scientific reports

Check for updates

OPEN

Random forest method for estimation of brake specific fuel consumption

Qinsheng Yun[1,2]✉, Xiangjun Wang[1]✉, Chen Yao[2] & Haiyan Wang[3]

The internal combustion engine is a widely used power equipment in various fields, and its energy utilization is measured using brake specific fuel consumption (BSFC). BSFC map plays a crucial role in the analysis, optimization, and assessment of internal combustion engines. However, due to cost constraints, some values on the BSFC map are estimated using techniques like K-nearest neighbor, inverse distance weighted interpolation, and multi-layer perceptron, which are recognized for their limited accuracy, particularly when dealing with distributed sampled data. To address this, an improved random forest method is proposed for the estimation of BSFC. Polynomial features are employed to increase higher dimensions of features for random forest by nonlinear transformation, and critical parameters are optimized by particle swarm optimization algorithms. The performance of different methods was compared on two datasets to estimate 20%, 30%, and 40% of BSFC data, and the results reveal that the method proposed in this paper outperforms other common methods and is suitable for estimating the BSFC map.

List of symbols

E	(Effective) work potential
E_0	Exergy
E_{00}	Energy of a system
K	Kelvin temperature scale
S	Entropy
T	Temperature or Celsius temperature scale
W	Effective work

The internal combustion engine finds extensive application in automobiles, ships, agriculture, modern industry, and construction machinery. It operates by converting gas expansion into mechanical energy and is considered the most promising product for energy conservation and emission reduction. To measure its energy efficiency, the brake specific fuel consumption (BSFC) is used. This refers to the fuel consumption of the engine per kilowatt-hour of work and is crucial for improving the engine's economy and thermal efficiency as a heat engine[1,2]. The BSFC map is generated by plotting the fuel consumption against the engine speed and load on the X and Y axes, respectively, over the engine's operating range. The map serves as an important tool for evaluating engine performance and enhancing its design and efficiency[3,4].

The BSFC map is a widely used tool in the analysis, optimization, and control of internal combustion engines. It serves multiple purposes, the first of which is to analyze engine performance and predict fuel consumption. This is exemplified by the analysis of a two-circuit bottom cycle system for a diesel engine[5], where fine-grain fuel consumption is predicted using the BSFC map[6]. The BSFC map can also be used to optimize fuel consumption and reduce engine emissions. For instance, in the automotive industry, the BSFC map is utilized to control diesel engines for minimum fuel consumption[7] and to obtain the optimal operating mode for the highest economic standard[8]. In addition, the BSFC map is helpful in studying the overall arrangement and system design of internal combustion engines, such as in the modeling and scheduling of fuel-efficient ships[9].

The BSFC map is an essential tool for internal combustion engine research, and its accurate representation is crucial for further development in this field. Accurate mapping of the BSFC map requires precise measurement of the BSFC, but in practice, some values can only be estimated due to cost and other constraints. Common calculation methods include the K-nearest neighbor (KNN) method, polynomial regression, inverse distance

[1]Naval University of Engineering, Wuhan 430000, China. [2]Shanghai Marine Diesel Engine Research Institute, Shanghai 200000, China. [3]Shanghai Maritime University, Shanghai 200000, China. ✉email: yunqinsheng@126.com; wxjnue@163.com

weighted (IDW) method, ordinary kriging (OK) method, and multi-layer perceptron (MLP) method. However, these methods are known to have large errors in estimating uniformly distributed data[10]. This is particularly problematic when drawing high-resolution prediction maps for weather data[11], where accuracy is essential. Research has shown that these common methods are insufficient for data estimation, especially when the sampled data is intimidatingly distributed, as reported in many fields such as agriculture and mining[12–14]. Recently, machine learning-based methods have become increasingly popular in various fields, including medical imaging and energy[15]. The random forest (RF) method, as an ensemble learning method, uses a decision tree classifier to achieve integrated decision-making[16]. Compared to other machine learning methods, this method has low computation requirements and high precision and is not sensitive to multicollinearity. It also demonstrates good robustness to missing and unbalanced data[10–13]. In this regard, an improved RF method has been introduced in this study to enhance the accuracy of BSFC estimation, marking the first application of this method in estimating the BSFC map. The results show that it outperforms other common methods on two different datasets. Therefore, it is a suitable method for estimating the BSFC map.

Methods for the estimation of BSFC

The aim of calculating BSFC is to predict the fuel consumption rate of an internal combustion engine under unknown operating conditions by learning the relationship between the fuel consumption rate and the known operating conditions. The relationship between the operating conditions and fuel consumption rate is usually determined through experiments. However, experimental limitations such as cost and conditions result in a limited amount of data. Accurately estimating fuel consumption under different operating conditions is crucial for fuel consumption control, which is a critical task in practical internal combustion engine work. Therefore, predicting fuel consumption under varying operating conditions is a fundamental task for optimizing internal combustion engines.

Consider the operating state of an internal combustion engine is represented by the combined measured state variables such as speed and power, denoted as x. The corresponding fuel consumption rate of the engine, denoted as y, and $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), ..., (\boldsymbol{x}_N, y_N)\}$. The task of estimating BSFC involves determining the fuel consumption rate, denoted as y, that corresponds to the unmeasured state variables, denoted as x, based on the dataset D. This problem involves establishing a mapping between the input variable and the output variable, and then using this mapping to predict the output value y for a given input value x. This problem can be classified as either a regression problem or an interpolation problem. There are several methods commonly used to solve this problem, including the KNN method, IDW method, OK method, and MLP method. The introductions are provided for each of these methods below.

KNN method

The KNN method is a conventional and efficient machine learning technique that operates on a simple concept. It calculates the average values of the points located in close proximity to the estimated points in the known dataset. Due to its speed and simplicity, the KNN method has found its application in various interpolation scenarios, such as cloud edge computing[17].

IDW method

The IDW method is a conventional and efficient technique for interpolation. Its fundamental concept involves assigning higher weights to the points in the training set closer to the interpolation points. Let the coordinates of n known points be (X_i, Y_i, Z_i), and i = 1, 2, 3, …, n, then the z value at the point (x, y, z) is given as

$$z(x,y) = \begin{cases} Z_i & x = X_i, y = Y_i \\ \frac{\sum_{i=1}^{n} Z_i d_i^{-2}}{\sum_{i=1}^{n} d_i^{-2}} & \text{otherwise} \end{cases} \quad (1)$$

where d_i^{-2} is the inverse of the Euclidean distance from (x, y) to (X_i, Y_i) squared. The weight in this method follows a normalization condition, and it is evident that the closer a point is to the interpolation point, the higher the weight assigned to it.

OK method

The OK method is based on the assumption that the data space has uniform expectations and variance. It uses optimal estimation to obtain the data for unknown points. This geostatistical technique is widely applied in fields such as geographical sciences, environmental sciences, and atmospheric sciences. The OK method has been utilized for deposit Cu concentration[14] and has been reported to provide high-fidelity uncertainty quantification in composite shell dynamics[18].

MLP method

The MLP method employs cascaded neurons that use a sigmoid nonlinear function to map the input to output, enabling the approximation of any nonlinear function. Thus, the neural network can approximate any given multivariable continuous function, including drawing characteristic curves for power machines. This method is highly flexible and possesses a strong nonlinear mapping ability, making it a broadly applicable computational technique. It has found use in numerous applications, such as predicting macroclimate index runoff in atmospheric science[19] and assessing the sensitivity to flood temperature in geographical research[20].

Improved RF method for the estimation of BSFC

KNN method

RF is a regression method based on trees and has the benefits of strong prediction ability, low overfitting risk, and high interpretability[8,9]. This method is computationally efficient and exhibits superior speed and accuracy[14,15]. It has been widely applied in various fields, including environmental science, agriculture, and engineering. For instance, it has been utilized to classify medical images[21] and predict indoor radon concentration[22].

RF is one of the widely used ensemble learning methods. It employs a large number of regression trees for ensemble learning, with random attribute selection during the training process. The regression tree serves as the fundamental learner for RF regression. As with other machine learning techniques, in RF, features and labels, are referred to as X and Y, respectively, while N represents the sample number and D represents the training data set. The representation is as follows: $X = \{x_1, x_2, \ldots, x_N\}, Y = \{y_1, y_2, \ldots, y_N\}, D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$. A regression tree corresponds to a partition of the feature space and labels on the partitioned units. Dividing the feature space into M units $R_1, R_2, \ldots, R_M$, each unit R_M with a fixed label C_m, the regression tree model can be represented as

$$f(\boldsymbol{x}) = \sum_{m=1}^{M} c_m I(\boldsymbol{x} \in R_m) \tag{2}$$

$$I(\boldsymbol{x} \in R_m) = \begin{cases} 0(\boldsymbol{x} \notin R_m) \\ 1(\boldsymbol{x} \in R_m) \end{cases} \tag{3}$$

The square error (E) is used to express the prediction error of the regression tree for the training data in the feature space whose partitioning method has been determined.

$$E = \sum_{\boldsymbol{x}_i \in R_m} (y_i - f(\boldsymbol{x}_i))^2 \tag{4}$$

This error is used to determine the optimal output value on each unit. In the RF method, the following algorithm is used to generate a regression tree.

Step 1: Select the j-th variable and its value s as the segmentation variable and segmentation point, respectively. The two regions are defined as follows:

$$\begin{aligned} R_1(j, s) &= \left\{ \boldsymbol{x} | \boldsymbol{x}^{(j)} \leq s \right\} \\ R_2(j, s) &= \left\{ \boldsymbol{x} | \boldsymbol{x}^{(j)} > s \right\} \end{aligned} \tag{5}$$

Step 2: Solve the following problem to obtain the optimal j and s values. These values divide the input space into two regions, R_1 and R_2.

$$\min_{j,s} \left[\min_{c_1} \sum_{\boldsymbol{x}_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{\boldsymbol{x}_i \in R_2(j,s)} (y_i - c_2)^2 \right] \tag{6}$$

It is easy to understand that the optimal value $\hat{c}_m$ of c_m on a unit R_m is the mean of the outputs y_i corresponding to all input instances x_i in the unit, which can be expressed as

$$\widehat{c}_m = \frac{1}{K} \sum_{k=1}^{K} y_k (\boldsymbol{x}_k \in R_m) \tag{7}$$

Step 3: Repeat steps 1 and 2 for R1 and R2, respectively, until the termination condition is reached. The termination condition can be that each interval contains one sample, all samples have been used, or the number of units has reached a specified number.

The RF method involves creating a training subset by randomly sampling D and evaluating the error of the remaining samples. Multiple random trees are then generated using the same method for generating random trees, except that instead of using all features, a specified number of features are randomly selected. A total of NT regression trees were generated, denoted as $\{f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_{N_T}(\boldsymbol{x})\}$. If the weight is set to $W = \{w_1, w_2, \ldots, w_{N_T}\}$ and $w_1 = w_2 = \cdots = w_{N_T} = 1/N\mathrm{T}$, the regression prediction result for feature $\boldsymbol{x}$ is

$$f(\boldsymbol{x}) = \sum_{i=1}^{N_T} w_i f_i(\boldsymbol{x}) \tag{8}$$

It is evident that the diversity in RF integration arises not only from sample disturbances but also from attribute disturbances. This results in a greater variation between individuals, leading to strong adaptability and anti-interference ability toward the data.

Improved RF method

In the RF algorithm, decision trees are generated directly from the features. In machine learning, adding some nonlinear features of input data can be an effective way to increase the complexity of the model. Therefore, this study introduces polynomial features that can generate higher dimensions of features and terms related to each other. Polynomial features are a method of increasing dimensionality and performing nonlinear transformations in machine learning. It combines and expands the original features, which improves the model's expression ability and fitting effect.

Let the feature vector be $\boldsymbol{x} = [x_1, x_2, \ldots, x_m]$, and define the feature of a polynomial of degree 0 as $\phi_0(\boldsymbol{x}) = 1$. The d-th polynomial feature can be represented by the following iterative formula.

$$\begin{aligned}\phi_d(\boldsymbol{x}) &= [\ \phi_{d-1}(\boldsymbol{x})\ \ x_1^n\ \ x_1^{n-1}x_2\ \ \cdots\ \ x_2^n\ \cdots\ \ x_{m-1}x_m^{n-1}\ \ x_m^n\ \ ] \\ &= [\ \phi_{d-1}(\boldsymbol{x})\ \ \phi'_d(\boldsymbol{x})\ ]\end{aligned} \tag{9}$$

$\phi'_d(\boldsymbol{x})$ is the row vector, that contains one or more variables from all possible $x_1, x_2, \ldots, x_m$ variables, with a degree of d as a monomial expression.

When the RF method is used to estimate BSFC, the d-degree polynomial feature $\phi_d(\boldsymbol{x})$ of feature $\boldsymbol{x}$ serves as the input feature of the RF regression model. This can incorporate more combinations of original features into the consideration of generating decision trees, enhancing their fitting and expression abilities.

The polynomial feature $\phi_d(\boldsymbol{x}_i)$ of each feature $\boldsymbol{x}_i$ is used to form a new training set $\Phi_d(D) = \{(\phi_d(\boldsymbol{x}_1), y_1), (\phi_d(\boldsymbol{x}_2), y_2), \ldots, (\phi_d(\boldsymbol{x}_N), y_N)\}$. The RF model F is trained with $\Phi_d(D)$, and the features with a proportion of p in all features are employed when the nodes split.

For a given feature vector $\boldsymbol{x}$, and its polynomial feature, denoted by $\phi_d(\boldsymbol{x})$, the predicted result value $\hat{y} = F(\phi_d(\boldsymbol{x}))$ is obtained using model F. The map from feature vector $\boldsymbol{x}$ to $\hat{y}$ is called a polynomial feature RF model $f_{(d,p)}(\boldsymbol{x})$ with hyperparameters (d, p).

Parameter optimization based on particle swarm algorithm

When polynomial features are introduced, the feature dimension for the feature vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_m]$ and polynomial feature $\phi_d(\boldsymbol{x})$ increases from m to $C_{m+d}^d = (m+d)!/(d!m!)$. However, too many polynomial features can cause slow training due to a large number of feature dimensions and may lead to overfitting, while too few features can result in underfitting. Thus, the degree d of the polynomial feature needs to be selected carefully.

Similarly, in decision tree generation, the parameter p represents the proportion of features considered to the total number of features. Too many features can lead to model complexity, which can be affected by noise and randomness, while too few features may cause under-fitting, making it difficult to capture complex relationships in the data. Therefore, when polynomial features are introduced, p needs to be selected more carefully.

Since both p and d are critical parameters, particle swarm optimization algorithms can be considered to optimize their combination. The object function is as follows.

$$L(d,p) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_{(d,p)}(\boldsymbol{x}_i))^2 \tag{10}$$

The optimization process begins with initialization, where the total number of particles and the number of iterations are specified. Each particle is randomly assigned a position $\boldsymbol{pi} = \{pi, di\}$ and a velocity $\boldsymbol{vi} = \{v_{pi}, v_{di}\}$. The objective function of each particle is then calculated to obtain the individual optimal solution of that particle, and the position of the particle with the smallest objective function is considered the global optimal solution.

In each iteration, the following calculations are performed.

For the i-th particle, the objective function of its particle is calculated. If the objective function result is less than the objective function at the position $\boldsymbol{g}_i^{best} = \{g_{pi}^{best}, g_{di}^{best}\}$ of the individual optimal solution, update the individual optimal solution to the current position. If the objective function result is less than the objective function at the global optimal solution position $\boldsymbol{g}^{best} = \{g_p^{best}, g_d^{best}\}$, update the global optimal solution to the current position. The velocity and position of the particles are updated as

$$v_{pi} \leftarrow \omega v_{pi} + c_1 r_1 (p_{pi}^{best} - p_i) + c_2 r_2 (g_p^{best} - p_i) \tag{11}$$

$$v_{di} \leftarrow \omega v_{di} + c_1 r_1 (p_{di}^{best} - d_i) + c_2 r_2 (g_d^{best} - d_i) \tag{12}$$

$$p_i \leftarrow p_i + v_{pi} \tag{13}$$

$$d_i \leftarrow d_i + v_{di} \tag{14}$$

In the above equation, ω is the inertia weight, generally set to 0.9. c_1 and c_2 are the acceleration coefficients, generally set to 2.0. r_1 and r_2 are randomly selected from [0, 1] at each update.

When the maximum number of iterations is reached, $\boldsymbol{g}^{best}$ is the optimal parameter of p and d.

Experiments and results

Experimental data

The data sets used in this paper were obtained from references[23,24], The data sets actual measurements of two gasoline internal combustion engines, including speed, power and fuel consumption rate. The two engines

produced a total of 52 and 80 measured data points, respectively. Tables 1 and 2 show the Speed, power and fuel consumption rate of the engines.

Figures 1 and 2 show the distribution of the first fuel engine in the speed-power plane and the distribution of the speed-power-fuel consumption in the three-dimensional space.

Figures 3 and 4 show the distribution of the second fuel engine in the speed-power plane and the distribution of the speed-power-fuel consumption in the three-dimensional space.

Evaluation index

In this paper, the following indicators are used for evaluation: root mean square error (RMSE), normalized mean square error (NMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and R-squared (R^2). Each indicator is calculated as follows.

The RMSE is defined as:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{15}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 \tag{16}$$

where n is the total number of the data to be estimated, *yi* is the real value to be estimated, and *yl* is the estimated value.

To compare the accuracy and degree of variation of different datasets, the NMSE) is proposed to compare the methods on different datasets. The calculation of NMSE is as follows.

$$\mathrm{NMSE} = \frac{\mathrm{MSE}}{\frac{1}{n}\sum_{i=1}^{n} y_i^2} \tag{17}$$

Rpm (r/min)					
P(kW)	be (g/kW h)	P(kW)	be (g/kW h)	P(kW)	be (g/kW h)
1200		1400		1600	
11.0	214.0	12.0	219.0	13.0	227.0
15.0	191.0	15.0	201.5	15.0	214.0
20.0	177.0	20.0	183.0	20.0	192.0
25.0	168.5	25.0	172.0	25.0	178.0
30.0	167.5	30.0	165.5	30.0	169.5
36.0	171.0	35.0	162.5	35.0	165.5
		41.0	163.5	40.0	165.0
				47.0	171.0
1800		2000		2200	
15.0	223.5	20.0	207.0	18.0	248.5
20.0	195.5	30.0	182.5	25.0	214.5
30.0	174.5	35.0	175.0	35.0	189.5
35.0	170.5	40.0	171.5	45.0	176.5
40.0	169.0	45.0	172.5	50.0	173.0
45.0	169.0	50.0	176.5	55.0	172.5
50.0	171.0	58.0	187.0	60.0	174.5
53.0	174.0			65.0	180.0
2500		/		/	
20.0	254.5				
30.0	211.0				
40.0	188.0				
50.0	177.0				
55.0	175.0				
60.0	175.5				
65.0	178.5				
70.0	185.0				
20.0	254.5				
30.0	211.0				

Table 1. Speed, power and fuel consumption rate of first engines[23].

Rpm (r/min)					
P(kW)	be (g/kW h)	P(kW)	be (g/kW h)	P(kW)	be (g/kW h)
1400		1600		1800	
58.61	222.8	68.54	222.0	76.96	226.0
51.51	220.4	61.27	221.7	69.42	225.3
46.69	232.4	55.00	235.4	61.88	226.4
40.77	228.5	47.60	226.5	54.47	233.9
34.63	227.8	40.83	230.5	46.06	242.1
29.85	232.6	34.04	236.8	39.35	283.3
27.16	248.5	27.53	249.1	31.61	253.9
23.05	245.9	20.76	276.1	24.90	271.4
17.18	272.4	13.99	407.9	16.87	323.5
11.85	329.7	6.65	487.0	8.69	468.6
2000		2200		2400	
89.13	206.5	96.92	234.7	101.68	174.2
79.64	231.1	87.45	259.8	90.60	242.2
69.68	231.1	77.08	235.5	81.10	252.1
60.92	233.0	67.17	237.6	71.12	287.4
51.18	242.0	56.30	242.8	61.14	253.6
42.95	244.9	46.72	292.3	51.64	263.6
33.55	265.0	36.28	277.9	40.74	290.6
23.98	299.8	26.72	308.7	31.34	316.8
2600		2800		/	
102.91	256.9	92.53	257.9		
93.85	253.7	80.77	295.3		
84.48	253.5	71.10	282.4		
71.96	260.0	61.66	288.7		
61.56	303.8	52.34	301.9		
50.86	280.7	42.69	329.7		
41.98	300.6	34.77	357.0		
31.39	346.6	21.29	475.4		
20.77	435.6	15.48	580.3		
9.28	812.9	6.57	1080.1		

Table 2. Speed, power and fuel consumption rate of second engines[24].

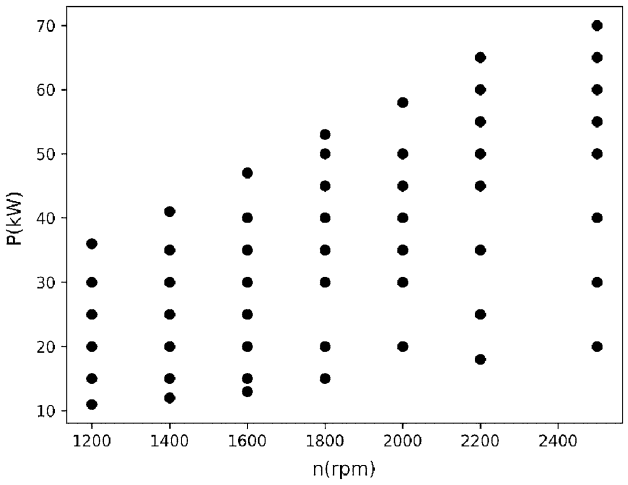

Figure 1. 2D distribution of all collected data for the first BSFC.

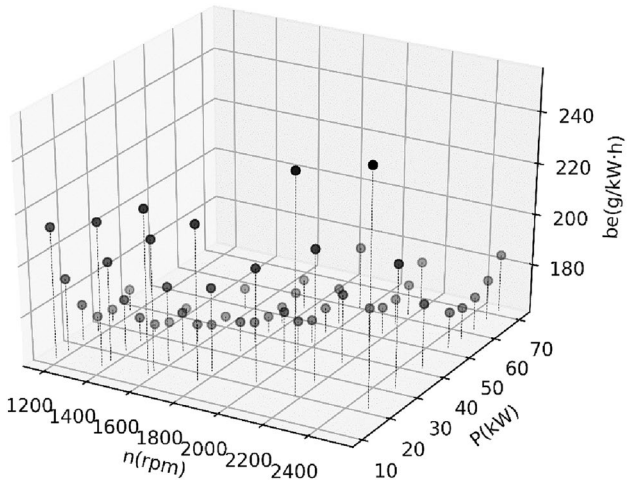

Figure 2. 3D view of all collected data for the first BSFC.

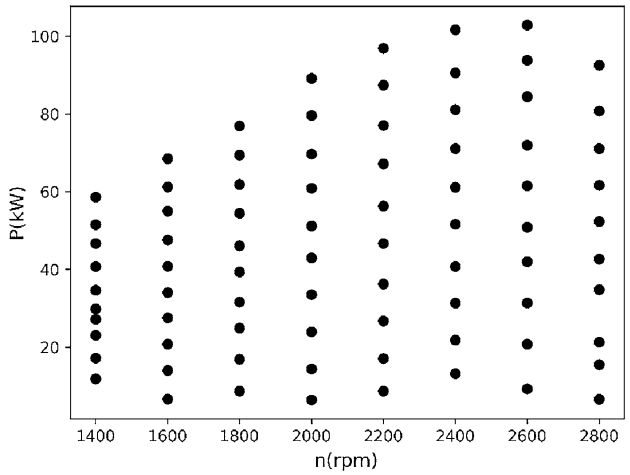

Figure 3. 2D distribution of all collected data for the second BSFC.

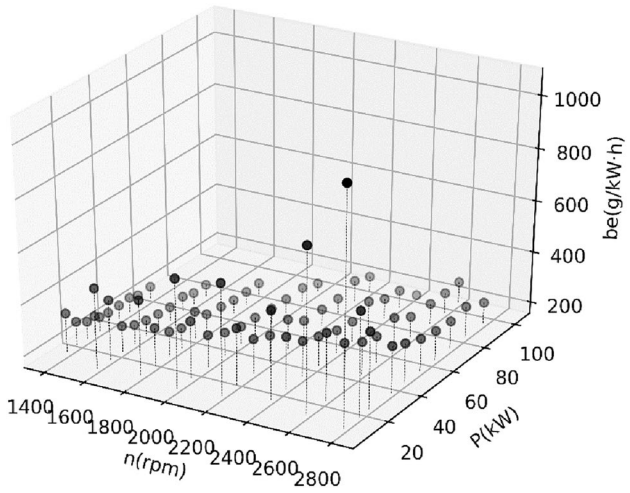

Figure 4. 3D view of all collected data for the second BSFC.

The MAE is also used to compare estimation errors. The calculation of MAE is

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right| \tag{18}$$

To evaluate and compare the accuracy of different algorithms and data sets, the MAPE is utilized in this study. The MAPE is considered more robust than the MAE, as it normalizes the error of each data point and can be used as an evaluation indicator. It is defined as

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{19}$$

R^2 is also used to evaluate different estimation methods, representing the proportion of estimated data information to original data information. The calculation of R^2 is as follows.

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \tag{20}$$

where $\bar{y}$ represents the average value of all the data to be estimated. The value range of R^2 is $(-\infty, 1]$. The closer R^2 is to 1, the more accurate the estimation method's results are. On the contrary, the farther R^2 is from 1, the greater the result error of the estimation method. When R^2 is less than 1, it indicates that the estimation error of the method is significant, even greater than using the mean as the estimation value.

In this paper, five indicators are used for evaluation, they are RMSE, NMSE, MAE, MAPE, and R^2. RMSE represents the standard deviation between the estimated value and the true value error, while NMSE represents the percentage of error. MAE represents the average error between the estimated value and the true value, while MAPE represents the percentage of this error. R^2 expresses the degree of fit between the data and the regression model. NMSE and MAPE can serve as the primary performance indicators, while other indicators can serve as secondary indicators.

Experimental results

To compare different estimation methods, the known data in this study were randomly divided into two groups at a ratio of 4:1, with 80% of the data being known and the remaining 20% being used for estimation. The data estimation methods compared in this study include KNN, IDW, OK, MLP, RF, and the proposed RF. The performance indicators compared in this study include RMSE, NMSE, MAE, MAPE, and R^2. To reduce the impact of grouping randomness on statistical results, the experiment was repeated 10 times, using the same ratio for random grouping each time. After each grouping, the known sample dataset and the estimated dataset used for testing have different data. The average of the performance metrics of the 10 experiments is used as the final indicator for performance comparison.

Estimating 20% of BSFC data

Tables 3 and 4 present the performance metrics of various estimation methods on Dataset 1 and Dataset 2 for estimating 20% of BSFC data, respectively. The reported values in these tables are the average results from ten experiments. Figures 5 and 6 display the estimated values of different methods for the BSFC of Datasets 1 and 2, respectively. These figures show the actual estimated result data and real data of a single experiment in the ten experiments.

The results of the experiment conducted on Dataset 1 indicate that the proposed RF method described in this paper outperforms RF method with an RMSE of 0.46 lower, and it outperforms other methods with an RMSE of 5.05 lower. And additionally, the other errors are similar, and the R2 value of RF is closest to 1. These indexes show that the proposed RF has a minimal error and the highest accuracy. Similar results were observed on Dataset 2, the proposed method outperforms other methods with an RMSE of 9.71 lower.

Method	RMSE	NMSE	MAE	MAPE	R2
KNN	15.28	0.0079	10.72	0.054	0.29
IDW	26.39	0.0228	10.01	0.105	−1.66
OK	19.38	0.0114	16.07	0.084	−0.22
MLP	13.69	0.0092	11.37	0.061	−0.24
RF	9.81	0.0032	7.00	0.036	0.68
Prposed RF	8.64	0.0025	6.35	0.032	0.75

Table 3. Performance comparison of different methods on Dataset 1 for estimating 20% of BSFC data.

Method	RMSE	NMSE	MAE	MAPE	R^2
KNN	84.19	0.0799	47.54	0.124	0.23
IDW	159.67	0.4993	134.59	0.459	− 5.16
OK	92.46	0.095	68.47	0.209	0.10
MLP	44.09	0.0248	36.00	0.119	0.76
RF	57.68	0.0388	35.12	0.097	0.64
Proposed RF	34.38	0.0143	23.75	0.073	0.87

Table 4. Performance comparison of different methods on Dataset 2 for estimating 20% of BSFC data.

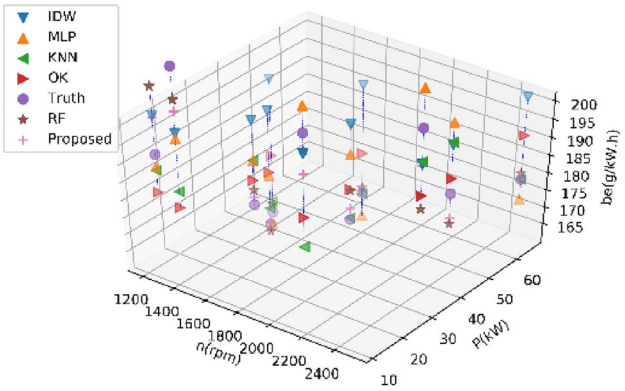

Figure 5. Results of different methods on Dataset 1 for estimating 20% of BSFC data.

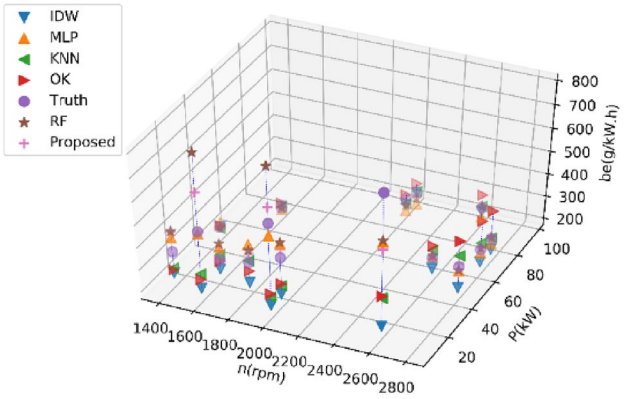

Figure 6. Results of different methods on Dataset 2 for estimating 20% of BSFC data.

Estimating 20% of BSFC data

Tables 5 and 6 present the average performance metrics of various estimation methods on Dataset 1 and Dataset 2 for estimating 30% of BSFC data after ten experiments, respectively. Figures 7 and 8 display the estimated values of different methods for the BSFC of Datasets 1 and 2 in a single experiment, respectively. The results of the experiment conducted on Dataset 1 indicate that the proposed RF method described in this paper outperforms other methods with an RMSE of 0.66 lower. The proposed method outperforms other methods with an RMSE of 23.84 lower on Dataset 2. All the indexes show that the proposed RF method has a minimal error and the highest accuracy.

Estimating 40% of BSFC data

Tables 7 and 8 present the average performance metrics of various estimation methods on Dataset 1 and Dataset 2 for estimating 40% of BSFC data after ten experiments, respectively. Figures 9 and 10 display the estimated values of different methods for the BSFC of Datasets 1 and 2 in a single experimen, respectively. The proposed RF method described in this paper outperforms other methods with an RMSE of 1.39 lower on Dataset 1. The

Method	RMSE	NMSE	MAE	MAPE	R^2
KNN	15.81	0.0076	11.53	0.060	−0.01
IDW	27.38	0.0302	22.98	0.124	−2.75
OK	17.97	0.0097	15.42	0.083	−0.32
MLP	8.58	0.0026	7.00	0.037	0.64
RF	10.34	0.0034	7.89	0.042	0.38
Proposed RF	7.92	0.0021	6.30	0.033	0.64

Table 5. Performance comparison of different methods on Dataset 1 for estimating 30% of BSFC data.

Method	RMSE	NMSE	MAE	MAPE	R^2
KNN	95.76	0.1048	51.53	0.133	0.06
IDW	143.36	0.2790	105.11	0.333	−2.04
OK	102.06	0.1220	69.59	0.210	−0.04
MLP	59.88	0.0450	45.53	0.148	0.55
RF	61.51	0.0465	33.97	0.094	0.56
Proposed RF	36.04	0.0178	21.16	0.064	0.86

Table 6. Performance comparison of different methods on Dataset 2 for estimating 30% of BSFC data.

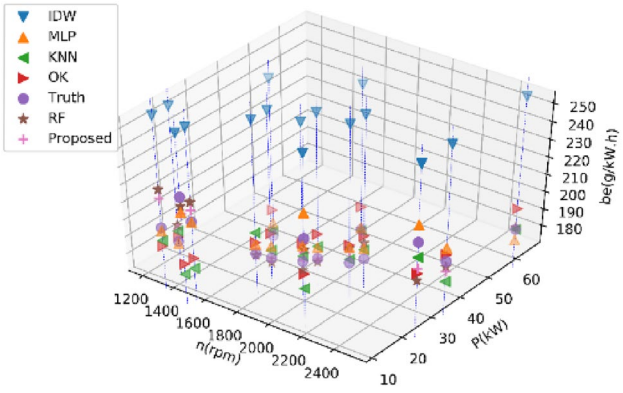

Figure 7. Results of different methods on Dataset 1 for estimating 30% of BSFC data.

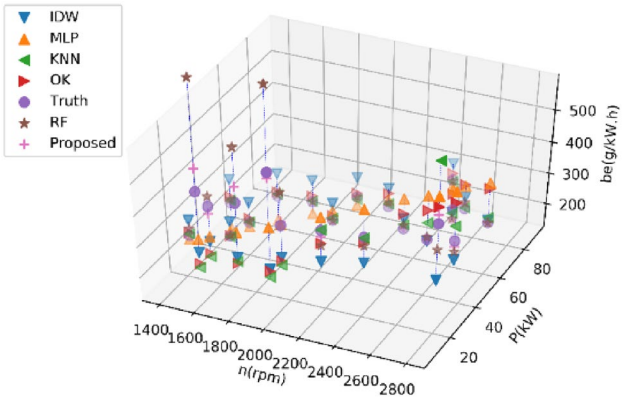

Figure 8. Results of different methods on Dataset 2 for estimating 30% of BSFC data.

Method	RMSE	NMSE	MAE	MAPE	R^2
KNN	19.95	0.0119	13.76	0.070	−0.26
IDW	25.97	0.0226	20.89	0.111	−1.37
OK	20.34	0.0121	15.96	0.083	−0.31
MLP	17.24	0.0114	14.20	0.076	−0.02
RF	10.93	0.0037	7.60	0.039	0.63
Proposed RF	9.54	0.0028	6.55	0.033	0.73

Table 7. Performance comparison of different methods on Dataset 1 for estimating 40% of BSFC data.

Method	RMSE	NMSE	MAE	MAPE	R^2
KNN	92.66	0.0952	56.46	0.158	0.08
IDW	159.96	0.3705	127.32	0.429	−3.23
OK	94.46	0.0988	69.50	0.214	0.04
MLP	59.27	0.0412	48.50	0.159	0.61
RF	72.29	0.0593	39.27	0.105	0.43
Proposed RF	41.02	0.0196	25.61	0.078	0.80

Table 8. Performance comparison of different methods on Dataset 2 for estimating 40% of BSFC data.

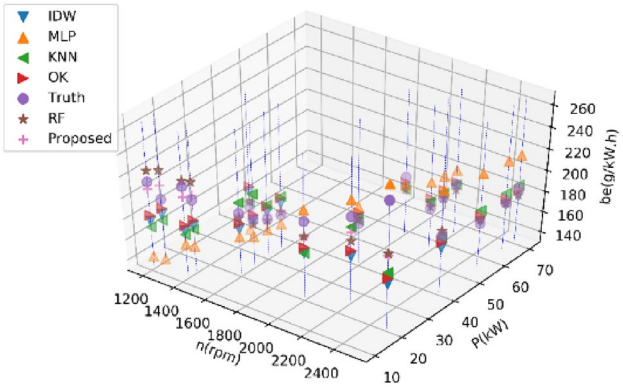

Figure 9. Results of different methods on Dataset 1 for estimating 40% of BSFC data.

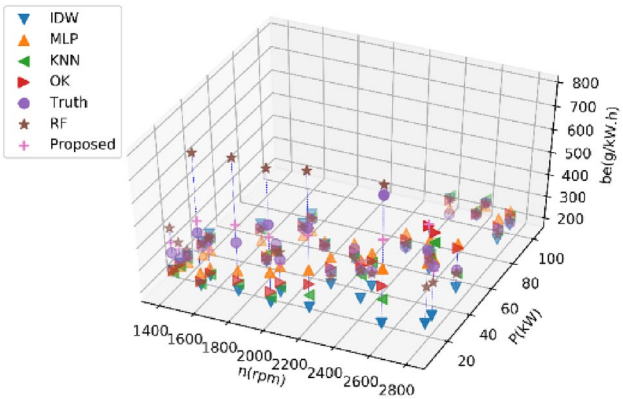

Figure 10. Results of different methods on Dataset 2 for estimating 40% of BSFC data.

proposed method outperforms other methods with an RMSE of 18.25 lower on Dataset 2. All the indexes show that the proposed RF has a minimal error and the highest accuracy.

The performance of different methods was compared on two datasets to estimate 20%, 30%, and 40% of BSFC data. All performance indicators indicate that the improved method proposed in this paper is the most accurate.

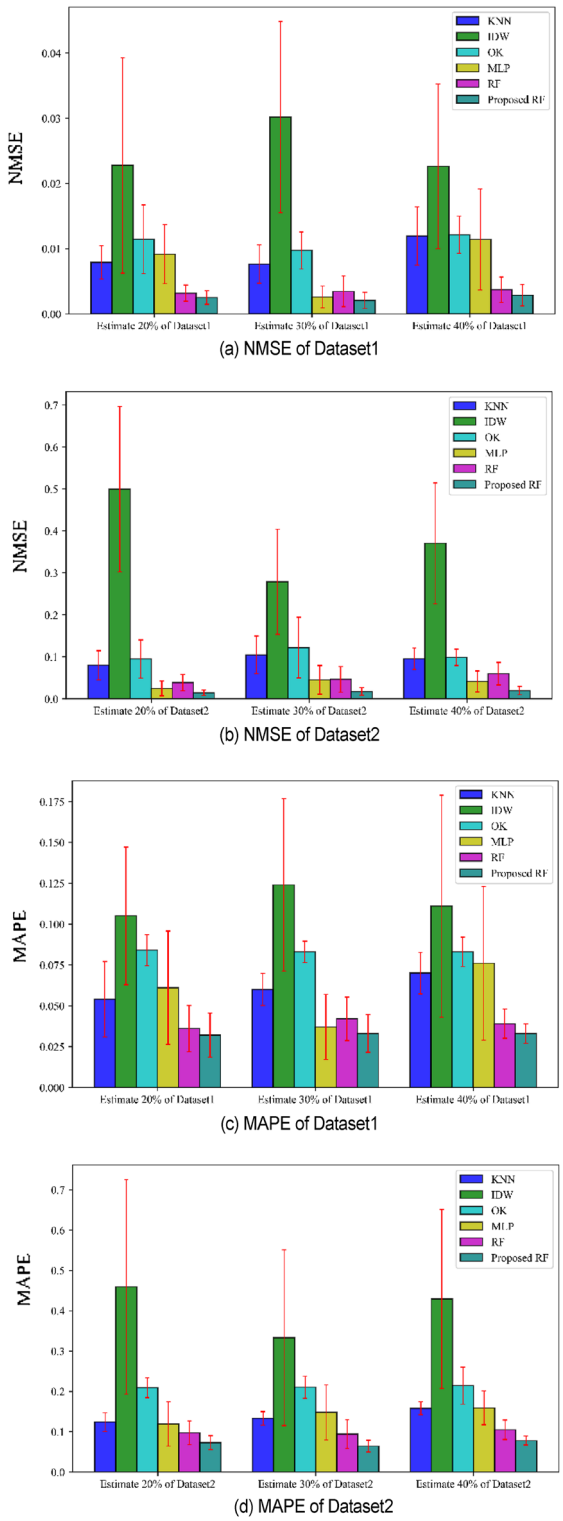

Figure 11. Comparison of NMS and MAPE with standard deviation.

Comparison of NMSE and MAPE

In order to analyze the distribution of performance indicators, the standard deviation of evaluation indicators is calculated. The two most important indicators, NMSE and MAPE, were selected to draw Fig. 11 and added to the paper. In this graph, bar charts with different methods, datasets, and estimated proportions of data were drawn, especially with standard deviations marked in the graph.

From Fig. 11, it can be seen that the improved random forest method proposed in this paper has the minimum average NMSE and MAPE on both dataset 1 and dataset 2. The sample standard deviations of NMSE and MAPE are also shown in the figure, indicating that the proposed method also has the smallest sample standard deviation. From Fig. 11, it can be seen that the improved random forest method proposed in this paper has the minimum average NMSE and MAPE on both dataset 1 and dataset 2. The sample standard deviations of NMSE and MAPE are also shown in the figure, indicating that the proposed method also has the smallest sample standard deviation. This analysis result is consistent with the previous analysis results.

Conclusions

The random forest method was introduced as an alternative approach for estimating brake-specific fuel consumption and was compared to commonly used calculation methods, such as the K-nearest neighbor method, inverse distance weighted method, ordinary kriging method, and multi-layer perceptron. The experimental results indicated that the proposed RF method outperformed the other methods in accuracy and precision. Therefore, it was concluded that the proposed RF method is more suitable for estimating the BSFC map compared to the other methods.

Data availability

All data generated or analyzed during this study are included in this published article.

Received: 12 August 2023; Accepted: 14 October 2023
Published online: 18 October 2023

References

1. Quan, R., Yue, Y. S., Huang, Z. K., Chang, Y. F. & Deng, Y. D. Effects of backpressure on the performance of internal combustion engine and automobile exhaust thermoelectric generator. *J. Energy Resour. Technol. Trans. ASME* **144**, 092301 (2022).
2. Bayat, Y. & Ghazikhani, M. Experimental investigation of compressed natural gas using in an indirect injection diesel engine at different conditions. *J. Clean. Prod.* **271**, 122450 (2020).
3. Mizythras, P., Boulougouris, E. & Theotokatos, G. A novel objective oriented methodology for marine engine-turbocharger matching. *Int. J. Engine Res.* **23**, 2105–2127 (2022).
4. Li, Y. Y. *et al.* Multi-objective energy management for Atkinson cycle engine and series hybrid electric vehicle based on evolutionary NSGA-II algorithm using digital twins. *Energy Convers. Manag.* **230**, 113788 (2021).
5. Zhang, H. G., Wang, E. H. & Fan, B. Y. A performance analysis of a novel system of a dual loop bottoming organic Rankine cycle (ORC) with a light-duty diesel engine. *Appl. Energy* **102**, 1504–1513 (2013).
6. Fang, C., Song, S., Chen, Z. *et al.* Fine-grained fuel consumption prediction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* 2783–2791 (2019).
7. Iwanski, G., Bigorajski, Ł & Koczara, W. Speed control with incremental algorithm of minimum fuel consumption tracking for variable speed diesel generator. *Energy Convers. Manag.* **161**, 182–192 (2018).
8. Durković, R. & Grujičić, R. An approach to determine the minimum specific fuel consumption and engine economical operation curve model. *Measurement* **132**, 303–308 (2019).
9. Satpathi, K., Balijepalli, V. & Ukil, A. Modeling and real-time scheduling of DC platform supply vessel for fuel efficient operation. *IEEE Trans. Transp. Electrification* **3**(3), 762–778 (2017).
10. Akar, A., Konakoğlu, B. & Akar, Ö. Prediction of geoid undulations: Random forest versus classic interpolation techniques. *Concurr. Comput. Pract. Exp.* **2022**, e7004 (2022).
11. Sekulić, A. *et al.* Random forest spatial interpolation. *Remote Sens.* **12**(10), 1687 (2020).
12. Silva, C. *et al.* Random forest techniques for spatial interpolation of evapotranspiration data from Brazilian's northeast. *Comput. Electron. Agric.* **166**, 105017 (2019).
13. Mariano, C. & Monica, B. A random forest-based algorithm for data-intensive spatial interpolation in crop yield mapping. *Comput. Electron. Agric.* **184**, 106094 (2021).
14. Daya, A. A. & Bejari, H. A comparative study between simple kriging and ordinary kriging for estimating and modeling the Cu concentration in Chehlkureh deposit, SE Iran. *Arab. J. Geosci.* **8**(8), 6003–6020 (2015).
15. Ma, L. Y., Liu, X. W., Zhang, Y. & Jia, S. L. Visual target detection for energy consumption optimization of unmanned surface vehicle. *Energy Rep.* **8**(4), 363–369 (2022).
16. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
17. Zhang, W. *et al.* A distributed storage and computation k-nearest neighbor algorithm based cloud-edge computing for cyber-physical-social systems. *IEEE Access* **8**, 50118–50130 (2020).
18. Mukhopadhyay, T. *et al.* A critical assessment of Kriging model variants for high-fidelity uncertainty quantification in dynamics of composite shells. *Arch. Comput. Methods Eng.* **24**(3), 495–518 (2017).
19. Panahi, F. *et al.* Streamflow prediction with large climate indices using several hybrid multilayer perceptions and copula Bayesian model averaging. *Ecol. Indic.* **133**, 108285 (2021).
20. Hadian, S., Shahiri, T. E. & Pham, Q. B. Multi attributive ideal-real comparative analysis (MAIRCA) method for evaluating flood susceptibility in a temperate Mediterranean climate. *Hydrol. Sci. J.* **67**(3), 401–418 (2022).
21. Zhu, M. Y. *et al.* Elastography ultrasound with machine learning improves the diagnostic performance of traditional ultrasound in predicting kidney fibrosis. *J. Formos. Med. Assoc.* **121**(6), 1062–1072 (2022).
22. Gummadi, J. *et al.* Interpolation techniques for modeling and estimating indoor radon concentrations in Ohio: Comparative study. *Environ. Prog. Sustain. Energy* **34**(1), 169–177 (2015).
23. Wan, D. Y., Liu, J. Q. & Ren, C. Y. Research of universal characteristics curve of internal combustion engines. *Middle South Autom. Transp.* (3), 5–8 (1998).
24. Zhou, G. M. *et al.* Universal characteristics curve plotting method based on MATLAB. *I.C.E. Powerpl.* **110**(2), 34–36 (2009).

Acknowledgements
This work supported by Research Program supported by the Department of Education and Technology (program name), Country Name.

Author contributions
QY wrote the main manuscript text, XW responsible for architecture design and overall review, CY responsible for diagram and drawing, algorithm verification, HW responsible for the design and implementation of algorithms. All authors reviewed the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Q.Y. or X.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.