Application of a PCA-ANN Based Cost Prediction Model for General Aviation Aircraft

XIAONAN CHEN, (Graduate Student Member, IEEE), MINGXU YI, AND JUN HUANG (Member, IEEE)
School of Aeronautic Science and Technology, Beihang University, Beijing 100191, China
Corresponding author: Jun Huang (hunagjun1964@yeah.net)

This work was supported in part by the National Postdoctoral Program under Grant 2017M610740.

ABSTRACT
The major objective of this paper is to build a cost prediction model for general aviation aircraft using artificial neural network (ANN) and principle component analysis (PCA) methods. A total number of 22 samples of general aviation aircraft collected from the literature are utilized to train and test the model. In the PCA, eigenvalues of PC1 and PC2 are 6.987 and 1.529, respectively, indicating that they have the strongest interpretation of the original variable information and are retained as cost influencing variables to train the ANN model. The pure multiple linear regression (MLR), stepwise regression (SR) and ANN models are built respectively for comparison. The comparative results reveal that the ANN method has better estimation effect than MLR and SR models in case of multi-collinearity of data. Combined with PCA, the ANN model is optimized, with MAPE, MAE, R and RMSE values of training and testing samples to be 0.009 and 0.015, 1.222 and 3, 0.9999 and 0.9994, 1.667 and 3.416, respectively. Finally, a more accurate and practical prediction model is developed. More importantly, this research can provide an important reference for general aviation aircraft companies in term of product cost planning and corporate sales strategy.

INDEX TERMS
General aviation aircraft, principle component analysis, artificial neural network, multi-collinearity, multiple linear regression, stepwise regression.

I. INTRODUCTION
The competition of general aviation aircraft production and sales companies mainly comes from the control of aircraft product prices, product quality, and production cycles. Therefore, having an accurate development cost prediction model to quickly formulate reasonable product sales price is extremely important for good decision-making and improving the competitiveness of companies. Although the cost of early manufacturing and design (development cost) of general aviation aircraft only accounts for a part of the life cycle cost (LCC), as proposed by Yeh and Deng [1], it determines the overall trend of LCC. However, due to the difficulty of data collection, late start time, multi-collinearity between variables, the relevant cost prediction models are limited [2], and there is currently no specific model to forecast the general aviation aircraft cost.

In the real world, cost data has the feature of multi-collinearity and is sensitive to noisy, variables number, and specific point. Numerous studies [2]–[4] have demonstrated that the quality of data can affect the prediction accuracy of the model. Since the 1960s, multiple linear regression (MLR) models have been widely utilized to predict aircraft cost. MLR is a statistical analysis system to determine the linear effect of independent variables on dependent variables. In recent decades, researchers have developed many MLR models such as DAPCA, PRICE H and SEER H to predict the aircraft costs [5]. Although MLR is very popular and useful in many cost prediction problems, it still has some drawbacks. For example, as proposed by [6], it is not suitable for systems with a large number of complex variables, and the multi-collinearity among the variables will lead to deviations in the interpretation of the results. However, the authors in [7] reported that the aircraft is a relatively complex system with many parameters that affect the cost and are correlated with each other. In recent years, researchers have tried to overcome the disadvantage of MLR by utilizing artificial neural network (ANN) models. The previous studies [8], [9] have demonstrated that the ANN method can obtain more accurate prediction performance owing to the deep learning ability.
of the model, which makes the ANN method not affected by the multi-collinearity of the cost data. The authors in [1] utilized the ANN method to predict product life cycle cost. The results indicated that the use of ANN method is feasible in the application of cost prediction of airframe structure manufacturing. The authors in [10] utilized the ANO method to predict the equipment cost of liquid crystal display; They reported that the method is fast and independent of data types. The authors in [11] studied the feasibility of ANN method for new construction cost prediction, and their approach can provide satisfactory prediction results. However, in the previous research [5], we have determined that the prediction results of the ANN method is poor in case of small sample size; Other works in [12], [13] also put forward the same conclusion.

In the prediction model, when multiple cost influencing variables are related to each other, that is, the variable information has the feature of multi-collinearity, this will bring many problems to the utilize of regression models, thereby affecting the prediction accuracy of the model. To solve this problem, the easiest way is to remove one or more variables that are less explanatory to the dependent variable and have stronger auto-correlation among the variables. The remaining variables are utilized to build a prediction model, thereby effectively avoiding the effects of multi-collinearity and improving the prediction accuracy of the model. Currently, researchers have utilized forward selection method, optimization subclass selection method, backward selection method and stepwise selection (SR) method to reduce the number of variables. Among them, the SR method is probably one of the most commonly used research practices in substantive and effectiveness studies [14]. It has the advantage of saving time for the individual. The SR is utilized to find and eliminate the parameters that cause multi-collinearity, so that the explanatory variables retained in the model are both important and irrelevant. The authors in [15] presented a SR model to estimate dominant modes, and the results proved that the developed method to be effective. The authors in [16] utilized the SR method to estimate the target parameters; they observed that inaccurate predictions are a result of missing and incomplete variables information. Meanwhile, the problem with this approach is that simply performing linear regression on the retained variables is not fully applicable to changes in the general aviation aircraft market, and the authors in [14] concluded that SR method itself has problems such as incorrect results and the inherent bias in the process itself.

Based on the limitations of the SR method, we hope to explore a more practical and accurate variable filtering method to solve the multi-collinearity problem of data while preserving variable information as much as possible, thereby improving the prediction accuracy of the model. Principal component analysis (PCA) is such an analysis method that can effectively reduce the dimension of parameters and retain the effective influence components in all cost influencing parameters. At present, the PCA method plays an important role in image compression [17], face recognition [18], [19], and image representation [20], [21]. However, it has not received much attention and application in solving the problems of multi-collinearity of data and improving the prediction accuracy of model. To our knowledge, there is no literature that has reported the combination of PCA and ANN methods to solve the problems of cost prediction for general aviation aircraft. Thus, a PCA-ANN based model can be considered to predict the development cost for general aviation aircraft. The main objective of this study is to verify whether the cost influencing parameters obtained through PCA method is significant for ANN modeling, and to verify whether the proposed PCA-ANN model has higher prediction accuracy than other existing models.

This work is organized as follows. Section 2 briefly introduces three cost prediction models: MLR, SR and ANN. The combined method and the modeling procedures of this research is presented in Section 3 followed by a case study to compare the training results of different prediction methods in Section 4. In Section 5, the fitting and prediction results are summarized, and the feasibility of PCA+ANN method is determined. Finally, concluding remarks and future works are discussed in Section 6.

II. COST ESTIMATION METHODS
A. MULTIPLE LINEAR REGRESSION

MLR is a statistical method that simulates the relationship between two or more independent variables and a dependent variable through the form of a linear equation [6]. In practical problems, when the dependent variable is affected by multiple factors, and the correlation between each factor and the dependent variable can be approximately regarded as linear, then a MLR model can be established for prediction and analysis.

Assuming that \( X \) and \( Y \) denote the independent variables \( x_{ij} \) with \( n \) samples and the dependent variable \( y \) with \( n \) samples, respectively.

\[
\begin{bmatrix}
y_1 \\
y_2 \\
\vdots \\
y_n
\end{bmatrix}
= \begin{bmatrix}
x_{11} & \cdots & x_{1m} \\
x_{21} & \cdots & x_{2m} \\
\vdots & \ddots & \vdots \\
x_{n1} & \cdots & x_{nm}
\end{bmatrix}
\begin{bmatrix}
b_0 \\
b_1 \\
\vdots \\
b_m
\end{bmatrix}
\]

The MLR model can be represented by the following equation.

\[
Y = XB + \Delta
\]

where \( B \) and \( \Delta \) denote the regression coefficient vector and the random error vector, respectively.

\[
B = \begin{bmatrix}
b_0 \\
b_1 \\
\vdots \\
b_m
\end{bmatrix}
\quad \Delta = \begin{bmatrix}
\varepsilon_1 \\
\varepsilon_2 \\
\vdots \\
\varepsilon_n
\end{bmatrix}
\]

To predict the value of matrix \( B \), the least squares method was used. \( \hat{B} = (\hat{b}_0, \hat{b}_1, \hat{b}_2, \ldots \hat{b}_m)^T \) denotes the values that
minimize the sum of squares of $\Delta$, which can be expressed as follows.

$$\min \sum_{i=1}^{n} \epsilon_i^2 = \min \epsilon^T \epsilon$$

$$= \min \sum_{i=1}^{n} (Y - XB)^T (Y - XB) \rightarrow 0$$

Thus, we obtain

$$\hat{B} = \begin{bmatrix} \hat{b}_0 \\ \hat{b}_1 \\ \vdots \\ \hat{b}_m \end{bmatrix} = (X^T X)^{-1} X^T Y$$

Finally, the MLR model can be converted into the following equation.

$$\hat{Y} = X\hat{B} = X (X^T X)^{-1} X^T Y$$

B. STEPWISE REGRESSION

The SR method is utilized to individually introduce variables into the model. After that, the model is subjected to the F test. If the initial interpretation parameter is no longer significant after the introduction of the subsequent explanatory variable, it is eliminated. This process is repeated until no significant explanatory parameters can be introduced into the regression model. The detailed description of each phase has been introduced in our previous work [5].

C. ARTIFICIAL NEURAL NETWORK

ANN methods are complex network systems that are widely connected by lots of simple processing units (called neurons). In practical applications, the ANN models trained by backpropagation (BP) algorithms are most commonly used [5]. BP model compares the output and target values and utilizes the gradient steepest descent method to reduce the model error. A typical ANN structure is shown in Fig. 1.

FIGURE 1. Architecture of an artificial neural network model.

As shown in Fig. 1, the ANN method contains three interrelated layers namely input, output and hidden layers. Besides, each layer has its own initial weight, number of neurons, function of neuron and biases. Summation and activation are the two functions performed by neurons in each layer. First, the weighted input is accumulated, and then the accumulated amount is compressed to obtain the output. Information flows between neurons, with each layer receiving signals from one neuron and then transmitting its output to the next. The BP algorithm is utilized to train the ANN model. The algorithm generates a cumulative error between the actual and predicted outputs and propagates the error back to adjust the weight value. The main purpose of network training is to minimize the network error. The criterion for stopping the training is the mean square error (MSE) which is calculated at the output layer [1]. If MSE reaches the set value, the training of the ANN model is terminated; otherwise, the weights between the output and hidden layers are reverse-transmitted and updated using the BP algorithm. The weights between the hidden and input layers are updated in a similar manner until the MSE value reaches the set value. In this way, the purpose of minimizing the error between the desired and predicted output values is achieved.

III. OUR PROPOSED METHOD

The basic idea of PCA is to reduce the dimension of variables. It transforms the multi-indicator problem into fewer comprehensive indicators, and reflects as much of the original information as possible, thereby simplifying the statistical characteristics of the variable system. PCA method uses orthogonal transformation to transform a number of random variables with a certain correlation into a set of relatively small numbers of uncorrelated new variables through a linear combination [22], that is, the principal component (PC). Generally, the principal components (PCs) have the following characteristics:

(1) The number of PCs is much less than that of original variables. Substituting the PCs for the original variables to participate in modeling will greatly reduce the amount of calculation in the model training phase.

(2) The method of obtaining variables by PCs is not the simple filtering of original information like the SR method, but the result of reorganizing the original data. Therefore, most of the original variables information is retained.

(3) The PCs are not related to each other. New indicators (principal components) obtained by PCA can be utilized to deal with the problems caused by variable information overlap and multi-collinearity in data modeling.

In other words, we want to reduce the raw data of $n \times m$ (here, $n$ denotes the sample number, $m$ denotes the number of variables) dimension to a smaller dimension of $n \times k$ ($k$ denotes the number of components). In this new dimension, the individual components will be uncorrelated to each other, while preserving the largest possible variance (the larger the variance, the more information the principal component contains) [23]. Generally, PC1 contains the largest amount of information, that is, PC1 has the largest variance among all original linear combinations, and is called the first principal component. When the information contained in the first principal component is insufficient to represent the
original variable information, continue to select the second principal component PC2. It should be noted that in order to effectively reflect the original information, the existing information in PC1 does not need to be re-represented in PC2, that is, they must be kept independent and uncorrelated. This relationship can be expressed by the covariance matrix $\text{cov}(\text{PC}_1, \text{PC}_2) = 0$. The mathematical expression of the PCA method can be expressed as follows:

$$
\begin{align*}
\text{PC}_1 &= \gamma_1 x_1 + \gamma_2 x_2 + \cdots + \gamma_j x_j + \cdots + \gamma_m x_m \\
\text{PC}_2 &= \gamma_1 x_1 + \gamma_2 x_2 + \cdots + \gamma_j x_j + \cdots + \gamma_m x_m \\
&\vdots \\
\text{PC}_k &= \gamma_1 x_1 + \gamma_2 x_2 + \cdots + \gamma_j x_j + \cdots + \gamma_m x_m
\end{align*}
$$

(7)

where, $x_j (1 \leq j \leq m)$ denotes the original variables, $\gamma_j (1 \leq i \leq k)$ represents the coefficient of the $j$th variable in the $i$th component structure, and $\text{PC}_i$ denotes the $i$th principal component.

In order to predict the vector $\gamma = (\gamma_1, \gamma_2, \cdots \gamma_k)^T$, the covariance matrix can be used. We assume $X_{m \times n}$ denotes the original data matrix, which can be expressed as,

$$
X = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1m} \\
x_{21} & x_{22} & \cdots & x_{2m} \\
& \vdots & \ddots & \vdots \\
x_{m1} & x_{m2} & \cdots & x_{nm}
\end{bmatrix} = (x_1, x_2, \cdots x_m)
$$

(8)

where, $x_j = (x_{1j}, x_{2j}, \cdots x_{nj})^T$. The covariance matrix $(s_{mj})_{k \times k}$ of the original data matrix $X$ can be expressed as,

$$
s_{mj} = \frac{1}{n-1} \sum_{t=1}^{n} (x_{tm} - \bar{x}_m) (x_{mj} - \bar{x}_j)
$$

(9)

Here, $\bar{x}_m$ and $\bar{x}_j$ represent the mean value of the $m$th and $j$th samples, respectively, as follows.

$$
\bar{x}_m = \frac{1}{n} \sum_{t=1}^{n} x_{tm}, \quad \bar{x}_j = \frac{1}{n} \sum_{t=1}^{n} x_{tj}
$$

(10)

We assume that $\lambda_i$ and $\gamma_i$ denote the eigenvalue and orthogonalized unit eigenvector of covariance matrix $s_{mj}$, respectively. Among them, the first $k$ eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k \geq 1$ of $s_{mj}$ are the variances corresponding to the first $k$ PCs. According to the authors in [23], select the first $k$ PCs to achieve the dimension reduction results from $X_{m \times n}$ to $PC_{n \times k}$. The corresponding $\gamma_i$ of $\lambda_i$ is the coefficient of $\text{PC}_i$ on the original variable, as follow.

$$
\text{PC}_i = \gamma_i X
$$

(11)

Take $k$ and $\text{PC}_i$ as the nodes number and input value of the input layer, respectively. $c_{ij}$ represents the weights connecting the hidden and input layers, and $\alpha_j$ denotes the biases on the hidden layer. The $j$th unit input value in the hidden layer can be expressed as:

$$
\text{net}_{\text{hidden}_j} = \sum_{i=1}^{k} c_{ij} \text{PC}_i + \alpha_j
$$

(12)

Consequently, the $u$th unit input value in the output layer can be expressed as,

$$
\text{net}_{\text{output}_u} = \sum_{j=1}^{p} d_{ju} \gamma_j + \beta_u
$$

(13)

where, $p$ denotes the nodes number of the hidden layer, $\gamma_j$ and $\beta_u$ denote the input value and biases on the hidden layer, respectively, $d_{ju}$ represents the weights connecting the hidden and output layers.

The output for hidden and output layers are expressed in Eqs. (14) and (15).

$$
a_j = f(\text{net}_{\text{hidden}_j}) = \sum_{i=1}^{k} c_{ij} \text{PC}_i + \alpha_j
$$

(14)

$$
b_u = f(\text{net}_{\text{output}_u}) = \sum_{j=1}^{p} d_{ju} \gamma_j + \beta_u
$$

(15)

where, $a_j$ represents $j$th node output value in the hidden layer, $f$ is the transfer function utilized for calculation, and $b_u$ is the output value of $u$th node in the output layer.

Then, the network error (MSE) is calculated as,

$$
\text{error} = \frac{1}{2} \sum_{u=1}^{l} (e_u - b_u)^2
$$

(16)

Here, $e_u$ denotes the desired output value, and $l$ is the nodes number of the output layer.

The back-propagation algorithm is utilized as the training algorithm to minimize the network error, $\mu$ denotes the learning rate of the network, the revised $\tilde{c}_{ij}$ and $\tilde{d}_{ju}$ can be expressed as.

$$
\tilde{c}_{ij} = c_{ij} - \mu \frac{\partial \text{error}}{\partial c_{ij}}
$$

(17)

$$
\tilde{d}_{ju} = d_{ju} - \mu \frac{\partial \text{error}}{\partial d_{ju}}
$$

(18)

When the error value or the number of training epoch reaches the network setting value, the training is stopped and the prediction result is obtained. Otherwise, network training continues from Eq. (14). Fig. 2 shows the modeling procedure for combining PCA and ANN methods.

Based on the multi-collinearity feature of general aviation aircraft cost data, MLR, SR and ANN models were established by using pre-processed samples, and the best prediction model (ANN model) was determined by comparing their estimation results. On the other hand, PCA method is utilized to reduce the dimension of cost influencing variables, the applicability of PCA method and the practicability of the combined model (PCA+ANN) are determined by observing the prediction effect of the combination of PCA and a single
optimal prediction model. Fig. 3 shows the analysis method adopted in this research.

IV. DEVELOPMENT COST ESTIMATION CASE STUDY

A. DATA COLLECTION

Unlike military aircraft, general aviation aircraft have a late start and a short development time, and most of the cost data are only for internal reference of the company. This directly brings about the fact that the cost data is difficult to obtain, the lack of personnel specializing in cost prediction, and the lack of a cost prediction model applicable to general aviation aircraft. In this case, we focus on solving the cost prediction problem for general aviation aircraft. Cost data of 22 general aviation aircraft in the open literature [24] were collected. There are numerous cost drivers for general aviation aircraft. Combining the variables introduced by Qu [24] and Chen et al. [5], and referring to the cost drivers of military aircraft, 10 variables are selected as the main driving factors. The collected cost data is shown in Table 1.

B. DATA PRE-PROCESSING

The general aviation aircraft cost data collected in Table 1 may be inaccurate or inconsistent. In the cost prediction problem, each sample is discrete and independent. Errors in individual cost data can also affect the whole trend of the data. It is crucial to pre-process the cost data and remove the incorrect samples before building the model. Data preprocessing can improve the quality of the cost data, thereby improving the prediction accuracy of cost prediction models.

| Sample | Design parameter |
|--------|------------------|
|        | $W_s$ | $L_f$ | $H_f$ | $W_e$ | $W_{to}$ | $W_{AP}$ | $v_{max}$ | $V$ | $C_e$ | $Re$ | $C_{act}$ |
| 1      | 9.88  | 7.715 | 3.035 | 840   | 1400   | 560     | 346       | 290  | 4200 | 1780 | 202   |
| 2      | 9.144 | 6.73  | 2.24  | 379   | 599    | 220     | 219       | 190  | 4724 | 870  | 90    |
| 3      | 11    | 8.28  | 2.72  | 745   | 1111   | 370     | 228       | 226  | 4115 | 1272 | 216   |
| 4      | 10.97 | 8.61  | 2.83  | 966   | 1632   | 666     | 280       | 265  | 4785 | 1278 | 276   |
| 5      | 13.42 | 8.5   | 2.6   | 1200  | 1700   | 500     | 350       | 335  | 5486 | 1857 | 356   |
| 6      | 10.21 | 8.38  | 2.62  | 1148  | 1655   | 507     | 326       | 253  | 5640 | 1702 | 360   |
| 7      | 11.4  | 8.69  | 2.84  | 760   | 1180   | 420     | 287       | 250  | 4267 | 1148 | 220   |
| 8      | 9     | 6.614 | 2.3   | 331   | 600    | 269     | 245       | 225  | 3996 | 1126 | 75    |
| 9      | 11.94 | 8.06  | 1.97  | 800   | 1150   | 350     | 272       | 263  | 5000 | 1056 | 220   |
| 10     | 14.97 | 6.6   | 1.7   | 285   | 473    | 188     | 220       | 200  | 8800 | 1200 | 55    |
| 11     | 6.7   | 5.4   | 1.6   | 275   | 522    | 247     | 317       | 200  | 5000 | 1000 | 45    |
| 12     | 8.68  | 5.02  | 1.71  | 126   | 240    | 114     | 120       | 76   | 3490 | 400  | 11    |
| 13     | 8.464 | 5.955 | 1.987 | 280   | 550    | 270     | 230       | 200  | 5000 | 1000 | 69    |
| 14     | 10.41 | 7.09  | 2.54  | 377   | 599    | 222     | 227       | 163  | 5200 | 724  | 63    |
| 15     | 9.8   | 6.8   | 2.2   | 190   | 420    | 230     | 107       | 62   | 4000 | 500  | 23    |
| 16     | 8.6   | 6.1   | 2.3   | 150   | 330    | 180     | 120       | 90   | 4000 | 550  | 16    |
| 17     | 8.7   | 6.8   | 2     | 210   | 420    | 210     | 160       | 134  | 4000 | 800  | 36    |
| 18     | 10.85 | 7.92  | 2.59  | 939   | 1360   | 421     | 259       | 239  | 5335 | 1480 | 230   |
| 19     | 8.1   | 5.98  | 2.34  | 262   | 450    | 188     | 270       | 204  | 5000 | 800  | 60    |
| 20     | 10.86 | 8.07  | 1.9   | 700   | 1200   | 500     | 407       | 374  | 6100 | 1900 | 292   |
| 21     | 8.66  | 6.17  | 2.67  | 703   | 1089   | 386     | 378       | 310  | 5500 | 1609 | 220   |
| 22     | 10.5  | 7     | 2.45  | 318   | 600    | 282     | 260       | 220  | 4000 | 1065 | 85    |

| Design parameter | Description | Unit | Min. | Max. | Mean |
|------------------|-------------|------|------|------|------|
| $W_s$            | Wing span  | m    | 6.7  | 14.97| 10.12|
| $L_f$            | Length of the fuselage | m | 5.02 | 8.69 | 7.11 |
| $H_f$            | Height of the fuselage | m | 1.6  | 3.04 | 2.32 |
| $W_e$            | Empty weight | kg | 126  | 1200 | 544.7|
TABLE 1. (Continued.) Statistical properties of experimental data.

| Design parameters | Description                   | Unit | Min.  | Max.  | Mean |
|-------------------|-------------------------------|------|-------|-------|------|
| $W_{TO}$          | Maximum take-off weight       | kg   | 240   | 1700  | 876.3 |
| $W_{AP}$          | Available payload             | kg   | 114   | 666   | 331.8 |
| $v_{\text{max}}$ | Maximum level speed           | km/h | 107   | 407   | 255.8 |
| $v$               | Cruise speed                  | km/h | 62    | 374   | 216.7 |
| Ce                | Service ceiling               | m    | 3490  | 8800  | 4892. |
| Re                | Maximum range                 | km   | 400   | 1900  | 1141. |
| $C_{\text{act}}$ | Actual development cost       | RMB  | 11    | 360   | 138.1 |

In this study, we use the judgment ellipse method in SIMCA-P software as the data pre-processing and screening method. 22 samples with 10 independent variables are subjected in the software with the confidence level set to 95%. Fig. 4 shows the 3D (a) and 2D (b) scatter plots of 22 data points obtained by the judgment ellipse method.

As shown in Fig. 4, the 10th sample point is not in the judgment ellipse of 95% confidence interval, which belongs to a specific point and should be removed to improve the overall quality of the data.

C. PREDICTION PERFORMANCE CRITERIA
Several metrics have been utilized to measure the estimation performance of different methods. We adopted the root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and correlation coefficient (R), and the mathematical definitions are shown in Table 2.

D. VARIABLE CORRELATION ANALYSIS
The correlation analysis method is utilized to determine the relationship between the cost influencing variables, and then to analyze whether the data is affected by multi-collinearity. The correlation coefficient (R) between variables ranges from $-1$ to $+1$, which is a value that determines the linear correlation between two parameters. When the negative correlation between the two parameters is stronger, the R value is closer to $-1$. On the contrary, the stronger the positive correlation, the closer R value is to $+1$. Table 3 listed the correlation coefficients between variables.

The higher the R value, the greater the correlation between the two parameters. When the cost impact parameters are strongly correlated, and these parameters are selected as independent variables to build the model, this will affect the model by multi-collinearity and the prediction accuracy will be affected. In general, R$>0.6$ and R$>0.8$ represent a certain and strong correlation between variables, respectively. As listed in Table 3, there are many correlation coefficients between pairwise parameters greater than 0.8, there are

$R (W_S : L_t) = 0.863$, $R (W_c : L_t) = 0.834$, $R (W_{TO} : L_t) = 0.844$, $R (W_{TO} : W_c) = 0.99$, $R (W_{AP} : L_t) = 0.8$, $R (W_{AP} : W_c) = 0.888$, $R (W_{AP} : W_{TO}) = 0.944$, $R (v : v_{\text{max}}) = 0.947$, $R (Re : W_c) = 0.837$, $R (Re : W_{TO}) = 0.859$, $R (Re : W_{AP}) = 0.842$, $R (Re : v_{\text{max}}) = 0.884$, $R (Re : v) = 0.924$

Therefore, great majority of the variables have a strong correlation. We can also draw this conclusion from the scatter plot between the variables. Fig. 5 shows the scatter plots established by some strong correlation variables. Through the above analysis, it can be obtained that the cost influencing variables have a strong cross-correlation relationship, so the data has a multi-collinearity feature.

TABLE 2. Estimation performance criteria.

| Definition                  | Abbreviation | Interpretation       | Formula |
|-----------------------------|--------------|----------------------|---------|
| Mean absolute percentage error | MAPE         | Smaller value is better | $\text{MAPE} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{c_{act,i} - c_{pre,i}}{c_{act,i}} \right|$ |
| Mean absolute error         | MAE          | Smaller value is better | $\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |c_{act,i} - c_{pre,i}|$ |
| Correlation coefficient     | R            | Value closer to one is better | $R = \frac{\sum_i (c_{act,i} - \bar{c}_{act}) (c_{pre,i} - \bar{c}_{pre})}{\sqrt{\sum_i (c_{act,i} - \bar{c}_{act})^2} \sqrt{\sum_i (c_{pre,i} - \bar{c}_{pre})^2}}$ |
| Root mean squared error     | RMSE         | Smaller value is better | $\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (c_{act,i} - c_{pre,i})^2}$ |

Notes: $n$ denotes the number of samples, $\bar{k}$ represents the number of terms in the data table, $c_{act}$ is the value of actual cost, $c_{pre}$ is the value of predicted cost, $\bar{c}_{act}$ denotes the mean of the actual cost values, $\bar{c}_{pre}$ is the mean of the predicted cost values.
TABLE 3. Correlation matrix.

|   | Wₙ | Lₙ | Hₙ | Wₑ | W_TO | W_AP | v_max | v | Ce | Re |
|---|----|----|----|----|------|------|-------|---|----|----|
| Wₙ | 1  |    |    |    |      |      |       |   |    |    |
| Lₙ | 0.86 | 1  |    |    |      |      |       |   |    |    |
| Hₙ | 0.48 | 0.63 | 1  |    |      |      |       |   |    |    |
| Wₑ | 0.73 | 0.83 | 0.61 | 1  |      |      |       |   |    |    |
| W_TO | 0.71 | 0.84 | 0.62 | 0.99 | 1  |      |       |   |    |    |
| W_AP | 0.6  | 0.8  | 0.62 | 0.89 | 0.94 | 1  |       |   |    |    |
| v_max | 0.29 | 0.43 | 0.31 | 0.68 | 0.7  | 0.7 | 1    |   |    |    |
| v | 0.5  | 0.6  | 0.38 | 0.76 | 0.79 | 0.78 | 0.95 | 1 |    |    |
| Ce | 0.25 | 0.32 | 0.05 | 0.56 | 0.54 | 0.46 | 0.73 | 0.68 | 1 |    |
| Re | 0.49 | 0.63 | 0.47 | 0.84 | 0.86 | 0.84 | 0.88 | 0.92 | 0.64 | 1 |

TABLE 4. Eigenvalues of principal components.

| Principal component (PC) | Eigenvalue | Proportion of total variance (%) | Cumulative explained variance (%) |
|--------------------------|------------|----------------------------------|----------------------------------|
| PC1                      | 6.987      | 69.872                           | 69.872                           |
| PC2                      | 1.529      | 15.291                           | 85.163                           |
| PC3                      | 0.631      | 6.311                            | 91.474                           |
| ΣPC4 to PC10             | 0.854      | 8.526                            | 100                              |

E. PCA MODELING ANALYSIS

Through the correlation analysis, we have obtained that the data has a multi-collinearity feature. PCA method reduces the number of variables by dimensionality reduction on the premise of preserving data information as much as possible, so as to solve the problem of multi-collinearity among variables. The eigenvalue, proportion of total variance and cumulative explained variance of each principal component are given in Table 4.

In Table 4, each of the ten PC is a linear combination of the ten original cost influencing variables. In general, eigenvalue ≥ 1 indicates that this PC contains most of the original variables information, which has a strong interpretation of the original cost influencing variable and should be retained. The screen plot of eigenvalues of different PC is shown in Fig. 6.

As shown in Fig. 6, the eigenvalues of PC1 and PC2 are greater than 1, thus they are retained as new cost influencing variables. Table 5 shows the coefficient values (γ_ij) of eigenvectors of PC1 and PC2 calculated using Eqs. (7)-(11).

Thus, the mathematical equations of PC1 and PC2 can be expressed as follows.

PC1 = 0.268Wₙ + 0.317Lₙ + 0.232Hₙ + 0.361Wₑ + 0.367W_TO + 0.352W_AP + 0.306v_max + 0.338v + 0.236Ce + 0.348Re
(19)

PC2 = 0.391Wₙ + 0.351Lₙ + 0.414Hₙ + 0.098Wₑ + 0.093W_TO + 0.074W_AP − 0.419v_max − 0.279v − 0.486Ce − 0.202Re
(20)

F. COST MODELING

After data pre-processing, the 10th group of samples was removed. For the remaining 21 groups of samples, the first 18 samples were randomly utilized to train the models, while the remaining 3 samples were selected to test the model. In the training stage, the MLR, SR, ANN, PCA+ANN methods were utilized to established a development cost prediction model for general aviation aircraft.

1) MLR MODELING

In the MLR training stage, the predicted cost and 10 raw cost influencing parameters are selected as dependent and independent variables, respectively. Therefore, the mathematical
### TABLE 5. Coefficients of the loadings (Eigenvectors) for PC1 and PC2.

| Parameter | PC1   | PC2   |
|-----------|-------|-------|
| $W_\delta$ | 0.268 | 0.391 |
| $L_f$     | 0.317 | 0.351 |
| $H_f$     | 0.232 | 0.414 |
| $W_\tau$  | 0.361 | 0.098 |
| $W_{TO}$  | 0.367 | 0.093 |
| $W_{AP}$  | 0.352 | 0.074 |
| $v_{max}$ | 0.306 | -0.419 |
| $v$       | 0.338 | -0.279 |
| $Ce$      | 0.236 | -0.486 |
| $Re$      | 0.348 | -0.202 |

The equation of MLR is:

$$
c_{pre} = 42.158 - 7.64W_\delta + 6.83L_f - 12.19H_f + 2.57W_e - 2.2W_{TO} + 2.15W_{AP} - 0.11v_{max} + 0.17v - 3.22 \times 10^{-3}Ce - 2.54 \times 10^{-2}Re \quad (21)$$

Based on the MLR method, the comparison between predicted and actual costs is shown in Fig. 7.

2) SR MODELING

After screening the cost influencing parameters through the SR method, only $W_e$ is left to build the model. The mathematical equation can be expressed as:

$$
c_{pre} = -34.001 + 0.32W_e \quad (22)$$

Based on the SR method, the comparison between predicted and actual costs is shown in Fig. 8.

As shown in Fig. 8, the predicted results can only be kept within the general range of accuracy. This is because only uses the filtered $W_e$ variable to establish the model, which avoids the impact of data multi-collinearity on the model, but inevitably leads to the loss of variable information. Therefore, MLR and SR methods have their own shortcomings, which affects the fitting effect of the model.

3) ANN MODELING

In the ANN training stage, ten cost influencing parameters and the predicted cost are include in the input and output layers. According to the hidden layer element selection principle (The minimum number of hidden nodes should be half of the total number of output and input nodes.) proposed by Sodikov [25], nine is selected as the number of hidden layer nodes to get the optimal effect. Purelin and tansig transfer functions are employed for the output and hidden layers, respectively. The mathematical expression of these two transfer functions are shown in Eqs. (23) and (24). Furthermore, the learning rate, maximum operation iteration epoch and target error (MSE) are 0.01, 1000 and 0.001, respectively.

Tansig fuction:

$$f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (23)$$

Purelin function:

$$f(x) = x \quad (24)$$

Based on the ANN method, the comparison between predicted and actual costs is shown in Fig. 9.

Comparing the results of Figs. 8 and 9, we find that the estimation performance of ANN method is significantly better than the results obtained by the SR method. This is because the ANN model can not only better solve the problem of multi-collinearity of data by learning the relationship between variables, but also will not cause the loss of data information. However, this has not always been the case.
When the total number of samples is small, the ANN model will get poor prediction results due to insufficient training. The authors in [5] also mentioned that the ANN model is not suitable for prediction under small sample size.

4) PCA+ANN MODELING
In the PCA+ANN training stage, PC1 and PC2 are include in the input layer, while the number of hidden layer nodes is set to be 5. The remaining settings (output layer, transfer functions, the learning rate, maximum operation iteration epoch and target error) for training the PCA+ANN model are exactly the same as ANN model. Based on the PCA+ANN model, the comparison between predicted and actual costs is shown in Fig. 10.

A comparison of Figs. 9 and 10 indicates that the fitting effect of PCA+ANN method is better than that of the ANN method. In addition, the fitting accuracy of ANN and PCA+ANN models can be compared by establishing the regression graphs of the training data. Figure 11 shows the regression plots of training data for these two models obtained by using matlab software. A higher R value indicates a better fitting effect.

As can be seen from Fig. 11, the R value of PCA+ANN model is 0.999, which is higher than that of ANN model, indicating that its fitting effect is better. When the data has a multi-collinearity feature, ANN method can solve this problem through in-depth learning analysis, and then establish the corresponding nonlinear model to predict the cost of test samples. However, the existence of multi-collinearity is bound to take the model much time to learn the training samples. The variables obtained by PCA are independent of each other, which avoids the influence of multi-collinearity on the model. Therefore, we hypothesized whether the combination of the PCA and ANN methods will reduce the training time of the model, thereby playing an optimization role on the ANN model. The training curves of ANN and PCA+ANN models are shown in Fig. 12.

The training stage is completed when the model reaches the MSE value of 0.001. It can be seen from Fig. 12 that ANN and PCA+ANN models require 680 and 386 epochs to stop the training phase, respectively, which means that the PCA+ANN model spends less time to train the data. This is consistent with our guess: Based on the training data, combined with PCA method can not only get a better fitting effect, but also shorten the training time, so as to achieve the effect of optimizing the ANN model.

V. RESULTS AND DISCUSSION
According to the above analysis of the fitting effect of the four models, the estimation performance reflected by MAPE, MAE, R, and RMSE is as follows: MLR model (0.115, 10.579, 0.9924, 14.028); SR model (0.131, 11.475, 0.9906, 15.604); ANN model (0.079, 6.611, 0.9935, 12.628); and PCA+ANN model (0.009, 1.222, 0.9999, 1.667). Table 6 lists a comparability of the prediction performance of these four models.

As can be seen from Table 6, the PCA+ANN model has the minimum values of MAPE, MAE and RMSE, and the R value is the largest, indicating that the fitting effect on the training data is the best. Instead, the results show that the MLR and SR models reflect relatively poor prediction accuracy on the training samples. This is because the MLR and SR models are affected by the multi-collinearity of data.
and severe loss of variable information, respectively. On the contrary, the ANN model can obtain better prediction accuracy, and is not affected by the multi-collinearity feature of the data and the number of variables.

In the testing stage, the remaining 3 testing samples are predicted by the MLR, SR, ANN, and PCA+ANN models, respectively. The comparison between the predicted and actual cost based on the testing data is shown in Table 7.

According to the prediction results of these four models, the estimation performance reflected by MAPE, MAE, R, and RMSE is as follows: MLR model (0.316, 62.03, 0.9326, 72.373); SR model (0.232, 50.173, 0.9371, 62.715); ANN model (0.086, 16.667, 0.9946, 22.716); and PCA+ANN model (0.015, 3, 0.9994, 3.416). Table 8 lists a comparability of the prediction performance of these four models.

As shown in Table 8, the MAPE, MAE and RMSE values of the PCA+ANN model are 0.015, 3 and 3.416, respectively, which are lower than that of the other three models. This indicates that the PCA+ANN method has the best prediction effect on the testing data among the four models. Meanwhile, the R value of PCA+ANN method is 0.9994, which is much closer to 1 than the other three models. This also implies the best prediction effect. Based on Tables 6 and 8, we make a comprehensive comparison of the prediction performance of
each model. Fig. 13 shows the estimation performance of the four models to compare the predicted results of each model more directly.

It can be seen from Fig. 13 that when the MAPE value is smaller, it is accompanied by a smaller MAE value, a smaller RMSE value and a larger R value, which represents a much better prediction effect. Among the four models, PCA+ANN method has the best prediction effect on both the training and testing samples. This is because the combination of PCA method not only greatly shortens the time for the ANN model to study the training data, but also reduces the complexity of variables, so that the information of variables can be expressed more accurately under the condition of limited sample size, thus improving the prediction accuracy. In addition, the fitting and prediction effects of a single ANN model are superior to MLR and SR models, which reflects the superiority of the ANN model in solving the problem of multi-collinearity of data.

Through the above analysis, we find that when the data has the characteristics of multi-collinearity, compared with the linear model, ANN can better solve this problem and obtain more accurate prediction performance. By combining the ANN model with PCA method, the R values of the training and testing samples increased from 0.9935 and 0.9946 to 0.9999 and 0.9994, respectively, which reflects the optimization of the prediction performance. This proves that the PCA+ANN method is practical and can be utilized to accurately predict the general aviation aircraft development cost.

VI. CONCLUSION

In this research, a distributed PCA+ANN based method is developed for estimating the general aviation aircraft development cost. Through comparative analysis with the prediction results of the single MLR, SR, ANN model, the following conclusions are drawn.

1) The prediction effect of MLR method is affected by data multi-collinearity and the prediction accuracy is poor. At the same time, SR method can solve the problem of multi-collinearity of cost data, but it will inevitably cause a large amount of loss of variable information by simply deleting variables and then modeling, which will affect the prediction accuracy. Compared with the above two methods, the ANN method is not affected by the multi-collinearity of the data and the number of variables, a more satisfactory prediction result is obtained through deep-learning of the data information.

2) Combined with PCA method, the complexity of variable information is reduced, and the ANN model is optimized. The training time of the PCA+ANN model is shortened from the original 680 epochs to 386 epochs, which greatly improves the model fitting speed. At the same time, compared with the ANN model, the combined model has more accurate prediction performance for the training and testing samples, which is reflected by having smaller MAPE, MAE, RMSE values and higher R value.

As part of future work, more cost data will be collected to explore whether the PCA+ANN method is common and whether it will be limited by data information or sample size. In addition, the SR+ANN model will be established to explore the applicable scope of SR+ANN and PCA+ANN methods. We are committed to building a more accurate and convenient general aviation aircraft cost estimation software to help companies improve their competitiveness while contributing to the promotion of a green society.

APPENDIX

ABBREVIATIONS

| Abbreviation | Description |
|--------------|-------------|
| PCA          | Principal component analysis |
| ANN          | Artificial neural network |
| BP           | Back-propagation |
| MLR          | Multiple linear regression |
| SR           | Stepwise selection |
| PC           | Principal component |
| WS           | Wing span |
| Lf           | Length of the fuselage |
| Hf           | Height of the fuselage |
| We           | Empty weight |
| WTO          | Maximum take-off weight |
| WAP          | Available payload |
| Vmax         | Maximum level speed |
| v            | Cruise speed |
| Ce           | Service ceiling |
| LCC          | Life cycle cost |
| Re           | Maximum range |
| c<act>       | Actual development cost |
| c<pre>       | Predicted development cost |
| MAPE         | Mean absolute percentage error |
| MAE          | Mean absolute error |
| R            | Correlation coefficient |
| RMSE         | Root mean squared error |
| n            | Number of samples |
| k            | Number of terms in the data table |
| c<act>       | Mean of the actual cost values |
| c<pre>       | Mean of the predicted cost values |
| γ<ij>        | Coefficient values |
| PC<sub>s</sub> | Principal components |
| λ            | Eigenvalue |
| γ            | Eigenvector |

ACKNOWLEDGMENT

The authors would like to thank all of the team members who assisted with the data collection. Also, the authors would like to acknowledge the supports from Hefei General Aviation Research Institute, Beihang University.

REFERENCES

[1] T.-H. Yeh and S. Deng, “Application of machine learning methods to cost estimation of product life cycle,” Int. J. Comput. Integ. Manuf., vol. 25, nos. 4–5, pp. 340–352, Apr. 2012.
[2] X. Chen, J. Huang, and M. Yi, “Cost estimation for general aviation aircrafts using regression models and variable importance in projection analysis,” J. Cleaner Prod., vol. 256, May 2020, Art. no. 120648.
[3] J. S. Challa, P. Goyal, S. Nikhil, A. Mangla, S. S. Balasubramaniam, and N. Goyal, “DD-Rtree: A dynamic distributed data structure for efficient data distribution among cluster nodes for spatial data mining algorithms,” in Proc. IEEE Int. Conf. Big Data (Big Data), Washington, DC, USA, Dec. 2016, pp. 27–36.

[4] A. Jungeja and N. N. Das, “Big data quality framework: Pre-processing data in weather monitoring application,” in Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput. (COMITCom), Faridabad, India, Feb. 2019, pp. 559–563.

[5] X. Chen, J. Huang, and M. Yi, “Development cost prediction of general aviation aircraft projects with parametric modeling,” Chin. J. Aeronaut., vol. 32, no. 6, pp. 1465–1471, Jun. 2019.

[6] W. Jiang, X. Xing, S. Li, X. Zhang, and W. Wang, “Synthesis, characterization and machine learning based performance prediction of straw activated carbon,” J. Cleaner Prod., vol. 212, pp. 1210–1223, Mar. 2019.

[7] S. Altavilla, F. Montagna, and M. Cantamessa, “A multilayer taxonomy of cost estimation techniques, looking at the whole product lifecycle,” J. Manuf. Sci. Eng., vol. 140, no. 3, p. 19, Mar. 2018.

[8] N. Jin and D. Liu, “Wavelet basis function neural networks for sequential learning,” IEEE Trans. Neural Netw., vol. 19, no. 3, pp. 523–528, Mar. 2008.

[9] Y. Huang, L. F. Capretz, and D. Ho, “Neural network models for stock selection based on fundamental analysis,” in Proc. IEEE Can. Conf. Electr. Comput. Eng. (CCECE), Edmonton, AB, Canada, May 2019, pp. 1–4.

[10] J.-S. Chou, Y. Tai, and L.-J. Chang, “Predicting the development cost of TFT-LCD manufacturing equipment with artificial intelligence models,” Int. J. Prod. Econ., vol. 128, no. 1, pp. 339–350, Nov. 2010.

[11] H. M. Günaydın and S. Z. Doğan, “A neural network approach for early cost estimation of structural systems of buildings,” Int. J. Project Manage., vol. 22, no. 7, pp. 595–602, Oct. 2004.

[12] T. Agarwal, P. Niknejad, M. R. Barzegaran, and L. Vanfretti, “Multi-level time-sensitive networking (TSN) using the data distribution services (DDS) for synchronized three-phase measurement data transfer,” IEEE Access, vol. 7, pp. 131407–131417, 2019.

[13] W. Song, Y. Fu, Q. Chen, L. Li, and C. Zhang, “ANN based adaptive successive cancellation list decoder for polar codes,” in Proc. IEEE 13th Int. Conf. ASIC (ASICON), Chongqing, China, Oct. 2019, pp. 1–4.

[14] K. Y. Chan, H. K. Lam, T. S. Dillon, and S. H. Ling, “A stepwise-based fuzzy regression procedure for developing customer preference models in new product development,” IEEE Trans. Fuzzy Syst., vol. 23, no. 5, pp. 1728–1745, Oct. 2015.

[15] N. Zhou, J. W. Pierre, and D. Trudnowski, “A stepwise regression method for estimating dominant electromechanical modes,” IEEE Trans. Power Syst., vol. 27, no. 2, pp. 1051–1059, May 2012.

[16] P. Tahmassabi, F. Javadpour, and M. Sahimi, “Data mining and machine learning for identifying sweet spots in shale reservoirs,” Expert Syst. Appl., vol. 88, pp. 435–447, Dec. 2017.

[17] L. Kanwal and M. U. Shahid, “Denoising of 3D magnetic resonance images using non-local PCA and transform-domain filter,” in Proc. 19th Int. Multi-Topic Conf. (DMIC), Islamabad, Pakistan, Dec. 2016, pp. 1–5.

[18] K. Rujirakul, C. So-In, B. Arnonkijapintich, K. Sunat, and S. Poolsguang, “PDF-PCA: Parallel fixed point PCA face recognition,” in Proc. 4th Int. Conf. Intell. Syst., Modeling Simulation, Bangkok, Thailand, Jan. 2013, pp. 409–414.

[19] S. Rajendran, A. Kaul, R. Nath, A. S. Arora, and S. Chauhan, “Comparison of PCA and 2D-PCA on Indian faces,” in Proc. Int. Conf. Signal Propag. Comput. Technol. (ICSPCT), Ajmer, India, Jul. 2014, pp. 561–566.

[20] Z. Zhu, A. Su, H. Liu, Y. Shang, and Q. Yu, “Vision navigation for aircrafts based on 3D reconstruction from real-time image sequences,” Sci. China Technol. Sci., vol. 58, no. 7, pp. 1196–1208, Jul. 2015.

[21] M. Arafa and Q. A. Moghli, “Efficient image recognition technique using invariant moments and principle component analysis,” J. Data Anal. Inf. Process., vol. 5, no. 1, pp. 1–10, 2017.

[22] A. Alver and Z. Kazan, “Prediction of full-scale filtration plant performance using artificial neural networks based on principal component analysis,” Separat. Purification Technol., vol. 230, Jan. 2020, Art. no. 115868.

[23] A. Lee, S. E. Kim, and K.-D. Suh, “Estimation of stability number of rock armor using artificial neural network combined with principal component analysis,” Proc. Eng., vol. 116, pp. 149–154, 2015.

[24] J. W. Qu, “The world general aviation aircraft,” in China: Aviation Industry Press, 2014, pp. 20–73.

[25] J. Sodikov, “Cost estimation of highway projects in developing countries: Artificial neural network approach,” J. Eastern Asia Soc. Transp. Stud., vol. 6, pp. 1036–1047, 2005.