasters in recent years. X_{41}, X_{42} and X_{43} could be calculated statistically, or referred to the cost information issued by the local government.

It should be emphasized that the rationality of the prediction index system of the construction cost has a significant effect on the subsequent prediction accuracy. Section 6.1 elaborates on the importance of each prediction index to reveal the regular pattern of the effect of the prediction index system on the accuracy of prediction.

3.3. Data processing methods

a) Type of structure (X_{11})

The type of structure determined the choice of construction technology and building materials, affecting the direct construction cost. The most common types were the frame structure 1), steel structure 2), shear wall structure 3), the brick–concrete structure 4) and tube structure 5). The values in brackets represented the index scores corresponding to the different types. The X_{11} data point was obtained by consulting the project management information and was an index without the measurement unit.

b) Total height (X_{12})

The total height of the buildings was calculated as the height of the highest point to the outdoor ground as the reference. The higher the building, the more rigorous the performance requirements set on the materials and the higher the cost of construction management. The X_{12} data point, the area of the standard floor (X_{13}) and the area of the basement (X_{14}) were determined by consulting design drawings. The measurement unit of X_{13} was m, whereas that of X_{13} and X_{14} was m^2.

c) Area of the standard floor (X_{13})

The area of the standard floor was a critical factor in the construction cost. The larger the area of the standard floor, the higher the labor, material and machinery costs needed. Notably, the cost estimation method commonly used in engineering practice is to approximate the construction cost based on the standard floor building area.

d) Area of the basement (X_{14})

The construction of the foundation and the basement entailed difficulty, comprising about 15–30% of the total construction cost. The construction cost of the foundation and the basement was approximately considered proportional to the basement area.

e) Type of contract (X_{21})

The type of contract determined the organization and management system of the project and the contracting method, which was crucial for the implementation of the project. If the same construction project adopted different contract types, the construction costs would vary. Common contracts included fixed lump-sum contract 1), fixed unit-price contract 2), variable lump-sum contract 3) and variable unit-price contract 4). The X_{21} data point was obtained by consulting the project management information and was an index without a measurement unit. The values in brackets represented the index scores corresponding to the different types.

f) Difficulty in resource scheduling (X_{22})

Resource scheduling is the rational allocation and mobilization of construction machinery and materials. Reasonable resource scheduling enables site coordination and reduces the construction cost.

In this study, the difficulty of resource scheduling was selected as a comprehensive qualitative index because of the complexity of resource scheduling at the construction site. The data of X_{22}, the difficulty of quality and safety management (X_{24}) and the rationality of the site layout (X_{33}) were obtained using the questionnaire survey or the expert interview. They were indexes without the measurement unit. In this study, it is very easy, easy, general, difficult and very difficult to divide the evaluation results of X_{22} into five grades. The quantitative evaluation interval of very easy was $(80, 100]$, and its qualitative language description was that the coverage of resources was small and the quantity was small, which would not cause trouble to the project cost management. The quantitative evaluation interval of easy was $(60, 80]$, and its qualitative language description was that the resource coverage was moderate and the quantity was appropriate, and the demand could be met through simple scheduling. The quantitative evaluation interval of general was $(40, 60]$, and its qualitative language description was that resources covered a wide range and had a large number, which required more detailed and comprehensive scheduling. The quantitative evaluation interval of difficult was $(20, 40]$, and its qualitative language description was that resources covered a wide range and a large number, and scheduling required a high level of technology and management, which was easy to encounter bottlenecks. The quantitative evaluation interval of very difficult was $[0,20]$, its qualitative language description was that resources covered a very wide range and had a large number, scheduling needed superb technology and management experience, and needed to overcome various complicated problems and difficulties. Experts could quantify the index according to the actual situation of the project and the qualitative language description of the index.

g) Proportion of managers (X_{23})

Organization and coordination refer to the organization and distribution of personnel inside and outside the project. A reasonable organizing staff can effectively avoid an increase in management costs caused by oversaturation. In this study, the proportion of managers was selected as an index describing the influence of personnel organization and coordination on the construction cost. X_{23} is calculated as follows:

$$X_{23} = \frac{X_{23}^1}{X_{23}^2} \times 100\%, \tag{1}$$

where X_{23}^1 is the number of managers, and X_{23}^2 is the number of construction workers. During the construction stage, the staff mobility of the project management team and the construction operation team is considerably strong. Thus, the data points X_{23}^1 and X_{23}^2 should be observed and averaged several times.

h) Difficulty of quality and safety management (X_{24})

Quality and safety management are important contents of project management. If quality management is improper, the completed construction content can be easily reworked, and the construction cost could increase. The occurrence of a construction safety accident can result in casualties and property losses. In this study, the difficulty of quality and safety management was selected as a comprehensive qualitative index because of the complexity of quality and safety management at the construction site. The evaluation grade of X_{24} was divided into five grades, which are very easy $(80, 100]$, easy $(60, 80]$, average $(40, 60]$, difficult $(20, 40]$ and very difficult $[0,20]$. The qualitative language description of very easy was that the quality and safety management in project implementation was very comfortable, easy to control and extremely low in risk. Qualitative language description of easy meant that the quality and safety management in project implementation was not

difficult, could be well controlled and has low risk. The qualitative language of general described that the quality and safety management in project implementation was not difficult and needed to be controlled through reasonable management measures, and there were certain risks. The qualitative language description of difficult was that the quality and safety management in project implementation is difficult, which required more powerful management measures to be controlled and might face certain risks. The qualitative language description of very difficult was that the quality and safety management in project implementation was extremely difficult, requiring extremely powerful management measures to be controlled, and might face high risks. Experts could quantify the index according to the actual situation of the project and the qualitative language description of the index.

i) Distance from the material supply place (X_{31})

The transportation cost of materials was an important component of the construction cost. The farther the distance between the construction site and the material supply location, the greater the transportation cost. The material loss in transit and the cost of material preservation measures were also proportional to the distance. X_{31} is calculated as follows:

$$X_{31} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} X_{31}^{ij}}{\sum_{i=1}^{n} z_i}, \tag{2}$$

where X_{31}^{ij} is the distance from the j-th supplier of the i-th material to the construction site; n is the number of materials; and z_i is the number of suppliers of the i-th material.

j) Frequency of disasters (X_{32})

Natural disasters interrupted the construction process and caused casualties and property losses within the construction scope. In addition, the higher the frequency of natural disasters, the more disaster prevention and mitigation measures were needed, which increased the construction cost. The data point of X_{32} was obtained by consulting local meteorological data and geological survey reports. This paper suggested that the annual frequency of natural disasters in recent 10 years should be the score of this index.

k) Rationality of the site layout (X_{33})

The layout of the construction site significantly affected the construction efficiency. The more reasonable the layout, the higher the construction efficiency, and the lower the amount of labor and machinery used. In this study, the layout of the construction site included too much content; thus, the rationality of the site layout was selected as a comprehensive qualitative index. The evaluation grade of X_{33} was divided into five grades, very reasonable $(80, 100]$, reasonable $(60, 80]$, half $(40, 60]$, unreasonable $(20, 40]$ and very unreasonable $[0,20]$. The qualitative language description of very reasonable was that the layout completely conformed to the regulations and standards, achieved the best design effect, maximized the space utilization and also considered the user's use needs and comfort. The qualitative language description of reasonable was that the layout basically conformed to the regulations and standards, the space utilization rate was high and the user's use needs were basically met. The qualitative language description of half was that the layout basically conformed to the regulations and standards, but the space utilization rate needed to be improved, and the user's use needs had been initially met. The qualitative language description of the unreasonable was that there were obvious violations or irrationalities in the layout, the space utilization rate was low and the user's use needs were not fully met. The qualitative language description of very unreasonable was that the layout was very unreasonable and there were great security risks, which could not meet the needs of

the users. Experts could quantify the index according to the actual situation of the project and the qualitative language description of the index.

l) Material price index (X_{41})

The cost of materials comprised about 30–50% of the total construction cost; thus, the price fluctuation of materials significantly influenced the construction cost. The material price index was selected to reflect the change in material cost during construction. The data calculation method for X_{41} is as follows:

$$X_{41} = \frac{\sum_{i=1}^{n} A_i * B_i}{\sum_{i=1}^{n} A_i * B_i^*}, \tag{3}$$

where n is the number of materials, A_i is the material planned consumption of the i-th material during reporting; B_i is the price of the i-th material during the reporting period, and B_i^* is the price of the i-th material during the reference period.

Notably, the data X_{41}, X_{42} and X_{43} could also refer to the construction price index information published by the local cost management department.

m) Labour price index (X_{42})

The labour price index was selected to reflect the changes in the labour price during the construction process. The data calculation method for X_{42} is given below:

$$X_{42} = \frac{\sum_{i=1}^{n} C_i * D_i}{\sum_{i=1}^{n} C_i * D_i^*}, \tag{4}$$

where n is the number of the main types of work; C_i is the labor planned consumption of the i-th work during the reporting period; D_i is the price of the i-th labour during the reporting period; and D_i^* is the price of the i-th labor during the reference period.

n) Machinery price index (X_{43})

The mechanical price index was selected to reflect the change in the mechanical price during the construction process. The data calculation method for X_{43} is given below:

$$X_{43} = \frac{\sum_{i=1}^{n} E_i * F_i}{\sum_{i=1}^{n} E_i * F_i^*}, \tag{5}$$

where n is the number of major construction machines; E_i is the budgeted consumption of the i-th machine during the reporting period; F_i is the price of the i-th machine usage fee during the reporting period; and F_i^* is the price of the i-th machine usage fee during the reference period.

The construction cost of a building has apparent time benefits. Thus, the construction cost of different time points should be converted to the same time point:

$$Y^* = \frac{Y}{(1+i)^n} \tag{6}$$

where n is the duration; i is the benchmark interest rate; Y^* is the present value of the construction cost; and Y is the present value of the construction cost.

4. Construction cost prediction model for building engineering

4.1. Introduction to RF

RF is a typical machine learning method [7]. It mainly uses the Bootstrap resampling method to extract multiple samples from the original data. RF builds the classification and regression trees (CARTs) for each Bootstrap sample. The predictions of all classification trees are combined, and the final result is derived by voting. With two classifications as an example, the calculation principle for RF is presented in this section.

Supposing two classes, c_1 and c_2, exist and S is the data set at the current tree node, then S is divided into S_1 and S_2 such that the condition $S = S_1 \cup S_2$ is satisfied. S_1 and S_2 represent the sample data assigned to c_1 and c_2, respectively; $\hat{P}(S_1) = \frac{|S_1|}{|S|}$ is the proportion of S_1 in S; and $P(c_i|S_j) = \frac{|S_j \cap c_i|}{|S_j|}$ is the proportion of c_i in S_j.

The variogram $g(S_j)$ in S_j is as follows [7]:

$$g(S_j) = \sum_{i=1}^{2} \hat{P}\left(c_i|S_j\right)\left(1 - \hat{P}\left(c_i|S_j\right)\right). \tag{7}$$

The Gini index G is the weighted sum of the variograms $g(S_1)$ and $g(S_2)$. The calculation method for G is as follows [7,10]:

$$G = \hat{P}(S_1)g(S_1) + \hat{P}(S_2)g(S_2) \tag{8}$$

After constructing a complete forest with k classification trees, it was used to verify or predict new known or unknown data. This forest synthesized the prediction results for k trees and determined the data category by voting. The mathematical expression for C_x is given below [7]:

$$C_x = argmax_c\left(\frac{1}{k}\sum_{i=1}^{k}\mu\left(\frac{C_{h_i,c}}{n_{h_i}}\right)\right), \tag{9}$$

where C_x is the classification result corresponding to the feature set x; k is the number of classification trees; μ is the indicator function; $C_{h_i,c}$ is the classification result for class C by the tree h_i; and n_{h_i} is the number of leaf nodes of h_i.

The forest classification model $H(x)$ is expressed as follows [7]:

$$H(x) = argmax_y \sum_{i=1}^{k} \mu(h_i(x) = y), \tag{10}$$

where h_i is the i-th taxonomic tree.

In addition, some data were not selected in each sampling, and the remaining out-of-bag data (OOB) were used for internal error estimation. Each classification tree had an OOB error estimation, and the average value was used as the generalization error of the model.

When a certain number of trees was reached, the OOB error of the model was considered similar

to that of the optimized model. The major mathematical processes included the following:

The empirical margin function of the sample data set (x, y) is defined as [7,28],

$$\hat{m}(x, y) = \hat{p}_k(h_k(x) = y) - \max_{j \neq y}(h_k(x) = y). \tag{11}$$

The generalization error e of the classifier set h is expressed as [7]:

$$e = p_{x,y}(\hat{m}(x, y) < 0), \tag{12}$$

where (x, y) is the feature-response space composed of x and y.

With an increase in k, the number of classification and regression trees increases [7,11]:

$$e_{k \to \infty} \to p_{x,y}\left[p_\Theta(h(x, \Theta) = y) - \max_{j \neq y}(h(x, \Theta) = j) < 0\right], \tag{13}$$

where p_Θ is the generalization error of the classification tree corresponding to the parameter Θ. In accordance with Eq (13), e should approach a finite upper boundary infinitely with an increase in DT to prevent overfitting.

According to the aforementioned analysis, the number of CARTs and the number of features in the RF method largely influenced the prediction accuracy. The number of CARTs is used to load training samples and their feature factors. The number of features is the number of randomly selected features during each node splitting operation. The flow chart of RF for data prediction was shown in Figure 1.

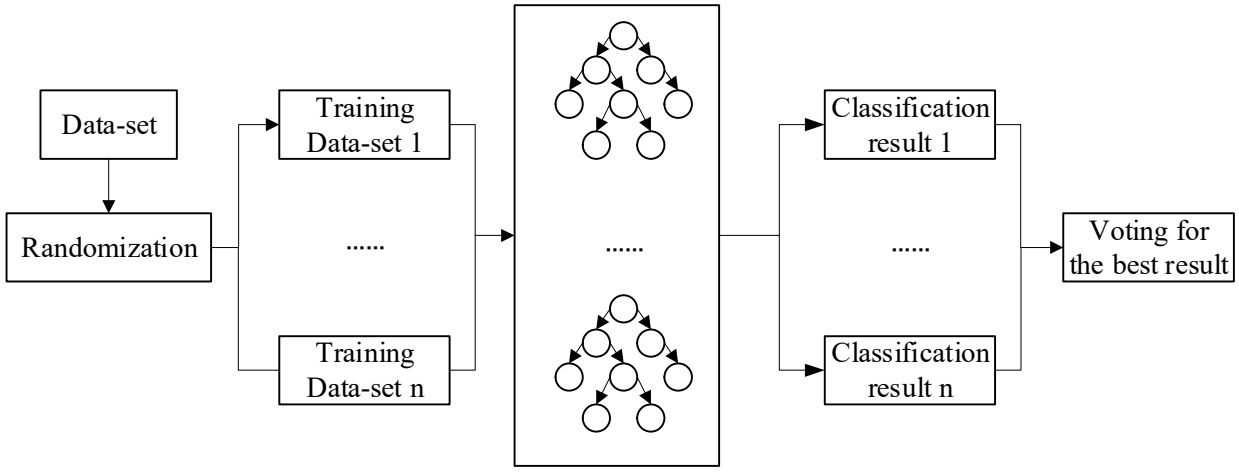

Figure 1. The flow chart of RF for data prediction.

4.2. Introduction to BSA

For a greater survival advantage, the social behavior of birds includes not only foraging behavior but alert behavior as well. BSA is an optimization algorithm derived from the social behavior of birds. The algorithm mostly simulates the division of labor and social interaction between different individuals in the sparrow population when it is looking for food (target). In BSA, the flight position of each bird represents the potential solution. A scattered flight search of birds cannot only maintain

the diversity of the population but also effectively avoids a locally optimal solution. The flow of BSA is presented in Figure 2 [13].

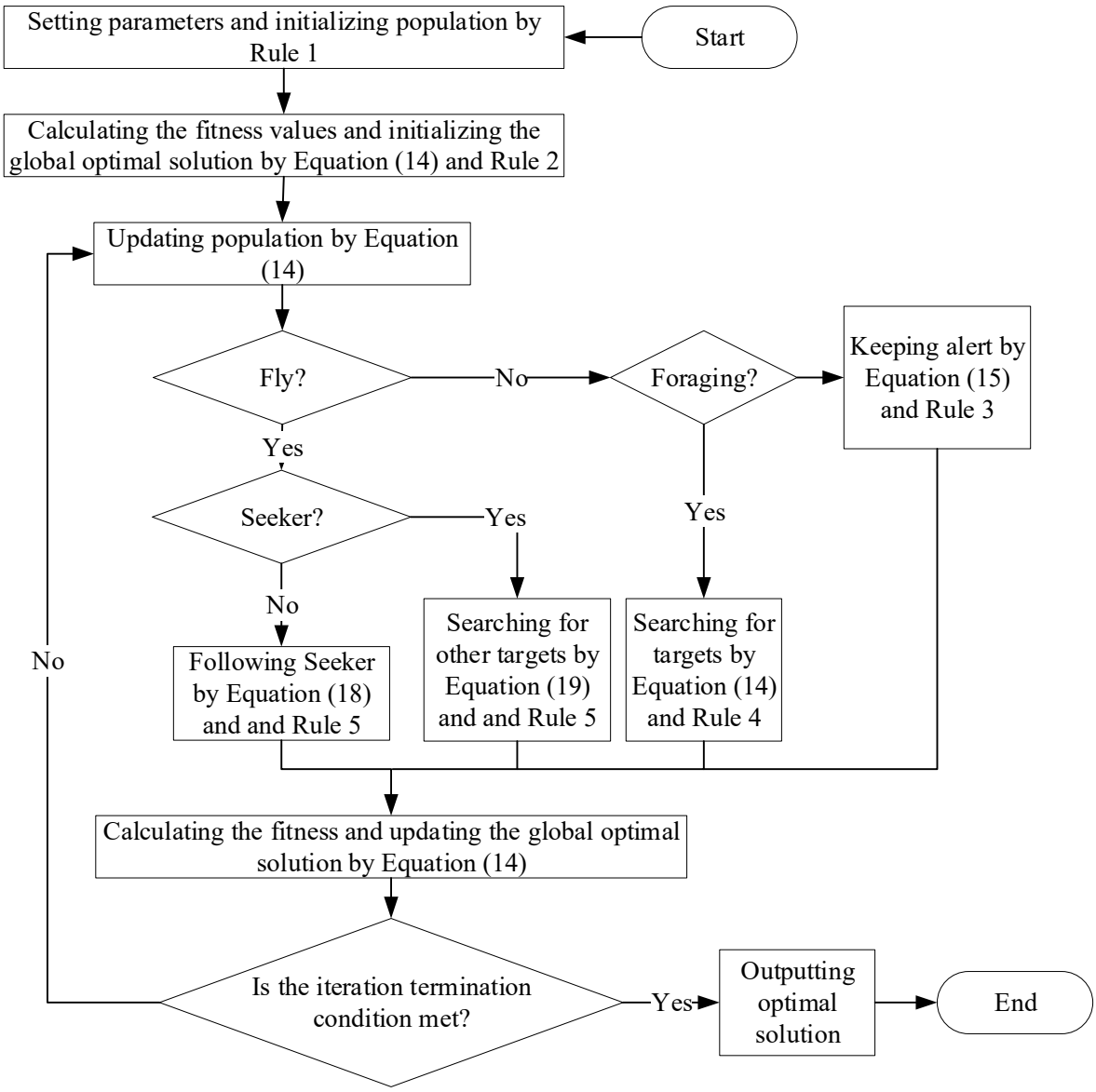

Figure 2. Flow chart of BSA.

The searching behavior of birds in nature, as determined from their social behavior, is transformed into the following five rules [13]:

Rule 1. Individual foraging behavior and alert behavior are random.

Rule 2. Foraging behavior. When a bird is foraging, it can quickly record the best position of the target and update the best position of the target. This information is dynamically shared within the whole flock of birds.

Rule 3. Alert behavior. The alert behavior of birds is the tendency to move to the center of the flock. The more vigilant the bird, the greater the tendency to move.

Rule 4. Flight behavior. Birds periodically fly to another location. When birds fly to a new place,

individuals in the flock switch identities between Seekers and Followers. The birds with high vigilance become Seekers, whereas those with low vigilance become Followers. The rest randomly choose between Seekers and Followers.

Rule 5. Seekers continue to forage, whereas Followers randomly follow a bird that becomes the individual search target of the Seeker.

Suppose there are N birds in the x_i^t $(i \in [1, \cdots, N])$ position at time t, foraging and warning in a D-dimensional space.

1) Foraging behavior

Rule 1 provides a basis for the random decision-making process of birds. When the random number between $(0,1)$ generated by equal probability is greater than the constant P, birds feed, otherwise they remain alert.

In accordance with Rule 2, every bird looks for its target by its own flight experience and group experience. Thus, the position $x_{i,j}^{t+1}$ at time $t+1$ is given below [16]:

$$x_{i,j}^{t+1} = x_{i,j}^t + C * rand(0,1) * \left(p_{ij} - x_{i,j}^t\right) + S * rand(0,1) * \left(g_j - x_{i,j}^t\right), \tag{14}$$

where $j \in [1, \cdots, D]$, $rand(0,1)$ represents the random number between 0 and 1; C is the cognitive coefficient of the individuals; S is the cognitive coefficient of the groups; p_{ij} is the best position of the *i*-th bird before its renewal; and g_j is the best position of the groups.

If the random number $rand(0,1)$ is less than a constant $p\left(p \in (0,1)\right)$, then birds start foraging; otherwise, they remain alert.

2) Alert behavior

In accordance with Rule 3, the alert behavior is described below [13]:

$$x_{i,j}^{t+1} = x_{i,j}^t + A_1 * rand(0,1) * \left(mean_j - x_{i,j}^t\right) + A_2 * rand(-1,1) * \left(p_{k,j} - x_{i,j}^t\right), \tag{15}$$

$$A_1 = a_1 * \exp\left(-\frac{pFit_i}{sumFit_i + \varepsilon} N\right), \tag{16}$$

$$A_2 = a_2 * \exp\left(\frac{pFit_i - pFit_k}{|pFit_i - pFit_k| + \varepsilon} * \frac{N * pFit_k}{sumFit_i + \varepsilon}\right), \tag{17}$$

where k is a positive integer from 1 to N, satisfying the condition $k \neq i$; a_1 and a_2 belong to $[0,2]$; $pFit_i$ is the optimal fitness value of the i-th bird; $sumFit_i$ is the sum of the optimal fitness values of the colony; and ε is a considerably small constant to avoid the situation in which $sumFit_i$ is 0; and $mean_j$ is the average fitness of the j-th birds.

In Figure 2, Keeping alert means that the bird performs alert behavior.

3) Flight behavior

In accordance with Rule 4, to avoid being chased or looking for food, birds will fly to other areas regularly, and the migration cycle is set as FQ. When they arrive in another area, they will feed again.

For a Seeker, its flying behavior is described as follows [14]:

$$x_{i,j}^{t+1} = x_{i,j}^t + randn(0,1) * x_{i,j}^t. \tag{18}$$

For a Follower, its flying behavior is as follows [14]:

$$x_{i,j}^{t+1} = x_{i,j}^{t} + rand(0,1) * \left(x_{k,j}^{t} - x_{i,j}^{t}\right) * FL, \tag{19}$$

where $rand(0,1)$ is a standard Gaussian random number, $k \in [1, N]$ and $k \neq i$; $FL(FL \in [0,2])$ indicates the interval between birds flying to a place to look for food.

4.3. Implementation of the proposed model

The flowchart of the prediction model proposed in this study is presented in Figure 3.

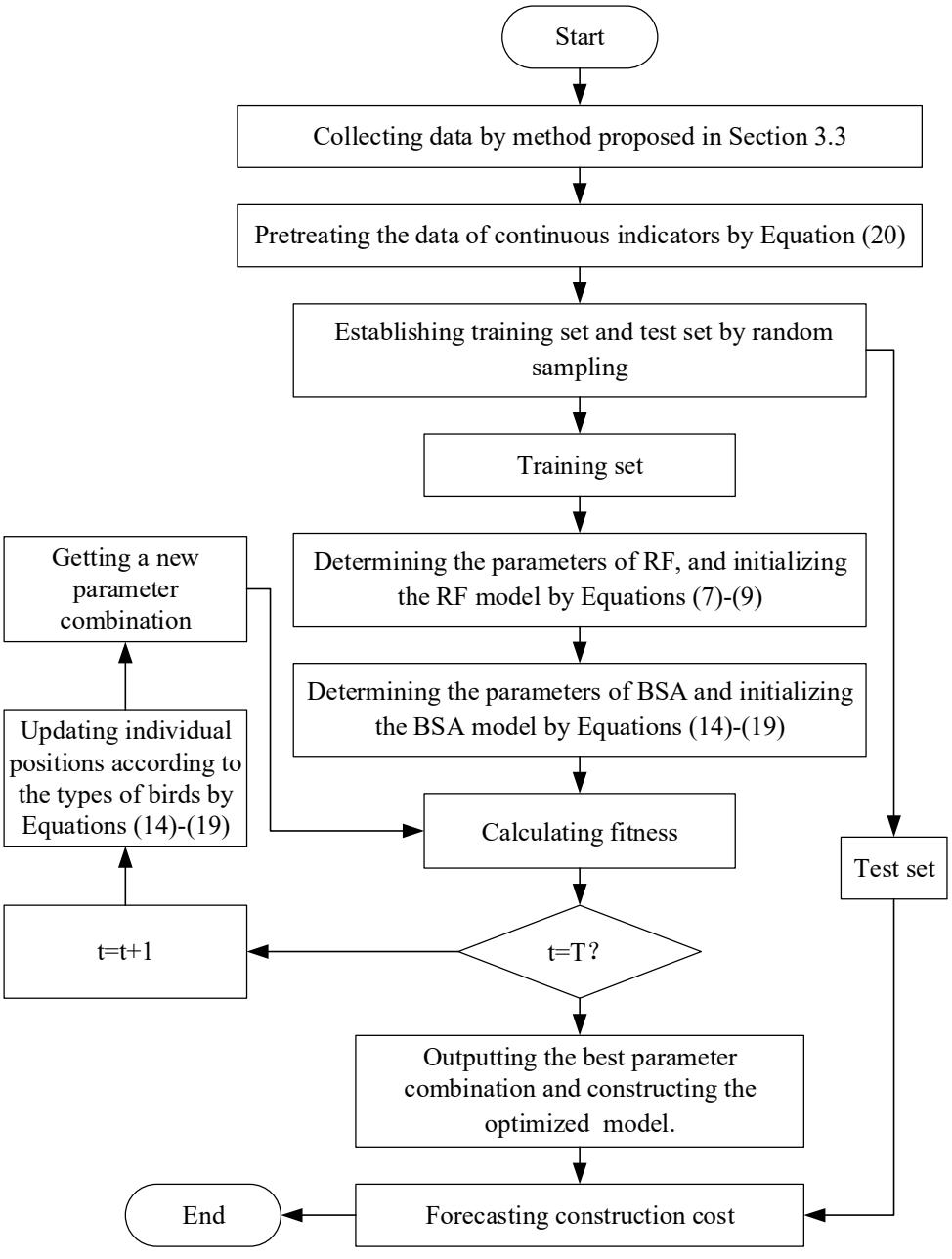

Figure 3. Flow chart of prediction model proposed in this paper.

Step 1. Data collection and preprocessing

As mentioned in Section 3.3, the data for all secondary indicators could be collected. Two types of secondary indicators, discrete indicators (X_{11} and X_{21}) and continuous indicators were identified (11 other secondary indicators). The continuous indicators were normalized to reduce the complexity of the model and improve the prediction accuracy. After normalization, the index value x_i^* is expressed as [29]

$$x_i^* = \frac{x_i - x_{min}}{x_{max} - x_{min}}, \tag{20}$$

where x_i is the value before normalization, x_{min} is the minimum value and x_{max} is the maximum value.

Step 2. Establishing the training set and the testing set.

The training set was used to train the prediction model, and the testing set was used to check its computational performance. The training set was obtained by random sampling, and the remaining data of the sample set were used as the testing set. The common ratios of the training set to the test set were 95%:5%, 90%:10%, 85%:15%, 80%:20% and so on.

Step 3. Determining the calculation parameters of RF and initializing the RF model.

In RF, the range and initial value of the DT and split features had to be set [7,10]. The RF model was initialized by bringing the testing set data and calculation parameters into Eqs (7)–(9).

The MSE of the training samples was selected as a fitness function. When the fitness function was minimum, BSA found the optimal combination of the calculation parameters of RF. The calculation method for MSE is given below [30]:

$$MSE = \sum_{i=1}^{n_1} \frac{(y_i - \hat{y}_i)^2}{n_1}, \tag{21}$$

where n_1 is the number of samples in the testing set; y_i is the predicted value and $\hat{y}_i$ is the real value.

Step 4. Determining the calculation parameters of BSA and initializing the BSA model.

In BSA, the parameters had to be set, including the following: the population size N; the maximum iteration number T; the individual cognitive coefficient C; and the cognitive coefficients S, a_1, a_2 and FL [16]. The testing set data and the calculation parameters were input, and the BSA model was initialized.

According to the optimization flow (Figure 2), the optimal combination of the RF calculation parameters was determined. If the convergence condition was not met, the number of iterations was increased by one, and the individual position and group knowledge of the flock were updated to obtain the new number of DTs and split features. If the convergence condition was met, Step 5 was performed.

Step 5. Outputting the optimal combination of the calculation parameters of RF, including the optimal DT ($nbest$) and the optimal split number ($mbest$).

The construction cost prediction model based on RF was constructed, depending on the best parameter combination.

Step 6. Forecasting the construction cost.

The testing set data was brought into the optimized construction cost prediction model. The $nbest$ DT was randomly selected using the Bootstrap method, and the $mbest$ features were randomly selected to split into leaf nodes. The final prediction result of the construction cost of the test set was determined by taking the arithmetic average of the results of each DT.

5. Case study

5.1. Data acquisition and preprocessing

A total of 48 construction cost data of a construction company in Xinyu City were selected as sample data. Quantitative indicators were obtained by referring to design documents, local yearbooks or project management data. For qualitative data acquisition, 50 engineering experts were invited to assign scores, and 43 valid questionnaires were collected. Personal information on the experts in the 43 valid questionnaires is listed in Table 2.

Table 2. Basic information of 43 experts.

Work unit	Years of work (Year)	Education background	Title	Involved in these projects?
Government 7 (16.28%)	$[0, 5)$ 0 (0%)	College 7 (16.28%)	Junior 0 (0%)	Yes 42 (97.67%)
Subcontractor 9 (20.93%)	$[5, 10)$ 0 (0%)	Bachelor 22 (51.16%)	Senior 43 (100%)	No 1 (2.33%)
Contractor 24 (55.81%)	$[10, 20)$ 12 (27.91%)	Master 10 (23.26%)		-
Owner 3 (6.98%)	$[20, +\infty)$ 31 (72.09%)	Doctor 4 (9.30%)	-	-

The following conclusions are listed in Table 2.

1) All experts of 43 valid questionnaires were from work units closely related to the management of the project construction cost and participated in the construction cost management. These experts were closely related to the object of the case study.

2) 100% of the experts who participated in this questionnaire had working experience of more than 10 years; more than 70% of the experts had working experience of more than 20 years; 93.72% of the experts attained a bachelor's degree or higher; and all of the experts had senior professional titles attached to their names. These findings indicated that most of the experts had solid professional backgrounds in project construction cost management.

Therefore, the questionnaire survey results for the 43 experts were considered qualitatively reliable.

Cronbach's α is the most commonly used method for reliability testing [31]. To quantitatively verify the reliability of qualitative index data, all data for the 43 valid questionnaires were loaded into the SPSS 21.0 software, and the Cronbach's α coefficients of the indexes X_{22}, X_{24} and X_{33} were 0.7511, 0.8124 and 0.7202, respectively. All Cronbach's α coefficients of the three qualitative indexes exceeded 0.7, which met the general requirements of questionnaire reliability testing [31]. Thus, the results of this questionnaire passed the reliability test.

The averages of the three qualitative indicators rated by 43 experts were used as their scores. The original data pertaining to the 48 projects are listed in Table 3. The construction cost (Y) was expressed in millions. Only some data are included in Table 3 because of space constraints.

Table 3. Raw data for the construction cost.

Index	1)	2)	3)	4)	5)	6)	⋯	42)	43)
X_{11}	1	1	3	3	3	1	⋯	1	3
X_{12}	42.5	58.5	103.6	99.5	117.75	26.5	⋯	15.6	45.8
X_{13}	784.41	1674.91	1819.03	2743.51	2321.98	915.47	⋯	687.58	1741.15
X_{14}	720.52	1563.26	3547.14	2402.00	4473.56	845.04	⋯	0	3248.6
X_{21}	4	3	4	3	4	3	⋯	1	3
X_{22}	74.74	57.98	43.12	75.14	25.86	57.00	⋯	63.21	75.14
X_{23}	14.69%	18.20%	16.25%	9.25%	24.89%	12.16%	⋯	15.43%	13.25%
X_{24}	69.05	53.84	84.79	49.93	28.23	62.49	⋯	59.26	48.98
X_{31}	24.52	27.66	12.76	58.07	21.01	40.25	⋯	18.59	15.74
X_{32}	1.74	2.03	1.63	5.50	5.27	1.62	⋯	1.33	0.97
X_{33}	61.58	33.07	72.67	66.95	62.30	69.33	⋯	29.63	73.49
X_{41}	101.4	109.5	116.5	118.4	106.4	118.2	⋯	106. 2	107.3
X_{42}	101.9	112.5	122.2	127.5	121.6	114.9	⋯	129.2	119.0
X_{43}	101.7	114.8	105.0	117.1	105.7	102.6	⋯	114.6	147.18
Y	39.22	25.32	135.63	164.99	149.28	24.14	⋯	17.15	127.48

In Table 3, almost all structural forms were frame structures 1) or shear wall structures 3). This phenomenon was explained in detail. At present, in the field of building engineering, frame structure 1) and shear wall structure 3) are common building structures. In newly-built projects in developed countries, frame structure 1) and shear wall structure 3) account for about 80%, steel structure 2) accounts for about 20%, brick-concrete structure 4), and tube structure 5) rarely appear. In new construction projects in developing countries, frame structures 1) and shear wall structures 3) account for about 50%, steel structures 2) account for about 10%, brick-concrete structures 4) account for about 40% and there are almost no tube structures 5). Among the 43 projects in this paper, there were only 2 steel structures 2) and 5 brick-concrete structures 4), and the remaining 36 were frame structures 1) or shear wall structures 3). We believed that the different structural types of 43 projects were representative.

The data for continuous variables in Table 3 are loaded into Eq (20), and the data after normalization are presented in Table 4, with the data for two discrete indexes (X_{11} and X_{21}) not normalized.

Table 4. Normalized data for the construction cost.

Index	1)	2)	3)	4)	5)	6)	⋯	42)	43)
X_{11}	1	1	3	3	3	1	⋯	1	3
X_{12}	0.82	0.71	0.41	0.43	0.31	0.93	⋯	1.00	0.80
X_{13}	0.88	0.61	0.56	0.28	0.41	0.84	⋯	0.91	0.59
X_{14}	0.89	0.77	0.47	0.64	0.34	0.87	⋯	1.00	0.52
X_{21}	4	3	4	3	4	3	⋯	1	3
X_{22}	0.17	0.39	0.59	0.16	0.82	0.41	⋯	0.32	0.16
X_{23}	0.53	0.34	0.44	0.81	0.00	0.66	⋯	0.49	0.60
X_{24}	0.33	0.52	0.13	0.57	0.84	0.41	⋯	0.45	0.58
X_{31}	0.84	0.79	1.00	0.37	0.88	0.62	⋯	0.92	0.96

Continued on next page

X_{32}	0.76	0.70	0.78	0.00	0.05	0.79	⋯	0.85	0.92
X_{33}	0.21	0.70	0.01	0.11	0.19	0.07	⋯	0.76	0.00
X_{41}	0.85	0.48	0.16	0.07	0.62	0.08	⋯	0.63	0.58
X_{42}	0.95	0.64	0.37	0.21	0.38	0.57	⋯	0.17	0.46
X_{43}	1.00	0.80	0.95	0.76	0.94	0.99	⋯	0.80	0.30
Y	0.90	0.96	0.49	0.36	0.43	0.97	⋯	1.00	0.52

5.2. Acquisition of the optimal combination of parameters

The ratio of the training set to the testing set was 80%:20%. The data for 38 groups were randomly selected from the data in Table 4, and the remaining data for 10 groups were the testing set. Thus, the actual ratio of the training set to the testing set in this study was 79.17%:20.83%. The data for the 38 training sets were entered into the self-compiling program based on Matlab software.

In the RF method, the conditions were set as follows: the maximum range of the DT, 500; the initial value of the DT, 1; the maximum range of split features, 5; and the initial value of split features, 1. In BSA, the conditions were set as follows: population size, 50; maximum number of iterations, 500; the individual cognitive coefficient C and the group cognitive coefficient S were 2, a_1 and a_2 were 1; the convergence error was 10^{-8}, and FL, 0.5–0.9. An extremely wide range for the calculation parameters would reduce the efficiency of the algorithm, whereas an extremely narrow range might not find the optimal solution. Therefore, in this case study, a broad range was used for the calculation parameters of RF and BSA to ensure that the optimal calculation parameters would be determined.

The optimization results for BSA are presented in Figure 4. BSA determined the global optimal solution around the 120th generation.

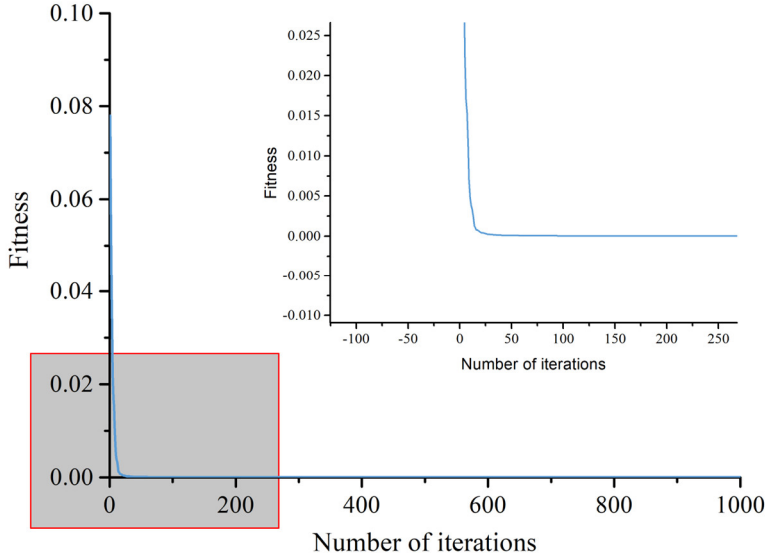

Figure 4. The optimization process of the BSA.

The details of the optimization of BSA are given in Table 5. With the termination iteration requirements considered, BSA found the best combination of RF parameters in the 118th generation.

Table 5. Details of the optimization of BSA.

Iteration (n)	Fitness (n-1)	Fitness (n)	Fitness (n) – Fitness (n-1)	Result
116	4.00741×10^{-5}	4.00436×10^{-5}	$3.05\times10^{-8}<10^{-8}$	Continue
117	4.00436×10^{-5}	3.94791×10^{-5}	$5.65\times10^{-7}>10^{-8}$	Continue
118	3.9101×10^{-5}	3.9101×10^{-5}	$0<10^{-8}$	Continue
1000	3.9101×10^{-5}	3.9101×10^{-5}	$0<10^{-8}$	Stop

In this study, the optimization of the RF algorithm was repeated 1000 times. BSA found the best parameter combination in 123.941 generations on average; the best solution was the 97th generation, and the worst solution was the 161st generation. The standard deviation of the convergent generation was only 14.747, and 99.96% of the 1000 calculations found the same optimal parameter combination.

On the basis of the aforementioned analysis, RF was considered to successfully determine the optimal solution. The optimal parameter combination of RF was as follows: the number of DTs, 124; and the number of split features, 1. The computational performances of BSA and other metaheuristic optimization algorithms are compared in Section 6.2.

5.3. Prediction of the construction cost

The 10 sets of testing data and the optimal calculation parameters of RF were brought into the RF model, and the prediction results of the 10 sets were obtained. When training the model, the normalized real cost was adopted; thus, the predicted value output by the model was also the normalized value. The forecast result was loaded into Eq (20), and the corresponding forecast cost was calculated (Table 6). The maximum errors appear bolded in Table 6.

Table 6. Prediction results for the testing set.

Real cost		Forecast cost		Absolute error	Relative error
Before normalization	After normalization	Before normalization	After normalization		
25.32	0.945	25.586	0.943	0.266	1.05%
135.63	0.199	134.640	0.205	0.990	0.73%
124.99	0.271	125.352	0.268	0.362	0.29%
134.78	0.204	135.934	0.197	1.154	0.86%
68.00	0.656	68.374	0.654	0.374	0.55%
79.54	0.578	78.999	0.582	0.541	0.68%
86.38	0.532	85.398	0.538	1.072	1.24%
115.40	0.335	116.727	0.326	1.327	1.15%
106.38	0.396	105.423	0.403	0.957	0.90%
59.85	0.711	60.305	0.708	0.455	0.76%

Calculated from the normalized data, the absolute error is the absolute value of the predicted value, minus the real value (Table 6). The relative error is the absolute error, divided by the real cost. The prediction results in Table 6 show that the maximum absolute error of 10 test sets is 1.154, and the maximum relative error is 1.24%. In this study, all errors in the prediction results met the requirements

of engineering practice [32].

To further analyze the prediction accuracy of the model, the coefficient of determination (R^2) is selected to analyze the correlation between the actual construction cost and the predicted construction cost. The calculation method for R^2 is given below [33]:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{22}$$

where n is the number of testing data; $\hat{y}_i$ is the predicted value of the i-th test set; y_i is the real value of the i-th test set; and $\bar{y}$ is the average of the real values.

The data in Table 6 were brought into Eq (22), and R^2 of the results predicted using the proposed model was 0.9997. The coefficient of determination between the predicted result and the real value was very close to 1, indicating that the predicted result was almost equal to the real value.

The Coefficient of Variation is also a commonly used error analysis tool, which is equal to the ratio of standard deviation to average value. After calculation, the Coefficient of Variation of the case study object was 6.329383364. This is very similar to the classic research result [34].

All samples were analyzed by tenfold cross-validation to verify the ability of the proposed model to extrapolate. In the tenfold cross-validation, the average absolute error was 0.974, and the average relative error was 1.05%. These results indicate that the proposed model could possessed the ability to generalize, was effective, and could complete the forecasting task with high accuracy. The proposed model is compared with the classical and latest prediction models with respect to calculation performance in Section 6.3.

It is worth mentioning that the ratio of the training set and testing set may have an impact on the prediction results. Therefore, this paper selected 43:5 (89.6%:10.4%), 34:14 (70.8%:29.2%) and 29:19 (60.4%:39.6%). The results of their 1000 repeated calculations were shown in Table 7.

Table 7. Calculation results of different proportions.

Ratio	MAE in 1000 calculations	SDAE in 1000 calculations	SDSD in 1000 calculations	AVSD in 1000 calculations	Ten-fold cross-validation ratio
89.6%:10.4%	1.31%	0.000763471	0.00040985320	0.00718178858	100%
79.17%:20.83%	2.26%	0.001609144	0.00144919951	0.01312937482	100%
70.8%:29.2%	3.47%	0.008471769	0.04223584591	0.03598123199	99.7%
60.4%:39.6%	9.23%	0.026951901	0.06827255313	0.09305018740	86.2%

In the Table 7, MAE is the maximum average error, SDAE is the standard deviation of average errors, SDSD is the standard deviation of standard deviations and AVSD is the average value of standard deviations.

As could be seen from Table 7, with the decrease of the proportion of training sets, the prediction accuracy gradually decreased. When the proportion of training set was as high as 89.6%, the MAE in 1000 calculations was only 1.31%. If the proportion of training set was 60.4%, the MAE in 1000 calculations was increased to 9.23%. In addition, when the proportion of training sets was 70.8 and 60.4%, the model proposed in this paper might not pass tenfold cross-validation every time. Therefore, this paper suggested that the proportion of training set should not be less than 80% when BSA-RF model

was adopted.

It was worth mentioning that each calculation had an average error in 1000 repeated calculations. The average error was the average of the prediction errors of the test sets. Because the samples of the test set were randomly selected in each calculation, the average error of each calculation was different. MAE in 1000 calculations was the MAE of test set prediction results in these 1000 calculations. In this study, it was not to solve the average error of all the calculation results of 1000 calculations, so it was not a definite value.

To analyze the stability of the prediction results, the standard deviations of 1000 calculations under different ratios of the training set and testing set were calculated and shown in Column 3 of Table 7. The calculation results showed that the standard deviation increased rapidly from 0.000763471 to 0.026951901 with the decreasing proportion of training sets. It could be considered that with the decreasing proportion of training sets, the prediction stability of the proposed model was decreasing. This was consistent with the results of similar studies [35,36].

It should be emphasized that in Column 3 of Table 7, the SDAE in 1000 calculations was to directly find the variance of the average error obtained by 1000 repeated calculations, and it studied the deviation between the average errors and their mathematical expectation in 1000 repeated calculations. Specifically, when the ratio was 89.6%:10.4%, five errors and an average error of five test samples were obtained by the one-time repeated calculation, and the SDAE in 1000 calculations was the standard deviation of these 1000 average errors. Similarly, at 70.8%:29.2%, it was also the standard deviation of 1000 average errors. In each of the 1000 calculations, 14 errors and an average error were generated, so it was 1000.

Studying the distribution of standard deviations in 1000 repeated calculations was also helpful to reveal the stability of the proposed model in 1000 repeated calculations. For this reason, the SDSD in 1000 calculations was also calculated. It should be emphasized that standard deviations in 1000 calculations was the standard deviation of test set prediction errors in 1000 repeated calculations, which had 1000 data. SDSD in 1000 calculations was used to describe the distribution of standard deviation in 1000 calculations, and there was only one data. The relevant calculation results were in Column 4 of Table 7. Its calculation process was to solve the standard deviation of all test set errors obtained in a repeated calculation, which had 1000 variance results, and then calculated the standard deviation of these 1000 standard deviation results. Specifically, when the ratio was 89.6%:10.4%, five errors and an average error of five test samples were obtained by repeated calculation. First, the standard deviation of the errors of five test samples was solved, and then the standard deviation of these standard deviations was solved, so that the data in Row 2 and Column 4 in Table 7 was obtained. In Table 7, with the decrease of the proportion of test sets, the number of standard deviations in 1000 calculations was increasing. The calculation results showed that the stability of the prediction results of the model proposed in this paper was positively correlated with the proportion of test sets.

In addition, the AVSD in 1000 calculations under different ratios of the training set and testing set was further calculated. With the proportion of training set reduced from 89.6% to 60.4%, the AVSD in 1000 calculations increased from 0.00718178858 to 0.09305018740. This calculation also showed that the stability of prediction results was decreasing with the decrease of the proportion of test sets.

6. Discussion

6.1. Analysis of the importance of the prediction index

RF can overcome the interference of the possible complex linear relationship between the characteristic variables; however, the influence of the scale of the characteristic variables on the performance of the model still needs to be considered. The index system proposed in Section 3 was adjusted, based on the difference in the degree of importance of the variables, to determine the relatively most appropriate variable scale.

The Mean Decrease Accuracy Index (MDAI) is one of the most common tools used in Variable Importance Analysis [37]. MDAI indicates the extent to which the prediction accuracy decreases when a randomly selected indicator is removed. The larger the MDAI, the stronger its effect, and the greater the importance of the index. The index importance analysis of this study is shown in Figure 5.

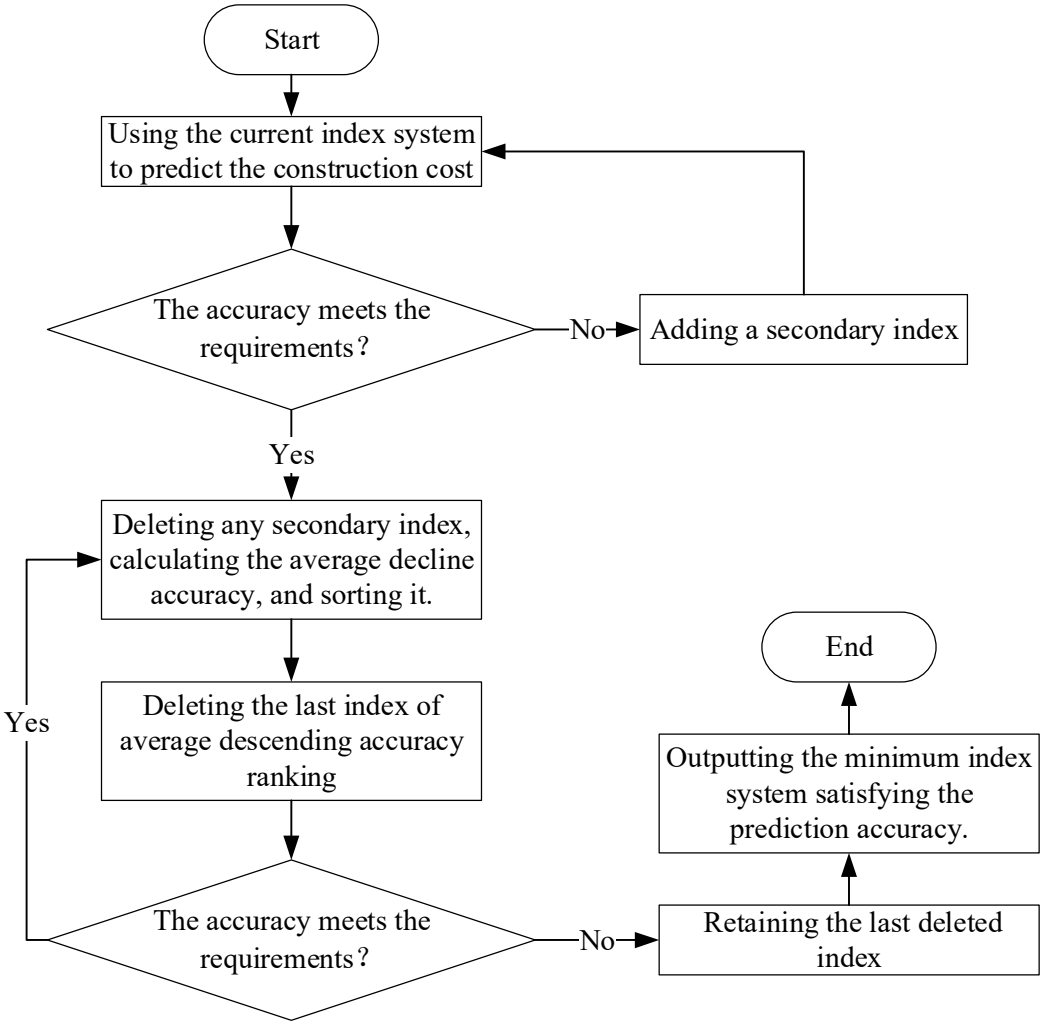

Figure 5. Flowchart of indicator importance analysis by the MDAI.

The prediction accuracy requirement was set to 10%, allowing the average relative error to fall within 10%. The research results in Section 5 indicate that the current predictive index system met the

accuracy requirements, and no secondary indexes had to be added.

Any one of the secondary indexes (Table 1) was deleted, successively, to reconstruct the predictive index system. The newly constructed index system was adopted, and the prediction model described in Section 4 was used for iterative calculations. The calculation results are shown in Figure 6(a). The ordinate in Figure 6 represents the secondary index deleted in this step.

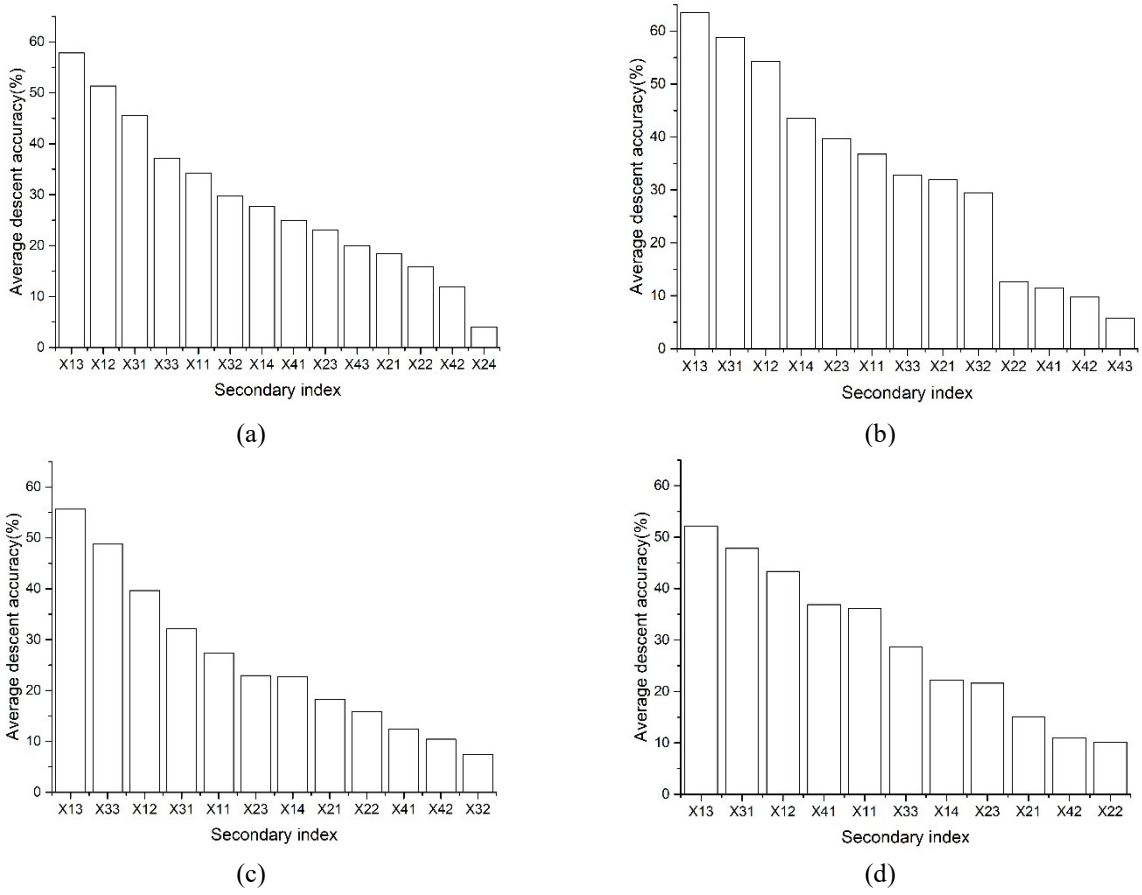

Figure 6. Ranking of indexes importance during gradual dimensionality reduction.

In Figure 6(a), the prediction accuracy of the 14 cases met the requirements when only one index was deleted. The average reduction accuracy of X_{24} was the smallest, only 4.041%. Thus, X_{24} was deleted on the premise of satisfying the prediction accuracy.

On the basis of deleting X_{24}, the successive deletion of a secondary index in Table 1 continued to reconstruct the predictive index system. The calculation results are shown in Figure 6(b). The prediction accuracy of all cases satisfied the requirements. X_{43} had the lowest average reduction accuracy, which was 5.778%; thus, X_{43} was deleted on the premise of satisfying the prediction accuracy.

On the basis of deleting X_{24} and X_{43}, one more secondary index was deleted to reconstruct the predictive index system. The prediction calculation results are shown in Figure 6(c). All 12 cases met the requirements with respect to the prediction accuracy. X_{32} had the lowest average reduction accuracy, which was 7.403%; thus, X_{32} was further deleted.

On the basis of deleting X_{24}, X_{43} and X_{32}, the deletion of one more secondary index (Table 1) proceeded. The prediction calculation results are shown in Figure 6(d). However, the average relative error was 12.34%, which failed to satisfy the preset accuracy requirements.

Therefore, the smallest set of input parameters in the case study was the predictor system, excluding X_{24}, X_{43} and X_{32}.

6.2. Comparison of different optimization algorithms by performance

PSO [20], GA [20], Tabu Search (TS) [38], Simulated Annealing (SA) [39], Ant Colony Optimization (ACO) [40], Differential Evolution (DE) [41], Chicken Swarm Algorithm (CSA) [42], Artificial bee colony algorithm (ABC) [43], Covariance Matrix Adaptation Evolutionary Strategies (CMAES) [44], Wolf pack algorithm (WPA) [45], Whale Optimization Algorithm (WOA) [46] and Artificial Fish School (AFS) [47] were selected to compare their computing performance. The calculation parameters of these algorithms are presented in detail in the corresponding references. All optimization calculations were conducted using the same personal computer (i7-10510U, 1.8GHz, acceleration frequency 4.9GHz quad-core 8MB, 512G SSD, DDR4-2400 8GB). The results of 1000 repetitions of all optimization algorithms are listed in Table 8.

Table 8. Computational performance of different optimization algorithms.

Algorithm	Best result	Worst result	Average result	Standard deviation	Correct rate
BSA	97	161	123.941	14.747	99.96%
PSO	173	249	201.928	13.134	98.69%
GA	267	368	314.428	26.354	98.47%
TS	243	492	298.286	39.582	97.25%
SA	394	647	541.776	30.305	97.58%
ACO	127	247	202.415	35.386	97.89%
DE	194	381	270.897	25.426	97.13%
CSA	122	262	185.527	37.835	98.55%
ABC	234	316	294.893	52.322	93.77%
CMAES	282	405	362.816	46.268	96.98%
WPA	245	439	324.164	47.139	97.68%
WOA	212	648	387.322	50.106	95.15%
AFS	186	326	279.637	27.884	98.76%

As shown in Table 8, BSA determines the best parameter combination of RF in 123.941 generations, on average, which was 77.987 generations faster, compared with PSO, which had the second-highest calculation speed. In the 1000 repeated calculations, the accuracy of BSA optimization was also significantly higher, compared with other metaheuristic optimization algorithms, reaching 99.96%. However, PSO exhibited the smallest standard deviation, indicating that the computational stability of BSA in this case study was slightly lower than that of PSO.

From the perspective of optimization, BSA includes the advantages of PSO and DE, which improved the search efficiency and had relatively good average stability [14]. Therefore, the results in Table 8 are interpretable. The research results in this section are similar to previous research results [20], which proves not only the correctness of the research results but also the more efficient computational performance of BSA.

6.3. Comparison of different prediction models by performance

Data forecasting methods, such as BPNN [14], SVM [5], Stacked Auto-Encoders (SAE) [48] and Extreme Learning Machine (ELM) [49], were selected to compare the computing performance. The parameter settings of the three prediction models referred to the corresponding literature, and BSA was used to determine the best calculation parameters of these three models. In this study, the average value of the relative error, the standard deviation of the relative error, R^2 and the average calculation time were selected as the evaluation indexes of the calculation performance of the prediction model [28]. The results of 1000 calculations of the four prediction models are listed in Table 9.

Table 9. Computational performance of different prediction models.

Model	Average relative error	Standard deviation of relative error	SDSD in 1000 calculations	AVSD in 1000 calculations	R^2	Average calculation time
RF	1.05%	0.000763471	0.00144919951	0.000699460	0.9997	2789.72 s
BPNN	3.87%	0.009122608	0.00789358696	0.010262891	0.9503	4188.89 s
SVM	6.97%	0.012512118	0.00889102508	0.075292102	0.9284	8673.31 s
SAE	5.43%	0.008754721	0.00402412706	0.006578722	0.9674	5742.98 s
ELM	4.84%	0.005957893	0.00227706946	0.006064134	0.9780	1122.95 s

The four prediction models could effectively predict the project construction cost of buildings, but their prediction accuracies noticeably varied. The average relative error of the prediction mode based on RF was only 1.05%. Relative to those of BPNN, SVM and ELM, the average relative error of RF was reduced by 2.82, 5.92 and 3.79%, respectively. The prediction accuracy based on SVM was the lowest, probably because SVM was not suitable for the prediction of small sample data in the current study [5]. The standard deviation of the relative error of the prediction model based on RF was at least one order of magnitude smaller than that of the other three models. Compared with those of BPNN, SVM and ELM, R^2 of RF was increased by 0.0494, 0.0713 and 0.0217, respectively. Among the four prediction models, the average calculation time of ELM was the shortest, only 1122.95 s. The calculation principle of ELM was to randomly select the input weight and analyze it to determine the output weight of the network. Therefore, ELM can provide the ultimate performance in learning speed [49]. On the basis of the aforementioned analysis, RF exhibited better computing performance than BPNN, SVM, or ELM.

In Table 9, the calculation errors of RF, ELM and BPNN were all less than 5%, which all met the practical requirements. The main reason was that BSA was used to determine the optimal combination of calculation parameters of ELM and BPNN, which made the calculation accuracy of ELM and BPNN very good. Therefore, this paper further adopted the GA, a classical meta-heuristic optimization algorithm, to determine the best calculation parameter combination of RF, ELM and BPNN. The average errors of GA-BPNN, GA-ELM and GA- RF were 6.47, 8.53 and 4.39% respectively. This also proved the advanced nature of RF and BSA. As could be seen from Table 9, among many prediction algorithms, the SDSD in 1000 calculations bases on RF was also the smallest. This implied that the prediction result of RF was more stable. In addition, the AVSD in 1000 calculations based on different forecasting models was analyzed in this paper. The average values of standard deviations in 1000 calculations based on RF, BPNN, SVM, SAE and ELM were 0.000699460, 0.010262891,

0.075292102, 0.006578722 and 0.006064134, respectively. Obviously, the average value based on the RF was the minimum, which also exhibited that RF had better prediction performance.

Different algorithms might have different calculation results for different data sets. The calculation result of RF was better than that of BPNN, which is a classical supervised learning algorithm. There are great differences between RF and BPNN in their implementation methods and characteristics, which affect their performance in this case study [50]. RF makes the decision more robust by training multiple decision trees and voting or averaging. Each decision tree is trained with random samples and features, so it has good generalization ability and robustness. In contrast, the performance of BPNN depends on the structure of neurons, activation function, initial weight and other factors and the effect may be poor for more than two types of classification tasks [51].

6.4. Sensitivity analysis of indicators

The construction cost forecasting system is a typical multi-parameter nonlinear system. In order to improve the analysis efficiency and system performance, this section used the Sobol index method to analyze the sensitivity of indicators [52]. Sobol index method is a global sensitivity analysis method proposed by Russian scholar Sobol in 1993. The core of this method is to analyze the sensitivity of parameters by calculating the influence of variance of single parameters and combined parameters on the total variance [53].

In this paper, the BSA-RF model was used as the benchmark model, and the quasi-Monte Carlo method was used for specific sensitivity analysis, and the sensitivity of each index was sorted, as shown in Table 10. In addition, BSA-SAE and BSA-BPNN were also used for sensitivity analysis.

Table 10. Sensitivity coefficient and ranking of secondary indicators.

Index	BSA-RF Sensitivity coefficient	Ranking	BSA-SAE Sensitivity coefficient	Ranking	BSA-BPNN Sensitivity coefficient	Ranking
X_{11}	0.0566	9	0.0741	8	0.0917	5
X_{12}	0.1314	4	0.1103	4	0.1078	4
X_{13}	0.0882	8	0.0487	9	0.0738	8
X_{14}	0.1574	2	0.1302	1	0.1272	2
X_{21}	0.1078	6	0.0870	6	0.0382	10
X_{22}	0.1578	1	0.1301	2	0.1274	1
X_{23}	0.0968	7	0.0782	7	0.0837	7
X_{24}	0.0026	14	0.0081	14	0.0063	14
X_{31}	0.0483	10	0.0430	10	0.0367	11
X_{32}	0.0321	12	0.0290	12	0.0302	12
X_{33}	0.0466	11	0.0385	11	0.0491	9
X_{41}	0.1123	5	0.0911	5	0.0894	6
X_{42}	0.1455	3	0.1178	3	0.1213	3
X_{43}	0.0131	13	0.0139	13	0.0171	13

It could be seen from Table 10 that the global sensitivity analysis results of the three models were basically the same. X_{22}, X_{14} and X_{42} were the most sensitive. Combined with the construction

practice of building engineering, the possible reasons why these three factors had the greatest influence on the construction cost error were as follows. 1) Reasonable resource scheduling could improve the construction efficiency, avoided the waste and lack of resources and reduced the construction cost. However, if the resources were not properly scheduled, it would lead to the delay of working procedure, the delay of construction period and the waste of resources, thus increasing the construction cost. 2) The basement was an important part of the building, and the construction of the basement was relatively difficult, and there were many materials and human resources needed, so the change of the basement area had a great influence on the error of the construction cost. 3) Reasonable site layout could make full use of the site, optimize the structure and design of the building. In addition, X_{24}, X_{32} and X_{43} were the least sensitive. In BSA-RF, their sensitivity coefficients were only 0.0026, 0.0321, and 0.0131 respectively. This was consistent with the analysis results of the importance of indicators in Section 6.1 of this paper.

7. Conclusions

The complex and unpredictable construction costs of building projects were analyzed in this study. A prediction index system of the construction cost was established to address the complicated problem of construction cost. The system included 14 secondary indexes, such as the type of structure, total height and standard floor area. A novel prediction model of construction cost based on RF was proposed to solve the problem of low construction cost prediction accuracy. BSA was used to optimize the number of DTs and split features of RF, considering the tendency of randomly set parameters in RF to cause low prediction accuracy. A case study of a construction company in Xinyu, China showed that the optimal combination of RF calculation parameters was i) the number of DTs set to 124 and ii) number of split features equal to 1. The maximum absolute error of the proposed model was 1.154, and the maximum relative error was 1.24%. These findings confirmed the feasibility of the prediction model in the construction cost prediction of building projects. Compared with the classical metaheuristic optimization algorithm, BSA could more quickly determine the optimal combination of the calculation parameters, on average. Compared with the classical and recent predictive methods, the proposed model presented advantages in prediction accuracy and generalization ability.

The major limitations of the current study were as follows. 1) The index system constructed in this study was only applicable to housing construction projects, and a more general index system of construction cost prediction will be constructed in the future. 2) The case study presented in this article included a small data sample, and the prediction accuracy of the proposed model for large sample data is yet to be determined. 3) The relationship between the prediction index and the construction cost is highly complex. Thus, a hybrid model combining linear and nonlinear methods will be built in the future.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This manuscript is supported by the Science and Technology Project of Jiangxi Province Department of Education (GJJ2203012).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. L. F. Cabeza, L. Rincon, V. Vilarino, G. Perez, A. Castell, Life cycle assessment (LCA) and life cycle energy analysis (LCEA) of buildings and the building sector: a review, *Renewable Sustainable Energy Rev.*, **29** (2014), 394–416. https://doi.org/10.1016/j.rser.2013.08.037
2. M. Y, Cheng, H. C. Tsai, E. Sudjono, Conceptual cost estimates using evolutionary fuzzy hybrid neural network for projects in construction industry, *Expert Syst. Appl.*, **37** (2010), 4224–4231. https://doi.org/10.1016/j.eswa.2009.11.080
3. A. Mahdavian, A. Shojaei, M. Salem, J. S. Yuan, A. A. Oloufa, Data-driven predictive modeling of highway construction cost items, *J. Constr. Eng. Manage.*, **147** (2021), 04020180. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001991
4. A. Mahmoodzadeh, H. R. Nejati, M. Mohammadi, Optimized machine learning modelling for predicting the construction cost and duration of tunnelling projects, *Autom. Constr.*, **139** (2022), 104305. https://doi.org/10.1016/j.autcon.2022.104305
5. M. Juszczyk, On the search of models for early cost estimates of bridges: an SVM-based approach, *Buildings*, **10** (2020), 2. https://doi.org/10.3390/buildings10010002
6. S. Kim, C. Y. Choi, M. Shahandashti, K. R. Ryu, Improving accuracy in predicting city-level construction cost indices by combining linear ARIMA and nonlinear ANNs, *J. Manage. Eng.*, **38** (2022), 04021093. https://doi.org/10.1061/(ASCE)ME.1943-5479.0001008
7. L. Breiman, Random forests, *Mach. Learn.*, **45** (2001), 5–32. https://doi.org/10.1023/A:1010933404324
8. C. Pierdzioch, M. Risse, Forecasting precious metal returns with multivariate random forests, *Empirical Econ.*, **58** (2020), 1167–1184. https://doi.org/10.1007/s00181-018-1558-9
9. J. Yoon, Forecasting of real GDP growth using machine learning models: gradient boosting and Random forest approach, *Comput. Econ.*, **57** (2021), 247–265. https://doi.org/10.1007/s10614-020-10054-w
10. S. Dang, L. Peng, J. M. Zhao, J. J. Li, Z. M. Kong, A quantile regression random forest-based short-term load probabilistic forecasting method, *Energies*, **15** (2022), 663. https://doi.org/10.3390/en15020663
11. G. Tang, B. Pang, T. Tian, C. Zhou, Fault diagnosis of rolling bearings based on improved fast spectral correlation and optimized random forest, *Appl. Sci.*, **8** (2018), 1859. https://doi.org/10.3390/app8101859
12. H. Latifi, B. Koch, Evaluation of most similar neighbour and random forest methods for imputing forest inventory variables using data from target and auxiliary stands, *Int. J. Remote Sens.*, **33** (2012), 6668–6694. https://doi.org/10.1080/01431161.2012.693969
13. X. B. Meng, X. Z. Gao, L. Lu, Y. Liu, H. Z. Zhang, A new bio-inspired optimisation algorithm: Bird Swarm Algorithm, *J. Exp. Theor. Artif. Intell.*, **28** (2016), 673–687. https://doi.org/10.1080/0952813X.2015.1042530

14. C. Zhang, S. Yu, G. Li, Y. Xu, The recognition method of MQAM signals based on BP neural network and Bird Swarm Algorithm, *IEEE Access*, **9** (2021), 36078–36086. https://doi.org/10.1109/ACCESS.2021.3061585
15. Y. Yu, S. Liang, B. Samali, T. N. Nguyen, C. X. Zhai, J. C. Li, et al., Torsional capacity evaluation of RC beams using an improved bird swarm algorithm optimised 2D convolutional neural network, *Eng. Struct.*, **273** (2022), 115066. https://doi.org/10.1016/j.engstruct.2022.115066
16. J. H. Huan, D. H. Ma, W. Wang, X. D. Guo, Z. Y. Wang, L. C. Wu, Safety-state evaluation model based on structural entropy weight-matter element extension method for ancient timber architecture, *Adv. Struct. Eng.*, **23** (2020), 1087–1097. https://doi.org/10.1177/1369433219886085
17. Y. Elfahham, Estimation and prediction of construction cost index using neural networks, time series, and regression, *Alexandria Eng. J.*, **58** (2019), 499–506. https://doi.org/10.1016/j.aej.2019.05.002
18. Y. Cao, B. Ashuri, Predicting the volatility of highway construction cost index using long short-term memory, *J. Manage. Eng.*, **36** (2020), 1–9. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000784
19. S. Mao, F. Xiao, A novel method for forecasting construction cost index based on complex network, *Physica A*, **527** (2019), 121306. https://doi.org/10.1016/j.physa.2019.121306
20. E. Kaya, A comprehensive comparison of the performance of metaheuristic algorithms in neural network training for nonlinear system identification, *Mathematics*, **10** (2022), 1611. https://doi.org/10.3390/math10091611
21. S. Roh, S. Tae, R. Kim, S. Park, Probabilistic analysis of major construction materials in the life cycle embodied environmental cost of Korean apartment buildings, *Sustainability*, **11** (2019), 846. https://doi.org/10.3390/su11030846
22. Y. Liu, X. Y. Wang, H. Li, A multi-object grey target approach for group decision, *J. Grgy Syst.*, **31** (2019), 60–72.
23. T. Moon, D. H. Shin, Forecasting construction cost index using interrupted time-series, *KSCE J. Civ. Eng.*, **22** (2018), 1626–1633. https://doi.org/10.1007/s12205-017-0452-x
24. R. Slade, A. Bauen, Micro-algae cultivation for biofuels: cost, energy balance, environmental impacts and future prospects, *Biomass Bioenergy*, **53** (2013), 29–38. https://doi.org/10.1016/j.biombioe.2012.12.019
25. J. Hong, G. Q. Shen, Z. Li, B. Y. Zhang, W. Q. Zhang, Barriers to promoting prefabricated construction in China: a cost-benefit analysis, *J. Cleaner Prod.*, **172** (2018), 649–660. https://doi.org/10.1016/j.jclepro.2017.10.171
26. L. Liu, D. Liu, H. Wu, J. W. Wang, Study on foundation pit construction cost prediction based on the stacked denoising autoencoder, *Math. Probl. Eng.*, **2020** (2020), 8824388. https://doi.org/10.1155/2020/8824388
27. S. Hwang, Time series models for forecasting construction costs using time series indexes, *J. Constr. Eng. Manage.*, **137** (2011), 656–662. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000350
28. S. Punia, K. Nikolopoulos, S. P. Singh, J. K. Madaan, K. Litsiou, Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail, *Int. J. Prod. Res.*, **58** (2020), 4964–4979. https://doi.org/10.1080/00207543.2020.1735666

29. Z. Zou, Y. Yang, Z. Fan, H. M. Tang, M. Zou, X. L. Hu, et al., Suitability of data preprocessing methods for landslide displacement forecasting, *Stochastic Environ. Res. Risk Assess.*, **34** (2020), 1105–1119. https://doi.org/10.1007/s00477-020-01824-x
30. L. Endlova, V. Vrbovsky, Z. Navratilova, L. Tenkl, The use of near-infrared spectroscopy in rapeseed breeding programs, *Chem. Listy*, **111** (2017), 524–530. Available from: https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/5214159.
31. M. A. Bujang, E. D. Omar, N. A. Baharum, A review on sample size determination for Cronbach's alpha test: a simple guide for researchers, *Malays. J. Med. Sci.*, **25** (2018), 85–99. https://doi.org/10.21315/mjms2018.25.6.9
32. Y. Yu, B. Samali, M. Rashidi, M. Mohammadi, T. N. Nguyen, G. Zhang, Vision-based concrete crack detection using a hybrid framework considering noise effect, *J. Build. Eng.*, **61** (2022), 105246. https://doi.org/10.1016/j.jobe.2022.105246
33. T. Mitsul, S. Okuyama, Measurement data selection using multiple regression analysis for precise quantitative analysis, *Bunseki. Kagaku.*, **60** (2011), 163–170. https://doi.org/10.2116/bunsekikagaku.60.163
34. M. Skitmore, D. H. Picken, The accuracy of pre-tender building price forecasts: an analysis of USA data, in *Information and Communication in Construction Procurement CIB W92 Procurement System Symposium*, (2000), 595–606. Available from: https://eprints.qut.edu.au/9460/.
35. T. Jin, Y. Jiang, B. Mao, X. Wang, B. Lu, J. Qian, et al., Multi-center verification of the influence of data ratio of training sets on test results of an Al system for detecting early gastric cancer based on the YOLO-v4 algorithm, *Front. Oncol.*, **12** (2022), 953090. https://doi.org/10.3389/fonc.2022.953090
36. P. An, X. Li, P. Qin, Y. J. Ye, J. Y. Zhang, H. Y. Guo, et al., Predicting model of mild and severe types of COVID-19 patients using Thymus CT radiomics model: a preliminary study, *Math. Biosci. Eng.*, **20** (2023), 6612–6629. https://doi.org/10.3934/mbe.2023284
37. C. Benard, S. Da Veiga, E. Scornet, Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA, *Biometrika*, **109** (2022), 881–900. https://doi.org/10.1093/biomet/asac017
38. D. Karamichailidou, V. Kaloutsa, A. Alexandridis, Wind turbine power curve modeling using radial basis function neural networks and tabu search, *Renewable Energy*, **163** (2021), 2137–2152. https://doi.org/10.1016/j.renene.2020.10.020
39. K. M. El-Naggar, M. R. AlRashidi, M. F. AlHajri, A. K. Al-Othman, Simulated annealing algorithm for photovoltaic parameters identification, *Sol. Energy*, **86** (2012), 266–274. https://doi.org/10.1016/j.solener.2011.09.032
40. S. Gao, Y. Wang, J. Cheng, Y. Inazumi, Z. Tang, Ant colony optimization with clustering for solving the dynamic location routing problem, *Appl. Math. Comput.*, **285** (2016), 149–173. https://doi.org/10.1016/j.amc.2016.03.035
41. L. Tang, Y. Dong, J. Liu, Differential evolution with an individual-dependent mechanism, *IEEE Trans. Evol. Comput.*, **19** (2015), 560–574. https://doi.org/10.1109/TEVC.2014.2360890
42. Y. Yu, M. Rashidi, B. Samali, M. Mohammadi, T. N. Nguyen, X. X. Zhou, Crack detection of concrete structures using deep convolutional neural networks optimized by enhanced chicken swarm algorithm, *Struct. Health Monit.*, **21** (2022), 2244–2263. https://doi.org/10.1177/14759217211053546

43. C. Zhang, X. Wang, S. Chen, H. Li, X. X. Wu, X. Zhang, A modified random forest based on kappa measure and binary artificial bee colony algorithm, *IEEE Access*, **9** (2021), 117679–117690. https://doi.org/10.1109/ACCESS.2021.3105796
44. M. Reif, F. Shafait, A. Dengel, Meta-learning for evolutionary parameter optimization of classifiers, *Mach. Learn.*, **87** (2012), 357–380. https://doi.org/10.1007/s10994-012-5286-7
45. Y. Dong, J. Du, B. Li, Research on discrete wolf pack algorithm of mutiple choice knapsack problem, *Transducer Microsyst. Technol.*, **34** (2015), 21–23.
46. H. Naseri, H. Jahanbakhsh, A. Foomajd, N. Galustanian, M. M. Karimi, E. O. D. Waygood, A newly developed hybrid method on pavement maintenance and rehabilitation optimization applying Whale Optimization Algorithm and random forest regression, *Int. J. Pavement Eng.*, **2022** (2022). https://doi.org/10.1080/10298436.2022.2147672
47. D. Karaboga, B. Gorkemli, C. Ozturk, N. Karaboga, A comprehensive survey: artificial bee colony (ABC) algorithm and applications, *Artif. Intell. Rev.*, **42** (2014), 21–57. https://doi.org/10.1007/s10462-012-9328-0
48. Y. Yu, J. Li, J. Li, Y. Xia, Z. H. Ding, B. Samali, Automated damage diagnosis of concrete jack arch beam using optimized deep stacked autoencoders and multi-sensor fusion, *Dev. Built Environ.*, **14** (2023), 100128. https://doi.org/10.1016/j.dibe.2023.100128
49. G. Huang, G. B. Huang, S. Song, K. Y. You, Trends in extreme learning machines: a review, *Neural Networks*, **61** (2015), 32–48. https://doi.org/10.1016/j.neunet.2014.10.001
50. M. Kayri, I. Kayri, M. T. Gencoglu, The performance comparison of multiple linear regression, random forest and artificial neural network by using photovoltaic and atmospheric data, in *2017 14th International Conference on Engineering of Modern Electric Systems (EMES)*, (2017), 1–4. https://doi.org/10.1109/EMES.2017.7980368
51. Y. Wang, A. W. Kandeal, A. Swidan, S. W. Sharshir, G. B. Abdelaziz, M. A. Halim, et al., Prediction of tubular solar still performance by machine learning integrated with Bayesian optimization algorithm, *Appl. Therm. Eng.*, **184** (2021), 116233. https://doi.org/10.1016/j.applthermaleng.2020.116233
52. A. B. Owen, Better estimation of small sobol' sensitivity pndices, *ACM Trans. Model. Comput. Simul.*, **23** (2013), 1–17. https://doi.org/10.1145/2457459.2457460
53. S. Kucherenko, O. V. Klymenko, N. Shah, Sobol' indices for problems defined in non-rectangular domains, *Reliab. Eng. Syst. Saf.*, **167** (2017), 218–231. https://doi.org/10.1016/j.ress.2017.06.001

AIMS Press