Aircraft Noise Prediction Based on Machine Learning Model

FENG Hao, *ZHOU Yadong**, *DING Cong*, *ZENG Weili*, *GUO Wentao*

College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, P.R. China

(Received 23 June 2023; revised 7 September 2023; accepted 1 October 2023)

Abstract: In order to explore the aircraft noise prediction methods beyond the best practice model and scientific model, this paper uses multiple linear regression model and random forest regression model to predict the aircraft noise value of Seattle-Tacoma International Airport in the summer of 2020—2022. The experiment confirm the feasibility and advantages of the machine learning model in aircraft noise prediction tasks and find that the mean R^2 predicted by the random forest regression model is 74.469%, 5.361% higher than that of the multiple linear regression model. The mean RMSE predicted by the random forest regression model is 0.814, 0.106 lower than that of the multiple linear regression model.

Key words: aircraft noise emissions; aircraft noise prediction; multiple linear regression; random forest regression

CLC number: U8 **Document code:** A **Article ID:** 1005-1120(2023)S2-0054-08

0 Introduction

Aviation noise refers to the noise generated by various aircraft in the airport and its adjacent areas, most of which are aircraft noise, mainly airframe noise, fan noise, and jet noise[1]. The World Health Organization believes noise pollution is the second largest environmental risk factor next to air pollution[2], and aviation noise is the third largest source next to road traffic and railway traffic noise[3]. Relevant medical research shows that long-term exposure to aircraft noise will increase the risk of cardiovascular diseases such as hypertension, coronary heart disease, and heart failure[4-5]. Moreover, there is sufficient evidence to show that aircraft noise will not only affect children's learning ability and cognitive skills[6] but may even lead to children's mental health problems[7].

Regarding environmental costs, the social cost of noise in Taipei Songshan Airport (TSA) is, on average, three times that of airport emissions. In 2015, the social cost of aircraft noise in TSA is as high as 33 million euros[8]. When considering the expected environmental impact per capita, the aviation noise damage within 6 km of the airport is dominant, while people living within 5 km of the commercial airport bear the cost of aircraft noise disproportionately[9]. The cost per decibel of aviation noise is higher than road and railway noise[10]. For each decibel increase in aircraft noise, the cost is 0.4% to 0.6% higher than road noise[11].

Aviation noise has become an important social issue. On the one hand, scientific and reasonable noise prediction can provide a theoretical basis for land planning near the airport, thus ensuring community residents' physical and mental health and improving the quality of life. On the other hand, it can reduce airport noise control costs and social property losses. That will undoubtedly enhance the public's recognition of the civil aviation industry, promote the establishment of environmentally friendly urban airports, and promote the sustainable development of the civil aviation transport industry.

Therefore, based on the data of noise monitoring points near Seattle-Tacoma International Airport (SEA), this paper uses Python programming language to build multiple linear regression (MLR) and random forest (RF) regression models for noise prediction. It can not only investigate the feasibility and advantages of the machine learning model in

*Corresponding author, E-mail address: yzhou@nuaa.edu.cn.

How to cite this article: FENG Hao, ZHOU Yadong, DING Cong, et al. Aircraft noise prediction based on machine learning model[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2023, 40(S2): 54-61.

http://dx.doi.org/10.16356/j.1005-1120.2023.S2.008

noise prediction but also provide technical support for noise prediction of similar airports.

1 Literature Review

Aeronautical noise prediction models are mostly divided into two categories: Best practice model and scientific model. The former regards the aircraft as a single noise point source, which uses a database and sound propagation law to predict noise; the latter is a multi-source parameter model, which uses a component sound source model (including engine noise and fuselage noise) and sound propagation model to predict noise[12]. However, the best practice model will not be satisfactory unless its prediction results are modified many times to improve the accuracy; the scientific model is very complex, the required component parameters are difficult to be obtained, and the cost is high. With the development of the artificial intelligence technology, some scholars have made exciting progress in using machine learning models to predict aircraft noise. Zellmann et al.[13] applied two multiple linear regression models with different complexity and applicability to the fuselage and engine noise sources, respectively, and established a general aircraft noise emission model for 19 turbofan engine aircraft. Based on radar data, noise readings, trajectory data, weather data, and other data for more than ten months, the LSTM model developed by Vela et al.[14] accurately predicted aviation noise at ground monitoring points near Washington National Airport, with a mean absolute error of 2.3 dB. Revoredo et al.[15] established a multilayer feedforward neural network model, which considers the dynamic relationship between flight parameters and 4D tracks. It is found that the model can not only compare the noise impact of different arrival and departure procedures but also evaluate the overall noise level in the airport.

2 Research Materials and Methods

2.1 Overview of airport and monitoring points

SEA Airport is located in Seatac, Washington State. The airport has three runways: 16L/34R, 16C/34C, and 16R/34L, covering a total area of 2500 acres, about 40% of the area of Seatac. In 2019, SEA's passenger flow is 51.829 million person-times, making it the eighth busiest airport in the United States. At least 32 000 residents are affected by airport aircraft noise. The noise monitoring point is located at 1217 S 207th St, SeaTac, which is the No.17 monitoring point in the flight track monitoring system established by Seattle Port. As seen in Fig.1, the monitoring is in the south of the airport, 3.8 km away from the center of the airport, and 1.6 km from the nearest runway.

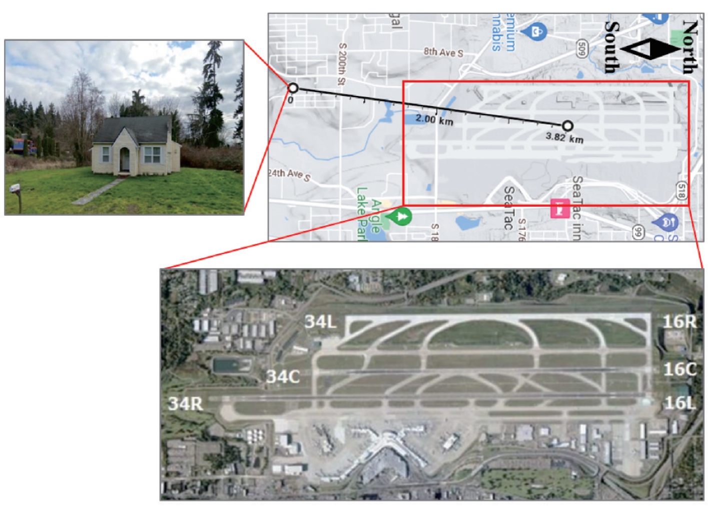

Fig.1 Overview of SEA airport and noise monitoring point

2.2 Dataset

The data set contains 20 869 data in six dimensions, including operation, aircraft model, engine model, runway, altitude, and noise, from July to September 2020 to 2022. The OrdinalEncoder (·) function in the preprocessing module of the sklearn library is used to convert the categorical features to integer codes. The frequency histogram of each dimension is shown in Fig.2.

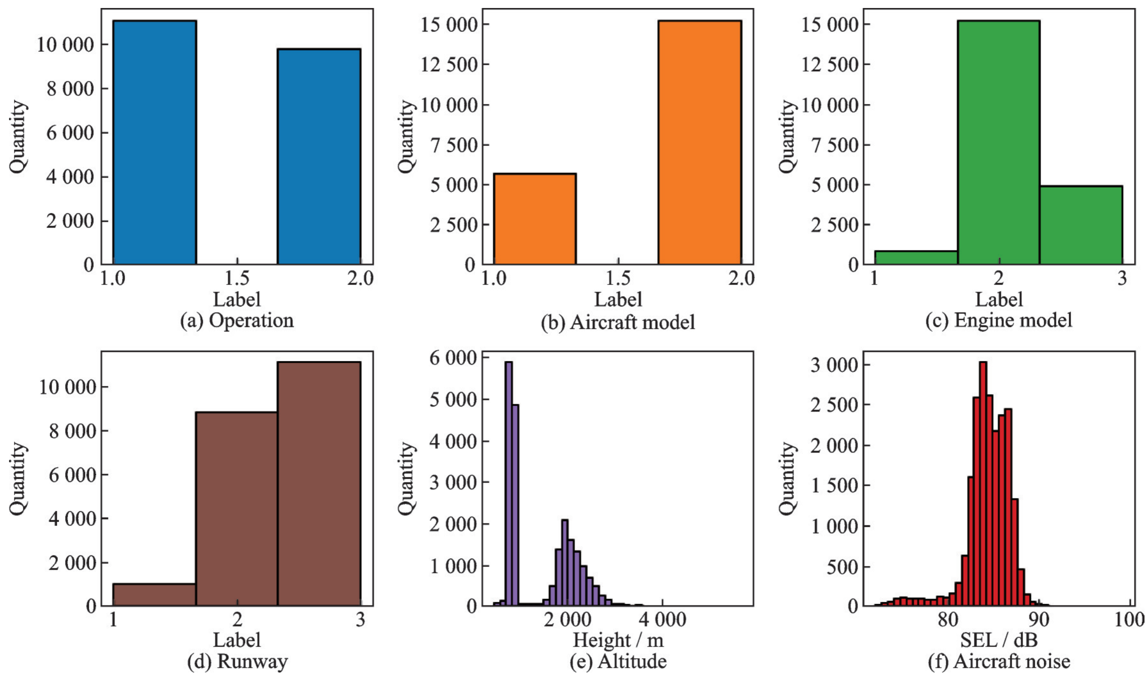

Fig.2 Feature and label histogram

The numbers 1 and 2 in "Operation" represent approach and departure; the numbers 1 and 2 in "Aircraft model" represent DH8D, E75L; the numbers 1, 2 and 3 in "Engine model" represent 2 PWC PW150A, CF34-8E5G01, PW150A; the numbers 1, 2 and 3 in "runway" represent 16C, 16L and 34L. The altitude of most samples is concentrated around 800 m and 2 000 m, and the noise of most samples is mainly distributed between 84—87 dB. The noise unit is sound exposure level (SEL), and its calculation formula is shown in

$$\mathrm{SEL}=\int_{t_0}^{t_f}10^{0.1L_{\mathrm{A,max}}}\mathrm{d}t \tag{1}$$

Where $L_{\mathrm{A,max}}$ represents the maximum A-weighted sound level, t_0 and t_f are the start and end time, respectively.

The cleaning of the original dataset includes three steps: Missing value test, type test, and outlier test. First, the missing value test is used to remove missing data and then unify data types. Then, the Z-score and inter-quartile range (IQR) methods are used to screen out suspected outliers. Finally, 6.9% of the screened samples are manually reviewed to determine whether to delete or retain them.

2.3 Principle of multiple linear regression model

The multivariate linear regression model is a classical mathematical, statistical model that explains the changes of dependent variables by constructing a linear equation with multiple independent variables. Its principle is shown in

$$\boldsymbol{y}=b_0+b_1\boldsymbol{x}_1+b_2\boldsymbol{x}_2+\cdots+b_n\boldsymbol{x}_n+\mu \tag{2}$$

where $\boldsymbol{x}_i(i=1,2,\cdots,n)$ and $\boldsymbol{y}$ are independent and dependent variables, respectively; b_0 is the constant, $b_i(i=1,2,\cdots,n)$ the partial regression coefficient, and μ the random error. This paper uses the ordinary least squares (OLS) method to calculate the partial regression coefficient in Eq.(2).

2.4 Principle of random forest regression model

The random forest model was proposed by Breiman in 2001. It originates from the idea of the Bagging algorithm and is a parallel integrated learning algorithm[16]. The model comprises multiple CART decision tree models and has good generalization ability and prediction performance[17]. The main steps of aircraft noise prediction based on the random forest model are as follows:

(1) Assuming that training set T contains W samples, the Bootstrap method is used to randomly select n groups of training subsets $\{t_1, t_2, \cdots, t_n\}$, t_i $(i=1, 2, \cdots, n)$ and T have the same capacity. The probability that each sample will not be selected is $(1-1/W)^W$, and the unselected data is called out of bag data.

(2) The feature dimension of the original training set and each training subset is H, and h features are randomly selected from H feature dimensions. According to the branch optimality criterion, the optimal feature is selected from h features for node splitting to build a decision tree.

(3) Each decision tree grows as far as possible without pruning according to the preset parameters, and n decision trees constitute a random forest regression model.

(4) The mean value of n decision trees' output values is taken as the final prediction result, and the prediction performance of the model can be evaluated with out-of-bag data.

The principle of the random forest regression model is shown in Fig.3.

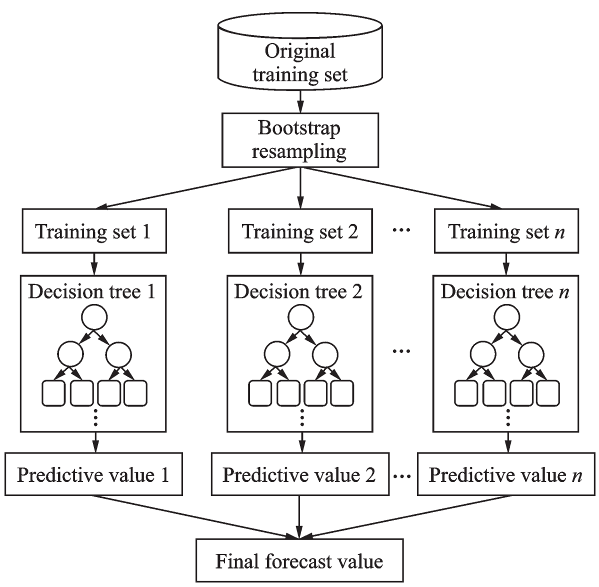

Fig.3 Principle of random forest regression model

2.5 Evaluation index

In order to evaluate the effect of the above two machine learning models on aircraft noise prediction quantitatively, the root mean square error (RMSE) and coefficient of determination R^2 are used as the basis.

Suppose y_i is the actual value, $\hat{y}_i$ the predicted value, $\bar{y}$ the average of the actual values, and n the total number. RMSE reflects the deviation degree between predicted and actual values[18]. The smaller the RMSE value, the better R^2 reflects the fitting degree of the predicted value to the actual value[19]. The closer the R^2 value is to 1, the better the fitting degree is. Their calculation formula is shown as

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2} \tag{3}$$

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}\right)^2} \tag{4}$$

3 Modeling Process and Discussion of Results

3.1 Modeling process of multiple linear regression

First, the Pearson correlation coefficient between each feature and the label is checked. The seaborn library's heatmap (·) function is used to make a correlation coefficient diagram. It can be seen from Fig.4 that the absolute values of the correlation coefficients of all features and noise labels are on the left and right of 0.3, showing a weak correlation.

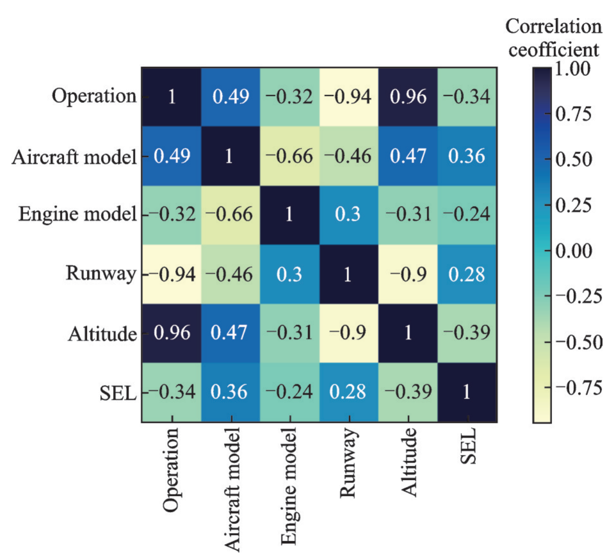

Fig.4 Pearson correlation coefficient diagram

Then, the variance inflation factor (VIF) is calculated to test the multicollinearity of the independent variable. If the VIF value is greater than or equal to 10, it indicates an obvious multicollinearity, and the independent variable needs to be eliminated. The calculation result of calling the variance_inflation_factor(·) function in the statsmodes library is shown in Table 1. The VIF value of the "Operation" is 16.529 970, which is greater than 10. After deleting the "Operation" feature column, the VIF value of the remaining features is less than 10.

Table 1 VIF factor value

Feature	Previous VIF factor	Rear VIF factor
Operation	16.530 034	—
Aircraft model	2.017 232	1.976 448
Engine model	1.772 234	1.772 228
Runway	8.141 861	4.077 224
Altitude	7.993 787	4.104 425

In order to avoid the impact of the numerical value of the category labels of "Operation" "Aircraft model" "Engine model" and "Runway" on the model training, the four columns of features are one-hot coded. The MinMaxScaler(·) function to is used to normalize the "Altitude" feature column, and then the formula.ols(·).fit(·) function in the API module of the statsmodes library is called to complete the modeling.

After the modeling is completed, the residual distribution histogram of the model is drawn. It can be seen from Fig.5 that the residual kernel density curve is close to the normal distribution curve. That is, the residual follows the normal distribution, indicating that the t value of the partial regression coefficient is valid, the partial regression coefficient is meaningful, the F value of the model is valid, and the model fitting equation is meaningful.

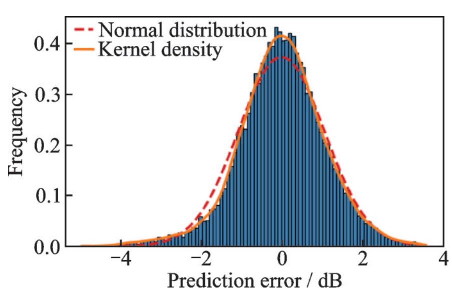

Fig.5 Residual distribution histogram

3.2 Modeling process of random forest regression

Before modeling, the one-hot encoding and normalization are also carried out, and then the GridSearchCV(·) function of the Sklearn library model_selection module is imported to search for the optimal parameters of the random forest model. That is, a variable containing the decision tree quantity list (n_estimators), the maximum tree depth list of each decision tree (max_depth), and the Scale list of the maximum number of features taken (max_features) is created. The list of minimum samples required by nodes (min_samples_split) are divided, and the optimal parameters through traversal operation are obtained, as shown in Table 2.

Table 2 RF model parameters

Keyword	Parameter list	Optimal value
n_estimators	[240, 250, 260, 270]	250
max_depth	[5, 6, 7, 8]	7
max_features	[0.4, 0.6, 0.8]	0.8
min_samples_split	[100, 200, 300, 400]	100

When setting hyperparameters, if the number of decision trees is too small, it will lead to insufficient training and a large deviation of prediction results. If it is too large, it will increase unnecessary calculations. If the number of features used is too large, the model will be overfitted, reducing the generalization ability of the model. If the number is too small, the model will be underfitted, reducing the prediction accuracy.

In a decision tree, the formula for calculating the importance of variable $\boldsymbol{X}_j$ at node m is shown in

$$V_{jm}^{\mathrm{Gini}}=\sum_{k=1}^{K}P_{mk}(1-P_{mk})-G_{ml}-G_{mr} \tag{5}$$

where $\sum_{k=1}^{K}P_{mk}(1-P_{mk})$ is the Gini index calculation formula of node m, K the number of classes in the decision tree, and P_{mk} the estimated probability that the sample belongs to class k at node m; G_{ml} and G_{mr} represent the Gini index of left and right nodes split by node m, respectively. The importance coefficients of variable $\boldsymbol{X}_j$ at each node are summed to get the importance coefficients of variable $\boldsymbol{X}_j$ in this decision tree, as shown in Eq.(6); the average value of the importance coefficient of the variable $\boldsymbol{X}_j$ in all decision trees is taken to obtain the importance coefficient of the variable $\boldsymbol{X}_j$ in this random forest model, as shown in Eq.(7).

$$V_{ij}^{\text{Gini}} = \sum_{m=1}^{M} V_{jm}^{\text{Gini}} \tag{6}$$

$$V_{j}^{\text{Gini}} = \frac{1}{n}\sum_{i=1}^{n} V_{ij}^{\text{Gini}} \tag{7}$$

where M is the number of nodes containing variable $\boldsymbol{X}_j$ in the ith tree, and n the number of decision trees in the random forest model.

The feature_importances_ function calculates the importance coefficients of five features, as shown in Fig.6. It can be seen that the importance coefficients of "Altitude" and "Aircraft model" features in the random forest model are higher, reaching 0.487 333 and 0.351 191, respectively, while the importance coefficients of "Operation" features are only 0.008 547.

The tree module of the sklearn library and the pydotplus library is imported, and the E function is called to visualize the specified decision tree in the random forest.

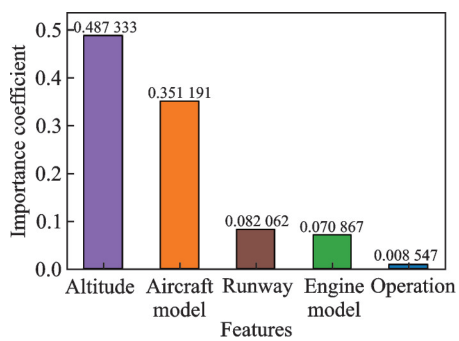

Fig.6 Importance coefficient diagram

3.3 Forecast results

The train_test_split (·) function of the model_selection module in the sklearn library is called to divide 10% of the data into a verification set, and the random seed "random_state" is not set. The two models are used to predict ten times, and the partial results of one experiment are shown in Fig.7.

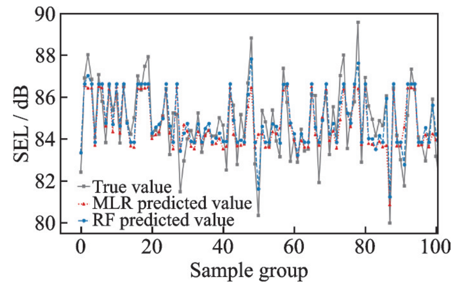

Fig.7 Partial forecast results

In Fig.7, the gray solid line is the actual value in the validation set, and the red dotted line and blue dashed line are the prediction values of the multivariate linear regression model and the RF model, respectively. It can be seen that the MLR model performs poorly in the peak part of the prediction, while the RF model performs relatively well.

The RMSE and R^2 values of the predicted data from ten experiments are calculated and a line chart and box chart are drawn, as shown in Fig.8.

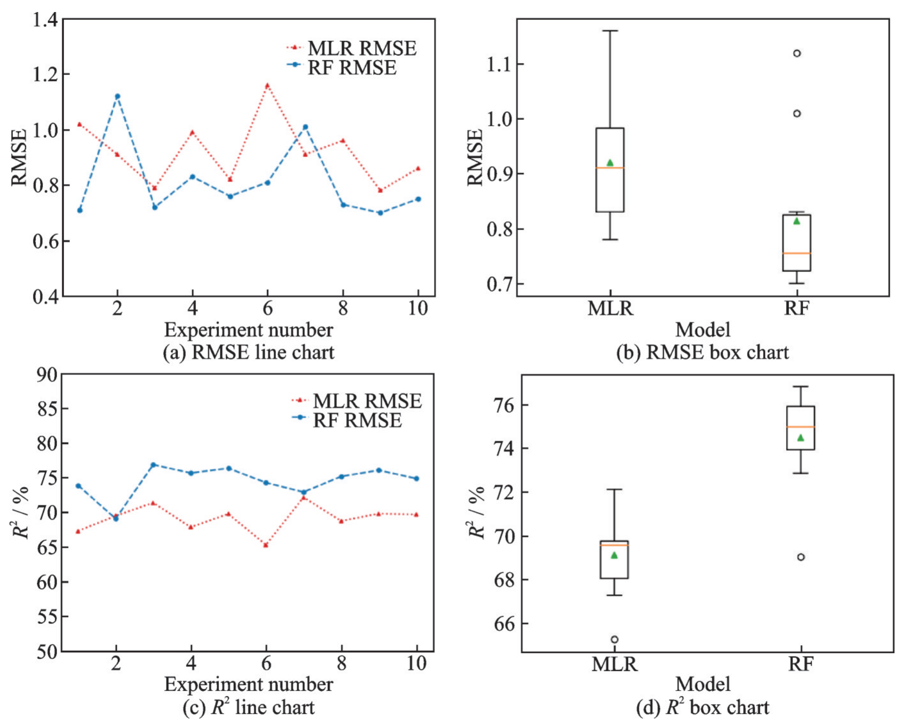

Fig.8 RMSE and R^2 value diagrams

In Figs.8(a) and (c), the red dotted line is the RMSE and R^2 value of the MLR model, and the blue dashed line is the RMSE and R^2 value of the RF model. In Figs. 8(b) and (d), the line in the middle of the box is the median, the triangle is the average value, and the dot is the abnormal value. The R^2 value of the MLR model fluctuates between 65.26% and 72.12%, with an average of 69.108%, and the RMSE value fluctuates between 0.78 and 1.16, with an average of 0.92. The R^2 value of the RF model fluctuates between 69.03% and 76.82%, with an average of 74.469%, and the RMSE value fluctuates between 0.7 and 1.12, with an average of 0.814.

4 Conclusions

According to the above experimental results, the following conclusions can be drawn.

(1) Not only the best practice model and scientific model but also the machine learning model can predict aviation noise and show good results. Compared with the scientific model, the data required by the machine learning model prediction is easier to be obtained. Compared with the best practice model, the machine learning model prediction process is more straightforward, and the prediction results do not need to be supplemented by various amendments.

(2) In the data set used in this experiment, "Operation" "Aircraft model" "Engine model" "Runway" and "Altitude" are the input features, and the SEL noise value is the label. MLR model is more sensitive to outliers in samples, while the integrated algorithm RF model has better anti-noise ability and generalization ability. On the above datasets, the prediction performance of the RF model is better than that of the MLR model. Specifically, the average R^2 value in the RF model is 5.361% higher than that in the MLR model, and the average RMSE value is 0.106 lower than that in the MLR model.

(3) This experiment uses the data from SEA airport in the summer, but it does not include meteorological data. That is to say, the impact of temperature difference between day and night on noise is not considered. The accuracy of model prediction can be further improved by integrating more meteorological data.

References

[1] GUO Y, THOMAS R H, CLARK I A, et al. Far-term noise reduction roadmap for the midfuselage nacelle subsonic transport[J]. Journal of Aircraft, 2019, 56(5): 1893-1906.

[2] World Health Organization. Environmental noise guidelines for the European region[M]. Denmark: World Health Organization Regional Office for Europe, 2018.

[3] PERIS E. Environmental noise in Europe: 2020[J]. Eur Environ, Agency, 2020, 1: 104.

[4] SAUCY A, SCHÄFFER B, TANGERMANN L, et al. Does night-time aircraft noise trigger mortality? A case-crossover study on 24 886 cardiovascular deaths[J]. European Heart Journal, 2021, 42(8): 835-843.

[5] MÜNZEL T, SCHMIDT F P, STEVEN S, et al. Environmental noise and the cardiovascular system[J]. Journal of the American College of Cardiology, 2018, 71(6): 688-697.

[6] KLATTE M, SPILSKI J, MAYERL J, et al. Effects of aircraft noise on reading and quality of life in primary school children in Germany: Results from the NORAH study[J]. Environment and Behavior, 2017, 49(4): 390-424.

[7] CLARK C, CRUMPLER C, NOTLEY H. Evidence for environmental noise effects on health for the United Kingdom policy context: A systematic review of the effects of environmental noise on mental health, wellbeing, quality of life, cancer, dementia, birth, reproductive outcomes, and cognition[J]. International Journal of Environmental Research and Public Health, 2020, 17(2): 393.

[8] LU C. Is there a limit to growth? Comparing the environmental cost of an airport's operations with its economic benefit[J]. Economies, 2017, 5(4): 44.

[9] WOLFE P J, YIM S H L, LEE G, et al. Near-airport distribution of the environmental costs of aviation[J]. Transport Policy, 2014, 34: 102-108.

[10] SPARROW V, GJESTLAND T, GUSKI R, et al. State of the science 2019: Aviation noise impacts[R]. [S.l.]: ICAO, 2019.

[11] KOPSCH F. The cost of aircraft noise—Does it differ

from road noise? A meta-analysis[J]. Journal of Air Transport Management, 2016, 57: 138-142.
[12] ISERMANN U, BERTSCH L. Aircraft noise immission modeling[J]. CEAS Aeronautical Journal, 2019, 10: 287-311.
[13] ZELLMANN C, SCHÄFFER B, WUNDERLI J M, et al. Aircraft noise emission model accounting for aircraft flight parameters[J]. Journal of Aircraft, 2018, 55(2): 682-695.
[14] VELA A E, OLEYAEI-MOTLAGH Y. Ground level aviation noise prediction: A sequence to sequence modeling approach using LSTM recurrent neural networks[C]//Proceedings of 2020 AIAA/IEEE 39th Digital Avionics Systems Conference (DASC). [S.l.]: IEEE, 2020: 1-8.
[15] REVOREDO T, MORA-CAMINO F, SLAMA J. A two-step approach for the prediction of dynamic aircraft noise impact[J]. Aerospace Science and Technology, 2016, 59: 122-131.
[16] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45: 5-32.
[17] CAO Z, FANG Z, YAO J, et al. Loess landform classification based on random forest[J]. Journal of Geo-information Science, 2020, 22(3): 452-463.
[18] RUDOY G I. On modification of the mean squared error loss function for solving nonlinear heteroscedastic errors-in-variables problems[J]. Informatics and Its Applications, 2017, 11(2): 74-84.
[19] ZENG Q, LIANG Y, CHEN G, et al. Noise prediction of chemical industry park based on multi-station Prophet and multivariate LSTM fitting model[J]. EURASIP Journal on Advances in Signal Processing, 2021, 2021: 1-18.

Acknowledgements This work was supported by the National Natural Science Foundation of China (No. 52202442) and National Key R&D Program of China (No.2022YFB2602403).

Authors Mr. **FENG Hao** received his B.S. degree in transportation from Nanjing Agricultural University in 2019. He is studying for a M.S. degree in transportation at Nanjing University of Aeronautics and Astronautics, and his research interests are aircraft noise assessment.

Dr. **ZHOU Yadong** received the B.S. degree in management from Southwest University in 2011 and the Ph.D. degree in engineering mechanics from Southeast University in 2018. He is currently a master's supervisor of the College of Civil Aviation at Nanjing University of Aeronautics and Astronautics. His research interests are random-vibration fatigue, bird-strike impact, buckling analysis, thin-walled structures, etc.

Author contributions Mr. **FENG Hao** constructed the model, analyzed the results, and wrote a manuscript. Dr. **ZENG Weili** and Dr. **ZHOU Yadong** participated in model building and polished the manuscript. Mr. **DING Cong** and Mr. **GUO Wentao** collected data and investigated the background of the problem. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: LIU Yandong)

基于机器学习模型的飞机噪声预测

丰　豪,周亚东,丁　聪,曾维理,郭文韬

(南京航空航天大学民航学院,南京211106,中国)

摘要: 为了探索最佳实践模型和科学模型之外的飞机噪声预测方法,采用多元线性回归模型和随机森林回归模型对西雅图-塔科马国际机场2020—2022年夏季的飞机噪声值进行了预测。实验验证了机器学习模型在飞机噪声预测任务中的可行性和优势,结果表明随机森林回归模型预测结果的R^2均值为74.469%,比多元线性回归模型预测结果的R^2均值高出5.361%;随机森林回归模型预测结果的RMSE均值为0.814,比多元线性回归模型预测结果的RMSE均值低0.106。

关键词: 飞机噪声排放;飞机噪声预测;多元线性回归;随机森林回归