World Scientific Research Journal Volume 11 Issue 9, 2025
ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

Cost Estimation Method for Missile Equipment in the Development Stage Based on Linear Regression - Random Forest Method

Yuanchang Li, Guiming Chen, Lingliang Xu

Rocker Force University of Engineering, Xian 710025, China

Abstract

This paper addresses the nonlinearity and small sample size characteristics of cost prediction in the missile equipment development stage. By using the linear regression method, 11 cost-related indicators were selected from 22 indicators. A cost estimation model based on the linear regression - random forest method was constructed using 20 sets of historical data from a certain missile development project. The estimation results of the established model were compared and analyzed with those of the neural network model. The root mean square error (RMSE) of the established model was 67.31% lower than that of the neural network model, and the mean absolute error (MAE) of the established model was 63.47% lower than that of the neural network model. This verified the superiority of the linear regression - random forest method in complex nonlinear prediction.

Keywords

Linear regression; Random forest model; Performance metrics; Cost estimation.

1. INTRODUCTION

In order to balance the demand for missile development with the associated costs, China has continuously researched and explored cost estimation methods over the past few decades, undergoing a series of methodological innovations. Traditional cost estimation methods include the engineering estimation method, parametric estimation method, and analogous estimation method, which were widely applied in early stages. For example, Reference [1] utilized the engineering estimation method to decompose remanufacturing costs and proposed a cost estimation scheme for remanufacturing; References [2] and [6] analogously adopted the Burns model to establish a production cost estimation model for an upgraded third-generation fighter jet, improving computational accuracy. These methods are linear approaches with certain predictive capabilities but are unable to simulate more complex relationships between product parameters. Subsequently, grey system theory emerged, focusing on analyzing systems with incomplete certainty and was applied to cost estimation [3]. Reference [4] employed grey system theory to estimate equipment costs, validating its effectiveness with practical samples. Meanwhile, neural networks estimate costs by learning relationships between sample attributes and costs. For instance, an analytical model based on RBF neural network theory [5] was developed to analyze the manufacturing costs of large commercial aircraft, demonstrating high estimation accuracy. However, both grey system theory and neural networks rely heavily on mathematical models and struggle to integrate practical sample data for significant parameter adjustments.

Traditional cost estimation methods exhibit significant limitations when applied to highly complex missile systems involving multidisciplinary and cross-domain technologies. Against this backdrop, there is an urgent need to develop a more scientific and accurate cost estimation

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

method tailored to missile systems. This paper aims to conduct in-depth research on missile cost estimation, exploring methodologies and models suitable for the characteristics of missile systems. The goal is to enhance the accuracy and reliability of cost estimation, providing robust support for missile development and the efficient utilization of military resources.

2. LINEAR REGRESSION AND RANDOM FOREST METHOD

2.1. Linear Regression Method

Linear Regression Method: Reference 1 adopted the linear regression method to establish a missile production cost model (1):

$$y=\begin{cases}-7.4890+5.5647x_2, m\le 30kg,\\ 192.8869+5.4830x_1+0.3692x_2+\\ 61.753x_3-27.2439x_4, m>30kg\end{cases} \tag{1}$$

In the equation, y represents the unit production cost of the missile, in 10,000 RMB; x_1 is the missile's launch mass, in kg; x_2 is the maximum range, in km; x_3 is the maximum Mach number during flight; and x_4 is the guidance error, in meters.

This model, due to the selection of typical and highly cost-related parameters, is relatively simple, computationally efficient, and physically meaningful. The drawback is that it requires fitting in two segments based on the missile's total mass. Nonlinear regression using a quadratic function and other nonlinear modeling approaches have also yielded good results, avoiding the need for segmented fitting in linear regression models. However, these methods are somewhat more complex and better suited for computer-based calculations.

The established quadratic function model (2):

$$\begin{aligned}y=&-221.64x_1-49.448x_2+609.5x_3+168.38x_4\\&-144.07x_5-0.18075x_1^{\ 2}-0.24901x_1x_2+\\&0.8163x_1x_3+72.346x_1x_4+3.5745x_1x_5-\\&0.035167x_2^{\ 2}-0.26965x_2x_3+11.512x_2x_4\\&+4.5989x_2x_5+10.215x_3^2-192.42x_3x_4-\\&60.829x_3x_5-105.38x_4^2+202.59x_4x_5\end{aligned} \tag{2}$$

In the equation: y is the unit production cost of the missile, in 10,000 RMB ; x_1 is the missile's launch mass, in kg; x_2 is the missile's maximum range, in km; x_3 is the warhead mass, in kg; x_4 is the maximum Mach number of the missile flight; x_5 is the missile guidance error, in meters.

The accuracy of this model is generally high, except for a few lightweight and short-range systems; however, the model is structurally complex, incorporating nearly all squared terms and interaction terms of the parameters.

The established nonlinear model (3):

$$\begin{aligned}y=&-71.589+8.692x_1-727.39x_4-\\&9.7104x_1^{0.5}x_3^{0.5}+749.1x_2^{0.05}x_4^{0.9}\end{aligned} \tag{3}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

In the equation: y is the unit production cost of the missile, in 10,000 RMB (yuan); x_1 is the missile's launch mass, in kg; x_2 is the missile's maximum range, in km; x_3 is the warhead mass, in kg; x_4 is the maximum Mach number during missile flight.

The model structure is simple and offers good accuracy, enabling a concise and precise expression of the unit production cost for all air defense missiles using a single equation. Although the selected cost variables in this study exhibit high correlation, other cost variables should also be considered in practical applications based on engineering requirements.

2.2. Random Forest method

Reference [1] provides a detailed introduction to the Random Forest method and related algorithms. In this study, the Random Forest model is employed for cost estimation. The Random Forest model is widely used for classification and regression tasks. It achieves higher accuracy and stability by constructing and averaging predictions from multiple decision trees, thereby enhancing the model's generalization capability. The core steps of the Random Forest model involve building decision trees from the input dataset. In this study, decision trees are constructed using the performance parameters of multiple missile models and their corresponding costs as input features. During the construction of each decision tree, the model randomly samples both data instances and features from the dataset. This process is known as Bootstrap Aggregating (Bagging), as shown in Equation (4). Specifically, the model samples with replacement from the original dataset to construct M independent training sets and trains corresponding decision trees. For each decision tree, the algorithm randomly selects N' candidate features out of the total N features for training. In this way, each decision tree independently predicts the cost, and the Random Forest model computes the average of all tree predictions to produce the final cost estimation.

$$f(x_i)=\frac{1}{n_{tree}}\sum_{i=1}^{n_{tree}} h_i(x_i) \tag{4}$$

In the equation: $f(x_i)$is the prediction result of decision tree t; x_i is the prediction result for instance, $h_i(x_i)$ from decision tree i;n_{tree} is the total number of decision trees.

$$VI_i=\frac{\sum(errOOB_2\text{-}errOOB_1)}{n_{tree}} \tag{5}$$

In the equation: $VI_i$is the importance of environmental covariate i, $errOOB_1$is the out-of-bag (OOB) accuracy of the decision tree.$errOOB_2$is the out-of-bag (OOB) accuracy after adding random perturbation to any feature in the OOB samples.

2.3. Model Evaluation

This study selects three widely recognized evaluation metrics to comprehensively assess the model's accuracy and applicability: the coefficient of determination (R^2), root mean squared error (RMSE), and relative root mean squared error (RRMSE), as defined in Equations (6) to (8).

Coefficient of Determination (R^2) : This metric evaluates the correlation between the model's predicted values and the actual observed values. It reflects the model's ability to explain variations in the data and is a key indicator of how well the model fits the data.Root Mean Squared Error (RMSE) : By directly quantifying the prediction errors, RMSE serves as a critical metric for measuring the model's predictive accuracy.Relative Root Mean Squared Error (RRMSE) : This metric evaluates the relative error of the model, helping to eliminate the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

potential impact of data scale on error assessment. The combined use of these three metrics enables a multidimensional and comprehensive evaluation of the model's performance.

$$R^2=1-\frac{\sum(y_i-\hat{y}_i)^2}{\sum(y_i-\bar{y})^2} \tag{6}$$

$$RMSE=\sqrt{[\frac{1}{n}\sum(y_i-\hat{y}_i)^2]} \tag{7}$$

$$RRMSE=\frac{RMSE}{\bar{y}} \tag{8}$$

In the equation: y_i——The true value of the i-th data point; $\hat{y}_i$——The predicted value of the i-th data point; $\bar{y}$——The mean of all data points.

3. MODEL CONSTRUCTION OF LINEAR REGRESSION - RANDOM FOREST

3.1. Model Construction

Following the principles and methodologies for establishing an evaluation index system, and considering the main factors influencing the performance of missile weapon systems, the performance evaluation index system for missile weapon systems was established through methods such as surveys, expert consultations, induction, and deduction, as shown in [5]。

3.2. Index Screening

To simplify the model, this chapter considers only the characteristic factors affecting cost. Taking the publicly available anti-submarine missile data as an example, we derive 11 indicators influencing cost through linear fitting methods. The validation results are presented in Table 1 below.

Table 1. Results of Linear Fitting Data

Parameter	Weight
'Standby time_h'	92166
' Critical component lifespan _year'	78657
'Maximum flight speed'	75799
' Penetration capability (1-10)'	23955
'Warhead Mass /kg'	20772
'Missile launch mass/t'	9417.3
'Missile diameter/m'	-6530.6
'Missile length/m'	-6948.2
'Maximum range /km'	-14912
'Accuracy/m'	-35278
'Transportation requirements (1-10)'	-38478

3.3. Parameter Settings

The key parameters influencing missile R&D costs are defined as follows:x_1 : Maximum range (km) ; x_2 : Missile length (m) ; x_3 : Missile diameter (m) ; x_4 : Missile launch mass (t) ; x_5 : Warhead

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

mass; x_6 : Maximum flight speed (Mach); x_7 : Accuracy (m); x_8 : Critical component lifespan (years); x_9 : Penetration capability (1-10 scale); x_{10} : Transportation requirements (1-10 scale)

x_{11} : Standby time (hours).These parameters encapsulate critical design and operational attributes that directly or indirectly impact cost estimation.

The input vector is: $\boldsymbol{X}=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9, x_{10}, x_{11}\}$

4. MODEL VALIDATION AND RESULT ANALYSIS

4.1. Model Framework

The basic algorithm flow is shown in Figure 3. It includes the input layer, optimization parameters, prediction layer, and cost output.

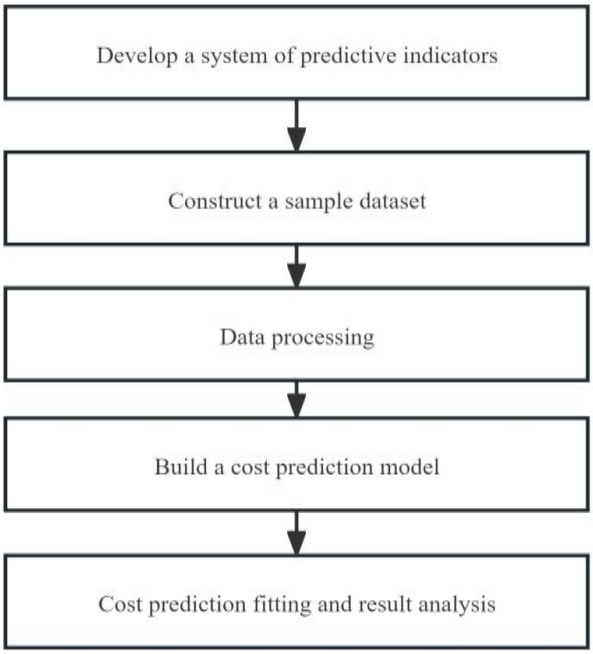

Figure 3. Basic Algorithm Flow

4.2. Data Preprocessing

Perform normalization

$$x_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)}$$

To ensure all parameters fall within the [0, 1] interval and eliminate differences in units of measurement.

4.3. Relevant Factors in Random Forest

Minimize the Mean Squared Error (MSE):

$$\text{Fitness} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

4.4. Cost Estimation

Integrating missile parameters with GRNN outputs, the final cost estimation model is:

$$C = \hat{y}_i \bullet K_{\text{scale}}$$

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

In the equation: K_{scale} As the Economic Adjustment Coefficient (determined based on historical data or industry benchmarks); $\hat{y}_i$ As the normalized cost prediction value output by GRNN.

5. EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

5.1. Experimental Design

missile	Unit Price / USD 10,000	Maximum range /km	Missile diameter /m	Missile length/m	Missile launch mass/t	Warhead Mass/kg	Maximum flight speed	Accuracy/m	Critical component lifespan_year	Penetration capability (1-10)	Transportation requirements (1-10)	Standby time_h
Pershing II	289	120	6.4	0.7	9.4	482	5.3	100	10	6	6	36
Polaris A-2	553	480	19.2	1.8	20.4	3000	6	300	10	6	6	36
Spider SS-23	600	1800	10.5	1.0	7.4	662	12	40	12	7	6	48
DF-21D	651	2200	8.7	1.4	12.9	450	6	1850	10	6	6	36
Red-stone PGM-11	1700	800	11.3	1.0	1	800	6	900	10	6	6	36
Pershing A-1	720	400	7.5	1.0	9.4	785	10	30	10	6	6	36
SS-12thin plate	772	2800	9.4	1.4	13.9	300	10.3	950	12	7	6	48
Pershing A-3	1430	4600	9.9	1.4	16.4	480	10.3	950	12	7	6	48
Scarab SS-21	2000	3100	12.8	1.4	15.0	600	10	10	12	8	7	60
Militia 1LGM-30A	5000	9500	33.0	3.0	285	3000	21	8000	12	7	7	48
SS-18 Satan (nuclear missile)	5418	12070	25.2	4.9	121	3600	24	1850	15	9	8.0	72
Militia 2LGM-30F	8000	3200	19.8	2.4	49.9	1800	20	4000	12	7	7	48
French S-3 missile	8000	10140	17.0	1.7	31.7	250	22	16000	12	7	7	48
Peacekeeper MGM-118	18000	10150	29.9	3.1	99.8	2000	21	2000	12	8	7	60
Scalpel SS-24 (nuclear)	22000	3500	13.8	1.5	31.9	700	15	1000	10	8	6	24

This experiment mainly uses publicly available anti-submarine missile data for validation. Reference [7] selects 20 sets of data from a certain missile development project as the dataset, as shown in Table 2.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

Table 2. Data Sample

missile	Unit Price / USD 10,000	Maximum range /km	Missile diameter /m	Missile length/m	Missile launch mass/t	Warhead Mass/kg	Maximum flight speed	Accuracy/m	Critical component lifespan_year	Penetration capability (1-10)	Transportation requirements (1-10)	Standby time_h
Thor RGM-17	29610	11100	21.6	2.3	88.5	3600	20	120	15	9	8	72
God of the Universe HGM-16	21000	15000	37.0	3.0	217.0	88000	23.5	400	15	9	8	72
Hercules 1HGM-25A	19000	11260	17.6	1.7	31.8	725.7	19	5600	12	8	7	60
Baton SS-6	20000	10000	23.8	2.4	104.5	40500	24	200	12	8	7	60
Hercules 2LGM-25C	22930	13000	18.3	1.7	34.5	955	20	227	12	8	7	60

Among them, the first 17 data samples are used as the training dataset, and the last 3 are used as the validation dataset. The experimental results are compared with those of the genetic algorithm. The evaluation metrics include MAE (Mean Absolute Error), RMSE (Root Mean Squared Error), MAPE (Mean Absolute Percentage Error), and R^2 (Coefficient of Determination).

$$\mathrm{MAPE} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%$$

MSE (Mean Squared Error): This refers to the average of the squared differences between predicted and actual values. The smaller the value, the higher the accuracy of the model.

RMSE (Root Mean Squared Error): This is the square root of the Mean Squared Error (MSE), known as the Root Mean Squared Error (RMSE). A smaller RMSE indicates higher model accuracy.

MAE (Mean Absolute Error): This is the Mean Absolute Error (MAE), which represents the average of the absolute differences between predicted and actual values. It intuitively reflects the average magnitude of prediction errors. A lower MAE indicates higher model accuracy.

MAPE (Mean Absolute Percentage Error): A variant of MAE expressed as a percentage, where lower values indicate higher accuracy. This refers to the Mean Absolute Percentage Error (MAPE).

R^2: This refers to the Coefficient of Determination, commonly denoted as R^2 (R-squared). It measures how well the model's predictions fit compared to the mean of the actual values. The closer the R^2 value is to 1, the higher the accuracy and explanatory power of the model.

5.2. Model Validation and Results Analysis

Training Time: From the training results, it can be seen that the traditional iterative algorithm achieves good prediction performance but takes the longest time. The newly adopted Random Forest method not only provides satisfactory prediction results but also reduces the time by 10.96%.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

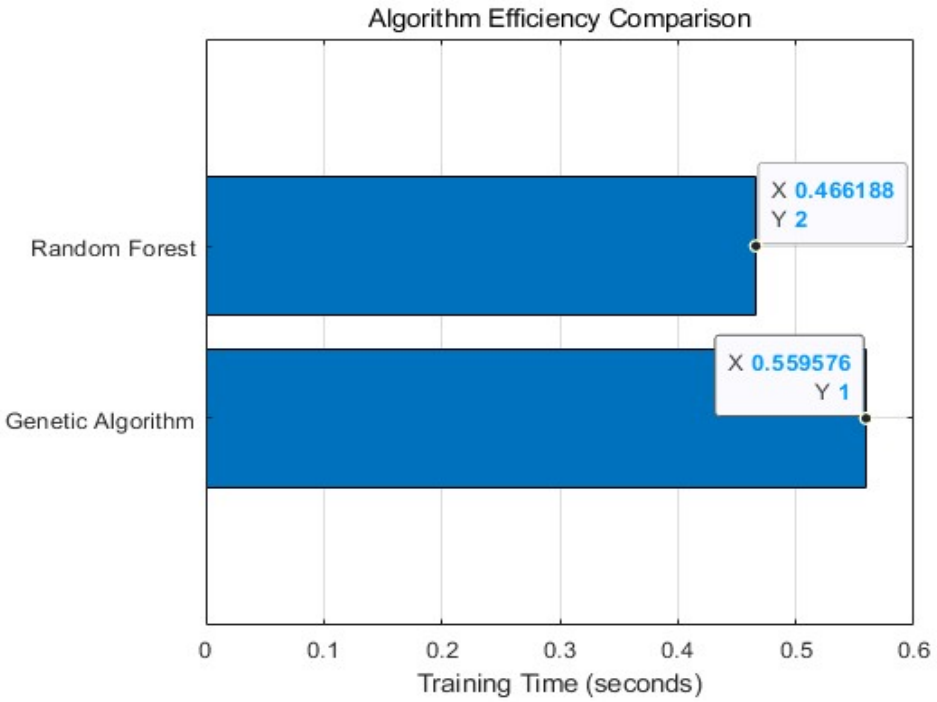

Table 3. Training Time

Convergence Comparison: The traditional genetic algorithm achieves a good convergence result after 79 iterations, whereas the newly improved Random Forest method reaches a similar level of convergence after only 22 iterations.

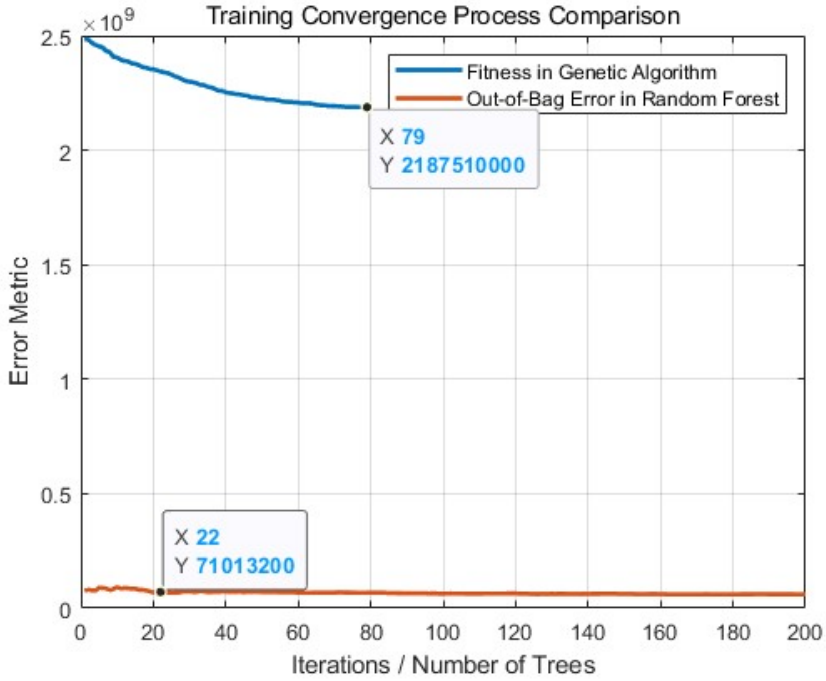

Table 4. Comparison of Iteration Performance

From the sample test results, the Random Forest method shows better prediction performance than the Genetic Algorithm.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

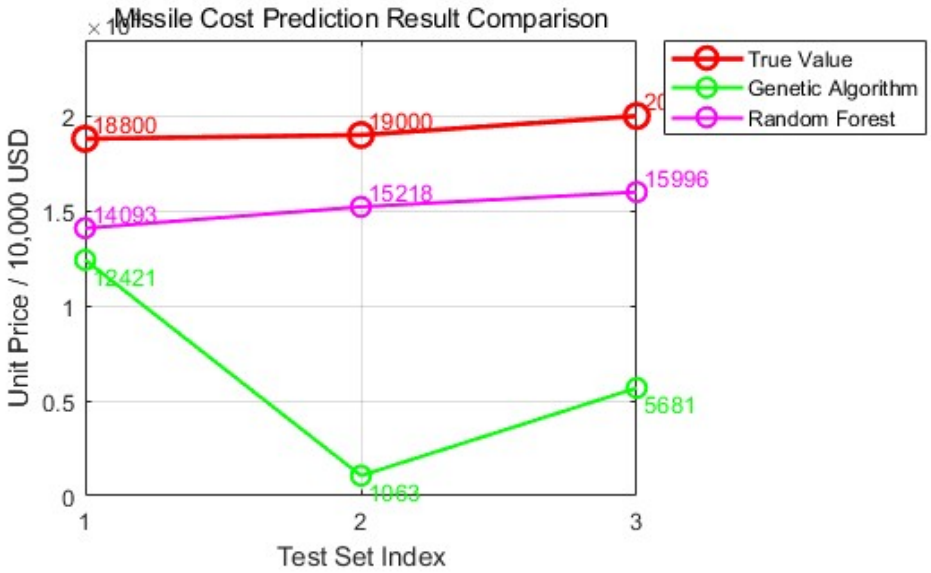

Table 5. Comparison of Fitting Performance on the Test Set

According to the data in Table 6, the Root Mean Squared Error (RMSE) of the Random Forest algorithm is 67.31% lower than that of the Genetic Algorithm, indicating a smaller prediction error deviation. At the same time, the Mean Absolute Error (MAE) of the Random Forest algorithm is 63.47% lower than that of the Genetic Algorithm, suggesting higher prediction stability. In terms of MAPE and R^2 values, the Random Forest algorithm also demonstrates greater accuracy compared to the Genetic Algorithm.

Table 6. Comparison of Evaluation Metrics for the Three Models

	Genetic Algorithm	Random Forest Algorithm
MAE	{'11640.20'}	'4252.15'
RMSE	'13106.57'	'4283.78'
MAPE	'60.18'	'22.09'
R^2	'-622.40'	'-65.60'

6. CONCLUSION

The Root Mean Squared Error (RMSE) of the proposed Random Forest method is reduced by 67.31% compared to the Neural Network approach, and the Mean Absolute Error (MAE) of the Random Forest algorithm is 63.47% lower than that of the Genetic Algorithm. This verifies its superiority in complex nonlinear prediction tasks. Furthermore, by controlling costs—either by selecting equipment with optimal performance or reducing cost investment under certain performance criteria—the practicality of the algorithm is enhanced, providing theoretical support for life cycle cost management of missile systems.

REFERENCES

[1] He Jiawu, Yao Jukun. Methods for Estimating and Predicting Equipment Remanufacturing Costs. Journal of the Academy of Armored Force Engineering, 2010, 24(6):89–91.

[2] Liu Hang, Yan Weitian, Wu Zhe. Aircraft Production Cost Estimation Model for Modern Fighter Jets. Aeronautical Computing Technology, 2011, 41(3): 62–65.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202509_11(9).0001

[3] Guo Jizhou, Song Guibao, Peng Shaoxiong. Grey Modeling and Analysis of Equipment Operational Support Costs. Systems Engineering and Electronics, 2004, 26(1): 64–67.

[4] Cao Guangsheng, Le Guang, Tao Jinliang, et al.Large Airliner Manufacturing Cost Analysis Based on RBF Neural Network. Electronic Design Engineering, 2013, 21(1): 41–46.

[5] Zhao Yueqiang, An Shi, Mai Qiang, et al. Air Defense Missile Cost Modeling Based on Linear and Nonlinear Regression Analysis. Modern Defense Technology, 2019, 47(2): 101–108.HUANG Jun, QU Dongcai, WU Xiaonan. Research on R&D Cost Estimation Model for Military Aircraft Based on RBF Neural Network. Flight Control Technology (Feijihui Jiakong Jishu), 2004(1): 42–46.

[6] Huang Jun, Qu Dongcai, Wu Xiaonan. Research on R&D Cost Estimation Model for Military Aircraft Based on RBF Neural Network.Flight Control Technology (Feijihui Jiakong Jishu), 2004(1): 42–46.

[7] Chen Xi. Research on Pricing Methods for Missile Weapon Systems Based on Performance[D]. Xi'an: Master's Thesis, The Second Artillery Engineering University, 2009.