Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2021.015378

ARTICLE

Variable Importance Measure System Based on Advanced Random Forest

Shufang Song[1,*], Ruyang He[1], Zhaoyin Shi[1] and Weiya Zhang[2]

[1]School of Aeronautics, Northwestern Polytechnical University, Xi'an, 710072, China

[2]AECC Sichuan Gas Turbine Establishment, Mianyang, 621700, China

[*]Corresponding Author: Shufang Song. Email: shufangsong@nwpu.edu.cn

Received: 14 December 2020 Accepted: 17 March 2021

ABSTRACT

The variable importance measure (VIM) can be implemented to rank or select important variables, which can effectively reduce the variable dimension and shorten the computational time. Random forest (RF) is an ensemble learning method by constructing multiple decision trees. In order to improve the prediction accuracy of random forest, advanced random forest is presented by using Kriging models as the models of leaf nodes in all the decision trees. Referring to the Mean Decrease Accuracy (MDA) index based on Out-of-Bag (OOB) data, the single variable, group variables and correlated variables importance measures are proposed to establish a complete VIM system on the basis of advanced random forest. The link of MDA and variance-based sensitivity total index is explored, and then the corresponding relationship of proposed VIM indices and variance-based global sensitivity indices are constructed, which gives a novel way to solve variance-based global sensitivity. Finally, several numerical and engineering examples are given to verify the effectiveness of proposed VIM system and the validity of the established relationship.

KEYWORDS

Variable importance measure; random forest; variance-based global sensitivity; Kriging model

Nomenclature

VIM	Variable Importance Measure
RF	Random Forest
DT	Decision Tree
MDI	Mean Decrease Impurity
MDA	Mean Decrease Accuracy
OOB	Out-of-Bag
SA	Sensitivity Analysis
MC	Monte Carlo
SDP	State-Dependent Parameter
HDMR	High Dimensional Model Representation
SGI	Sparse Grid Integration
ANOVA	Analysis of Variance

MSE	Mean Square Error
$\boldsymbol{X}$, Y	the input variable vector and output response
$g(\,)$	the response function
n	the dimension of input variables
g_0	the expectation of response function
$f_X(x)$	the probability density function of variable X
$E(\,)$, $Var(\,)$	the expectation and variance operator
$\boldsymbol{X}_{\sim i}$	the variable vector without X_i
$\boldsymbol{\mu}_{\sim i}$	the mean vector without μ_i
V, σ, ρ	the variance, standard variance and Pearson correlation coefficient of variable
$\boldsymbol{\mu}_{\boldsymbol{X}}$, $\boldsymbol{C}_{\boldsymbol{X}}$	the mean and covariance matrix of normal input variables
$\boldsymbol{\mu}_{\sim i\|i}$, $\boldsymbol{C}_{\sim i\|i}$	the conditional mean vector and conditional covariance matrix of dependent normal variables
$\mu_{i\|\sim i}$, $C_{i\|\sim i}$	the conditional mean and conditional covariance of dependent normal variable
$\boldsymbol{T}_m$	Bootstrap samples to train the m^{th} decision tree
h_m	the m^{th} decision tree of RF
η_i^T, η_i, η_{ij}	the defined variable importance measure of RF
N	the size of random samples
M	the number of decision trees of RF
S_i, S_{ij}	the variance-based global sensitivity indices
S_i^T, $S_{[i,j]}$	
ε_m, ε_m^i	the MSE of predicted values of RF
$\varepsilon_m^{\sim i}$, $\varepsilon_m^{\sim i,j}$	
$\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}_i$	the sample matrices of input variable samples
$\boldsymbol{X}_{OOB}$, $\boldsymbol{X}_{OOB}^i$	
$\boldsymbol{X}_{OOB}^{\sim i}$, $\boldsymbol{X}_{OOB}^{\sim i,j}$	
$\boldsymbol{y}_{\boldsymbol{A}}$, $\boldsymbol{y}_{\boldsymbol{B}}$, $\boldsymbol{y}_{\boldsymbol{C}_i}$, $\boldsymbol{y}$,	the response vectors of corresponding sample matrices
$\boldsymbol{y}_m$, $\boldsymbol{y}_m^i$, $\boldsymbol{y}_m^{\sim i}$, $\boldsymbol{y}_m^{\sim i,j}$	

1 Introduction

Sensitivity analysis can reflect the influence of input variables on the output response. The sensitivity analysis includes local sensitivity and global sensitivity analysis [1]. The local sensitivity can respond to the influence of input variables on the characteristics of output at the nominal value. The global sensitivity analysis, known as the importance measure analysis, can estimate the influence of input variables in the whole distribution region on the characteristics of output [2–4]. There are three kinds of importance measures: non-parametric measure, variance-based global sensitivity and moment-independent importance measure [1]. The variance-based global sensitivity is the most widely applied measure because it is generality and holistic, and it can give the contribution of group variables and the cross influence of different variables. There are plenty of methods to calculate variance-based global sensitivity indices, such as Monte Carlo (MC) simulation [5], high dimensional model representation (HDMR) [6], state-dependent parameter (SDP) procedure [7] and so on. MC simulation can estimate the approximate exact solution of total and main sensitivity indices simultaneously, but the amount of calculation is generally large, especially for high dimensional engineering problems. HDMR and SDP can calculate the main sensitivity indices by solving all order components of input-output surrogate models.

Random forest (RF) is composed by multiple decision trees (DTs), it is an ensemble learning method proposed by Breiman [8]. RF has many advantages, such as strong robustness, good tolerance to outliers and noise. RF has a wide range of application prospects, such as geographical energy [9], chemical industry [10], health insurance [11] and data science competitions. RF can not only deal with classification and regression problems but also analyze the critical measure. RF provides two kinds of importance measures: Mean Decrease Impurity (MDI) based on the Gini index and Mean Decrease Accuracy (MDA) based on Out-of-Bag (OOB) data [12]. MDI index is the average reduction of Gini impurity due to a splitting variable in the decision tree across RF [13]. MDI index is sensitive to variables with different scales of measurement and shows artificial inflation for variables with various categories. For correlated variables, the MDI index is related to the selection sequence of variables. Once a variable is selected, the impurity will be reduced by the first selected variable. It is difficult for the other correlated variables to reduce the same magnitude of impurity, so the importance of the other correlated variables will be decline. MDA index is the average reduction of prediction accuracy after randomly permuting OOB data [14,15]. Since MDA index can measure the impact of each variable on the prediction accuracy of RF model and have no biases, it has been widely used in many scientific areas. Although there are importance measures based on RF to distinguish the important features, there is no complete importance measure system to deal with nonlinearity and correlation among variables [16,17]. In addition, the similarity analysis process of MDA based on OOB data and Monte Carlo simulation of variance-based global sensitivity can be used as a breakthrough point to find their link [18]. With the help of variance-based sensitivity index system, the construction of variable importance measure system based on RF can be realized.

By comparing the procedure of estimating the total sensitivity indices and the MDA index based on OOB data, a complete VIM system is established based on advanced RF by using Kriging models, including single variable, group variables and correlated variables importance measure indices. The proposed VIM system combines the advantages of random forest and Kriging model. The VIM system can indicate the contribution of input variables to output response and rank important variables, and also give a novel way to solve variance-based global sensitivity with small samples.

This paper is organized as follows: Section 2 reviews the basic concept of variance-based global sensitivity. Section 3 reviews random forest firstly, presents MDA index and then proposes single variable, group variables and correlated variables importance measures respectively. Section 4 finds the link between MDA index and total variance-based global sensitivity index, and the relationship between VIM indices and variance-based global sensitivity indices is derived. In Section 5, several numerical and engineering examples are provided before the conclusions in Section 6.

2 Variance-Based Global Sensitivity

The variance-based global sensitivity, proposed by Sobol [19], reflects the influence of input variables in the whole distribution region on the variance of model output. The variance-based global sensitivity indices not only have strong model generality, but also can discuss the importance of group variables and quantify the interaction between input variables. ANOVA (Analysis of Variance) decomposition is the basic of variance-based global sensitivity analysis.

2.1 ANOVA Decomposition

Response function $Y = g(\boldsymbol{X})$ exists a unique ANOVA decomposition as follows:

$$g(\boldsymbol{X}) = g_0 + \sum_{i=1}^{n} g_i(X_i) + \sum_{1 \le i < j \le n} g_{ij}\left(X_i, X_j\right) + \ldots + g_{1\ldots n}(X_1, X_2, \ldots, X_n) \tag{1}$$

where n is the dimension of input variables, g_0 is the expectation of $g(\boldsymbol{X})$, $g_0 = \int_{R^n} g(\boldsymbol{x}) \prod_{i=1}^{n} \left[f_{X_i}(x_i)\,\mathrm{d}x_i\right]$, and $f_{X_i}(x_i)$ is the probability density function of variable X_i. The components in Eq. (1) are:

$$g_i(X_i) = \int_{R^{n-1}} g(\boldsymbol{x}) \prod_{j\neq i}^{n} \left[f_{X_j}(x_j)\,\mathrm{d}x_j\right] - g_0$$

$$g_{ij}(X_i, X_j) = \int_{R^{n-2}} g(\boldsymbol{x}) \prod_{k\neq i,j}^{n} \left[f_{X_k}(x_k)\,\mathrm{d}x_k\right] - g_i(X_i) - g_j(X_j) - g_0$$

2.2 Variance-Based Global Sensitivity Indices

The variance of response function can be expressed as:

$$V = Var(Y) = \int_{R^n} g^2(\boldsymbol{x}) \prod_{i=1}^{n} \left[f_{X_i}(x_i)\,\mathrm{d}x_i\right] - g_0^2 \tag{2}$$

Since the decomposition terms are orthogonal, the variance of the response function is the sum of variances of all individual decomposition terms:

$$V = \sum_{i=1}^{n} V_i + \sum_{1\leq i<j\leq n} V_{ij} + \ldots + V_{1,2,\ldots,n}$$

where

$$V_i = Var(g_i(X_i)) = \int_R g_i^2(x_i) f_{X_i}(x_i)\,\mathrm{d}x_i$$

$$V_{ij} = Var\left(g_{ij}(X_i, X_j)\right) = \iint_{R^2} g_{ij}^2(x_i, x_j) f_{X_i}(x_i) f_{X_j}(x_j)\,\mathrm{d}x_i\mathrm{d}x_j$$

Then the ratio of each variance component to variance of response function can reflect the variance contribution of each component, i.e., $S_i = V_i/V$, $S_{ij} = V_{ij}/V \cdots$

$S_i = V_i/V$ is the first order sensitivity index of variable X_i (also name S_i as main sensitivity index), it can reflect the influence of variable X_i on the response Y. $S_{ij} = V_{ij}/V$ is the second order sensitivity index, it can reflect the interaction influence of variables X_i and X_j on the response Y. The total sensitivity index S_i^T can be obtained by summing all the influence related to variable X_i:

$$S_i^T = S_i + \sum_{1\leq i<j\leq n} S_{ij} + \sum_{1\leq i<j<k\leq n} S_{ijk} + \ldots + S_{12\ldots n}$$

According to probability theory, the variance-based global sensitivity indices can be expressed as [20]:

$$S_i = \frac{Var\left[E(Y \mid X_i)\right]}{Var(Y)}$$

$$S_{ij} = \frac{Var\left[E\left(Y \mid X_i X_j\right)\right]}{Var(Y)}$$

$$S_i^T = \frac{Var(Y) - Var\left[E(Y \mid \boldsymbol{X}_{\sim i})\right]}{Var(Y)} = 1 - \frac{Var\left[E(Y \mid \boldsymbol{X}_{\sim i})\right]}{Var(Y)}$$

where $\boldsymbol{X}_{\sim i}$ indicates variable vector without X_i.

2.3 Simulation of Variance-Based Global Sensitivity Indices

Due to the enormous computational load, the traditional double-loop Monte Carlo simulation is not suitable for complex engineering problems [21]. The computational procedures of single-loop Monte Carlo simulation are listed as follows:

Step 1: Randomly generate two sample matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ based on the probability distribution of variables $\boldsymbol{X}$.

$$\boldsymbol{A} = \begin{bmatrix} x_{11} & \cdots & x_{i1} & \cdots & x_{n1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1N} & \cdots & x_{iN} & \cdots & x_{nN} \end{bmatrix}_{N\times n}, \quad \boldsymbol{B} = \begin{bmatrix} x_{1(N+1)} & \cdots & x_{i(N+1)} & \cdots & x_{n(N+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1(N+N)} & \cdots & x_{i(N+N)} & \cdots & x_{n(N+N)} \end{bmatrix}_{N\times n}$$

Step 2: Construct sample matrix $\boldsymbol{C}_i$, where the ith column of $\boldsymbol{C}_i$ comes from the ith column of $\boldsymbol{A}$, and the other columns come from the corresponding columns of $\boldsymbol{B}$.

$$\boldsymbol{C}_i = \begin{bmatrix} x_{1(N+1)} & \cdots & x_{i1} & \cdots & x_{n(N+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1(N+N)} & \cdots & x_{iN} & \cdots & x_{n(N+N)} \end{bmatrix}_{N\times n}$$

Step 3: The main and total sensitivity indices can be expressed as follows:

$$S_i = \frac{\frac{1}{N}\sum_{j=1}^{N} y_{\boldsymbol{A}}^j y_{\boldsymbol{C}_i}^j - g_0^2}{Var\,(Y)} \tag{3}$$

$$S_i^T = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} y_{\boldsymbol{B}}^j y_{\boldsymbol{C}_i}^j - g_0^2}{Var\,(Y)} \tag{4}$$

where $\boldsymbol{y_A} = \left[y_{\boldsymbol{A}}^1, \ldots, y_{\boldsymbol{A}}^N\right]$, $\boldsymbol{y_B} = \left[y_{\boldsymbol{B}}^1, \ldots, y_{\boldsymbol{B}}^N\right]$, $\boldsymbol{y}_{\boldsymbol{C}_i} = \left[y_{\boldsymbol{C}_i}^1, \ldots, y_{\boldsymbol{C}_i}^N\right]$ are the model outputs with the input matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}_i$ respectively. The computational cost of single-loop Monte Carlo simulation is $(n+2)\times N$.

3 Variable Importance Measure System Based on Random Forest

RF is an ensemble statistical learning method to deal with classification and regression problems [22]. Bootstrap sampling technique is firstly carried out to extract training samples from the original data, and these training samples are used to build a decision tree; the rest Out-of-Bag data are used to verify the accuracy of established decision tree.

There are M established decision trees by employing Bootstrap sampling technique M times. All decision trees are used to compose a random forest (shown in Fig. 1). And the final prediction results of RF are obtained by voting in the classification model or taking the mean in the regression model [23]. And the prediction precision of RF can be expressed by mean square error square error (MSE) between predicted values and true values of OOB data.

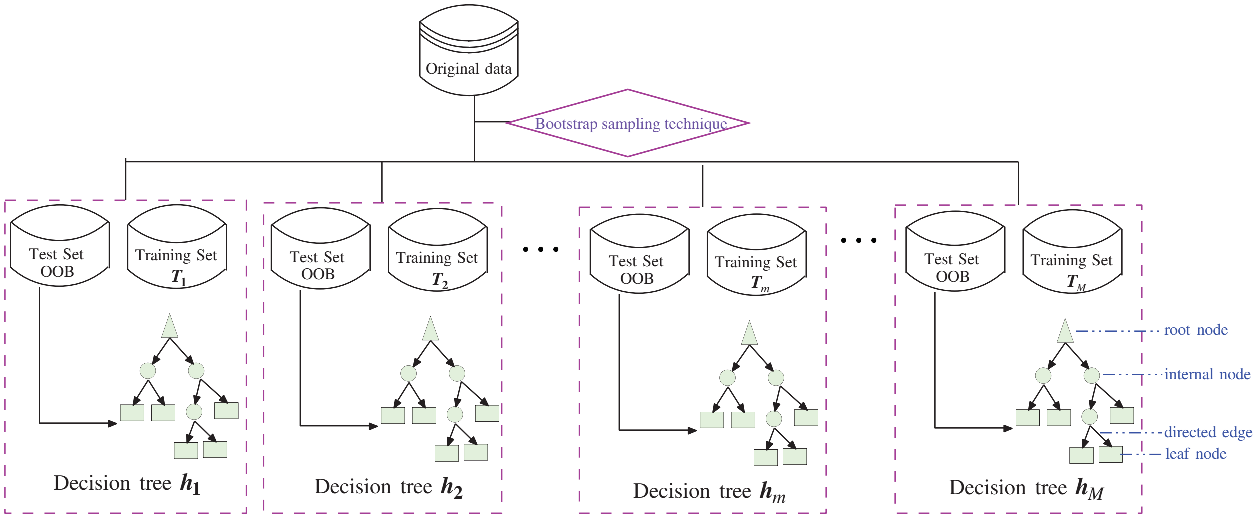


Figure 1: Random forest

Bootstrap technique can extract training points to build a decision tree h_m $(m = 1, 2, \ldots, M)$ and the corresponding OOB data of input $\boldsymbol{X}_{OOB}$ and output $\boldsymbol{y}$. The decision tree h_m is used to predict the forecast response $\boldsymbol{y}_m$ of $\boldsymbol{X}_{OOB}$. The MSE of decision tree h_m is $\varepsilon_m = \text{mean}\left(\boldsymbol{y}_m - \boldsymbol{y}\right)^2$. Obtain the MSEs of all decision trees ε_m $(m = 1, 2, \ldots, M)$, the average will be the total predicted error of RF model [24]:

$$MSE = \frac{1}{M}\sum_{m=1}^{M} \varepsilon_m \tag{5}$$

In order to improve the prediction precision of RF, a high-precision Kriging model is used as the model of leaf nodes in the decision tree, replacing the original average or linear regression. Next, a nonlinear discontinuous function is used to verify the prediction accuracy of Kriging model and linear regression model of decision tree.

$$Y = \begin{cases} -X^2 + 10\cos(2\pi X) - 30 & X < 0 \\ X^2 - 10\cos(2\pi X) + 30 & X \geq 0 \end{cases}$$

where the input variable X is uniformly distributed on $[-\pi, \pi]$.

A comparison of Kriging based decision tree (abbreviated as Kriging-DT) and linear regression based decision tree (abbreviated as Linear-DT) for prediction data are shown in Fig. 2. With the increase of training samples, the predicted errors of Kriging-DT and linear-DT are shown in Fig. 3. And it can be found that Kriging-DT can better approximate the original function. For the same training samples, Kriging-DT has higher prediction accuracy and faster decline rate of predicted error than Linear-DT. Kriging-DT inherits the advantages of Kriging model and has good applicability for nonlinear piecewise function.

There are two kinds of importance measures based on RF: Mean Decrease Impurity (MDI) based on Gini index and Mean Decrease Accuracy (MDA) based on OOB data. MDA index is widely used to rank important variables on the prediction accuracy of RF model [12].

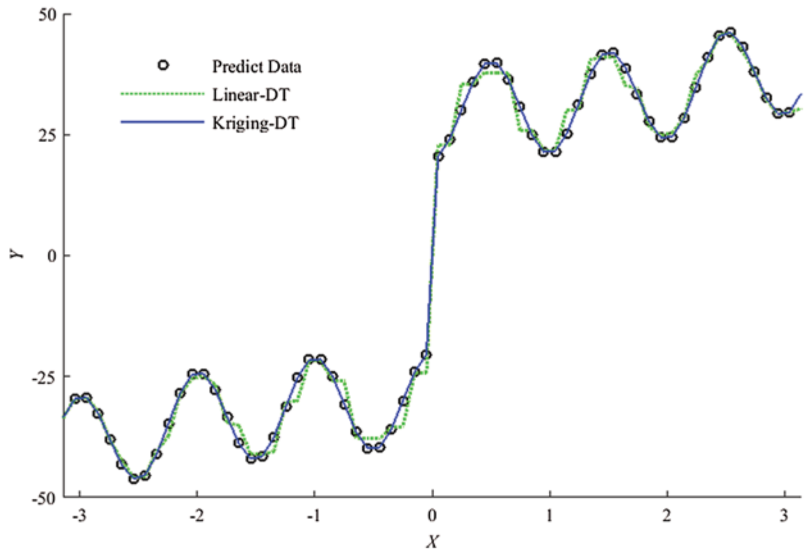


Figure 2: Comparsion of Kriging-DT, Linear-DT and predict data with 64 training samples

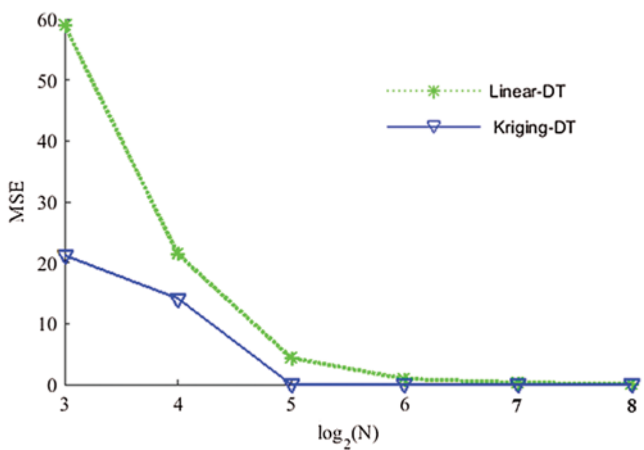


Figure 3: Predicted errors of Kriging-DT and Linear-DT *vs.* size of training samples

3.1 Mean Decrease Accuracy Index of Random Forest

MDA index is the average reduction of prediction accuracy after randomly permuting OOB data. Permuting the order of variable in OOB data, the corresponding relationship between the OOB sample and output will be destroyed. The prediction accuracy will be calculated after each permutation. The MSE between the paired predictions is taken as the importance measure.

For the decision tree h_m $(m = 1, 2, \ldots, M)$, the corresponding OOB input data is matrix $\boldsymbol{X}_{OOB} = \left(\boldsymbol{X}^1_{OOB}, \ldots, \boldsymbol{X}^i_{OOB}, \ldots, \boldsymbol{X}^n_{OOB}\right)$, $\boldsymbol{X}^i_{OOB}$ is the ith column of matrix $\boldsymbol{X}_{OOB}$. Permute the order of $\boldsymbol{X}^i_{OOB}$, decision tree h_m can obtain the new forecast response $\boldsymbol{y}^i_m$. The MSE of predicted values is $\varepsilon^i_m = \text{mean}\left(\boldsymbol{y}^i_m - \boldsymbol{y}_m\right)^2$. Obtain the influence of variable X_i in all decision trees

$\left(\varepsilon_1^i, \varepsilon_2^i, \ldots, \varepsilon_M^i\right)$, the average of ε_m^i $(m = 1, 2, \ldots, M)$ is the total impact of variable X_i based on the RF model:

$$\eta_i^T = \frac{1}{M}\sum_{m=1}^{M} \varepsilon_m^i \tag{6}$$

The subscript m of ε_m^i and $\boldsymbol{y}_m^i$ is the number of decision tree h_m $(m = 1, 2, \ldots, M)$, and the superscript i of ε_m^i and $\boldsymbol{y}_m^i$ indicates that the ith column of $\boldsymbol{X}_{OOB}$ is in disorder, corresponding to the variable X_i.

Based on the procedure of MDA index, the single variable, group variables and correlated variables importance measures are expanded to establish the variable importance measure system.

3.2 Single Variable Importance Measure of Random Forest

For the decision tree h_m $(m = 1, 2, \ldots, M)$, the order of OOB input data $\boldsymbol{X}_{OOB} = \left(\boldsymbol{X}_{OOB}^1, \ldots, \boldsymbol{X}_{OOB}^i, \ldots, \boldsymbol{X}_{OOB}^n\right)$ is randomly permuted expected $\boldsymbol{X}_{OOB}^i$, that is to say, the value of variable X_i is fixed, and the values of the other variables are randomly permuted. Then the decision tree can predict the modified OOB samples to get the predicted values $\boldsymbol{y}_m^{\sim i}$, the MSE of predicted values is $\varepsilon_m^{\sim i} = \text{mean}\left(\boldsymbol{y}_m^{\sim i} - \boldsymbol{y}_m\right)^2$. Obtain the influence of variable X_i in all decision trees, the average of $\varepsilon_m^{\sim i}$ is the main impact of variable X_i based on the RF model:

$$\eta_i = \frac{1}{M}\sum_{m=1}^{M} \varepsilon_m^{\sim i} \tag{7}$$

The superscript $\sim i$ of $\varepsilon_m^{\sim i}$ and $\boldsymbol{y}_m^{\sim i}$ indicates that the OOB data are permuted, expect for the ith columns.

3.3 Group Variable Importance Measure of Random Forest

The MDA index of group variables can be presented as follows. In the process of permuting OOB data, the values of variables X_i and X_j are fixed, and the values of the other variables are permuted. The decision tree can predict the modified OOB samples to get the predicted values $\boldsymbol{y}_m^{\sim i,j}$, the MSE of predicted values is $\varepsilon_m^{\sim i,j} = \text{mean}\left(\boldsymbol{y}_m^{\sim i,j} - \boldsymbol{y}_m\right)^2$. Obtain the influence of group variables $[X_i, X_j]$ in all decision trees, the average of $\varepsilon_m^{\sim i,j}$ is the main impact of group variables $[X_i, X_j]$ based on the RF model:

$$\eta_{ij} = \frac{1}{M}\sum_{m=1}^{M} \varepsilon_m^{\sim i,j} \tag{8}$$

The superscript $\sim i, j$ of $\varepsilon_m^{\sim i,j}$ and $\boldsymbol{y}_m^{\sim i,j}$ indicates that the OOB data are permuted, expect for the ith and jth columns.

3.4 Correlated Variable Importance Measure of Random Forest

With the past years, several techniques based on RF are proposed to measure the importance of the correlated variables [25,26]. However, these researches directly use the independent importance measure techniques to estimate the importance of the correlated variables, which is not reasonable. Reference [27,28] divided the variance-based sensitivity indices into correlated contribution and independent contribution. Moreover, sparse grid integration (SGI) is carried out to perform importance analysis for correlated variables [29]. In the paper, the correlation of correlated variables is considered in the process of the RF importance measure. The necessary procedure of a single decision tree of the RF model for estimating the VIM consists of the following steps:

Step 1: Estimate the covariance matrix $\boldsymbol{C_X}$ and mean vector $\boldsymbol{\mu_X}$ from the original data $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_i, \ldots, \boldsymbol{X}_n)$;

Step 2: Randomly extract the OOB data $\boldsymbol{X}_{OOB} = \left(\boldsymbol{X}^1_{OOB}, \ldots, \boldsymbol{X}^i_{OOB}, \ldots, \boldsymbol{X}^n_{OOB}\right)$ from the original data and use the other data to build the decision tree h_m $(m = 1, 2, \ldots, M)$. Use the decision tree h_m to predict the corresponding OOB data, and the prediction is $\boldsymbol{y}_m$;

Step 3: Split the matrix $\boldsymbol{X}_{OOB}$ into two parts: vector $\boldsymbol{X}^i_{OOB}$ and matrix $\boldsymbol{X}^{\sim i}_{OOB}$;

Step 4: Generate a new matrix $\boldsymbol{X}_{\sim i|i}$ and vector $\boldsymbol{X}_{i|\sim i}$ based on $\boldsymbol{X}^i_{OOB}$ and $\boldsymbol{X}^{\sim i}_{OOB}$, respectively. The mean vectors and covariance matrixes are different from the original $\boldsymbol{\mu_X}$ and $\boldsymbol{C_X}$, the new ones should be used in the transformation process. For the multivariate normal distribution, $\boldsymbol{\mu}_{\sim i|i}$, $\mu_{i|\sim i}$, $\boldsymbol{C}_{\sim i|i}$ and $C_{i|\sim i}$ can be acquired as follows:

The mean vector $\boldsymbol{\mu_X}$ and covariance matrix $\boldsymbol{C_X}$ of $\boldsymbol{X}$ can be separated as $\boldsymbol{\mu_X} = [\boldsymbol{\mu}_{\sim i}, \mu_i]$ and $\boldsymbol{C_X} = \begin{bmatrix} \boldsymbol{C}_{\sim i} & \boldsymbol{C}_{\sim i,i} \\ \boldsymbol{C}_{i,\sim i} & C_i \end{bmatrix}$. The conditional mean vector and covariance matrix can be obtained by the following formulas [30]:

$$\boldsymbol{\mu}_{\sim i|i} = \boldsymbol{\mu}_{\sim i} + \boldsymbol{C}_{\sim i,i} C_i^{-1} (X_i - \mu_i) \quad \mu_{i|\sim i} = \mu_i + \boldsymbol{C}_{i,\sim i} \boldsymbol{C}_{\sim i}^{-1} \left(\boldsymbol{X}_{\sim i} - \boldsymbol{\mu}_{\sim i}\right)$$

$$\boldsymbol{C}_{\sim i|i} = \boldsymbol{C}_{\sim i} - \boldsymbol{C}_{\sim i,i} C_i^{-1} \boldsymbol{C}_{i,\sim i} \quad C_{i|\sim i} = C_i - \boldsymbol{C}_{i,\sim i} \boldsymbol{C}_{\sim i}^{-1} \boldsymbol{C}_{\sim i,i}$$

After obtaining the corresponding $\boldsymbol{\mu}_{\sim i|i}$, $\mu_{i|\sim i}$, $\boldsymbol{C}_{\sim i|i}$ and $C_{i|\sim i}$, Nataf transform can be employed to extract normal correlation samples $\boldsymbol{X}_{\sim i|i}$ and $\boldsymbol{X}_{i|\sim i}$ directly.

Step 5: Combine matrix $\boldsymbol{X}_{\sim i|i}$ with vector $\boldsymbol{X}^i_{OOB}$ as the new matrix $\boldsymbol{X}^i_{OOBnew} = \left(\boldsymbol{X}^1_{\sim i|i}, \ldots, \boldsymbol{X}^{i-1}_{\sim i|i}, \boldsymbol{X}^i_{OOB}, \boldsymbol{X}^{i+1}_{\sim i|i}, \ldots, \boldsymbol{X}^n_{\sim i|i}\right)$, while combine vector with the matrix $\boldsymbol{X}^{\sim i}_{OOB}$ as $\boldsymbol{X}^{\sim i}_{OOBnew} = \left(\boldsymbol{X}^1_{OOB}, \ldots, \boldsymbol{X}^{i-1}_{OOB}, \boldsymbol{X}_{i|\sim i}, \boldsymbol{X}^{i+1}_{OOB}, \ldots, \boldsymbol{X}^n_{OOB}\right)$;

Step 6: $\boldsymbol{X}^i_{OOBnew}$ and $\boldsymbol{X}^{\sim i}_{OOBnew}$ are passed down the decision tree and the predicted values $\boldsymbol{y}^i_m$ and $\boldsymbol{y}^{\sim i}_m$ are computed, respectively. ε^i_m and $\varepsilon^{\sim i}_m$ of the correlated variables can be calculated by the following formula:

$$\varepsilon^{\sim i}_m = \text{mean}\left(\boldsymbol{y}^{\sim i}_m - \boldsymbol{y}_m\right)^2 \quad \varepsilon^i_m = \text{mean}\left(\boldsymbol{y}^i_m - \boldsymbol{y}_m\right)^2$$

Obtain the influence of variable X_i in all decision trees, the averages of $\varepsilon^{\sim i}_m$ and ε^i_m $(m = 1, 2, \ldots, M)$ are the main and total impact of variable X_i on the RF model.

The importance measure indices in correlated space and independent space are all given based on RF, which will establish the complete VIM system.

4 Link between VIM of RF and Variance-Based Global Sensitivity

The similarity analysis process of MDA index ε_m^i based on OOB data and single-loop Monte Carlo simulation of variance-based global sensitivity can be used as a breakthrough point to find their link. The relationship between MDA index and variance-based global sensitivity can be explored firstly.

1) MDA index ε_m^i can be decomposed as follows:

$$\begin{aligned}\varepsilon_m^i &= \text{mean}\left(\boldsymbol{y}_m^i - \boldsymbol{y}_m\right)^2 = \frac{1}{N}\sum_{j=1}^{N}\left(y_{m,j}^i - y_{m,j}\right)^2 \\ &= \frac{1}{N}\sum_{j=1}^{N}\left[\left(y_{m,j}^i\right)^2 + \left(y_{m,j}\right)^2 - 2y_{m,j}y_{m,j}^i\right] = \frac{1}{N}\sum_{j=1}^{N}\left(y_{m,j}^i\right)^2 + \frac{1}{N}\sum_{j=1}^{N}\left(y_{m,j}\right)^2 - \frac{2}{N}\sum_{j=1}^{N}y_{m,j}y_{m,j}^i \end{aligned} \tag{9}$$

When the sample size is large, $\frac{1}{N}\sum_{j=1}^{N}\left(y_{m,j}^i\right)^2$ asymptotically equals $\frac{1}{N}\sum_{j=1}^{N}\left(y_{m,j}\right)^2$, they are both second-order moment estimators of output response Y.

The total sensitivity index of single-loop Monte Carlo numerical simulation is:

$$S_i^T = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N} y_{\boldsymbol{B}}^j y_{\boldsymbol{C_i}}^j - g_0^2}{Var\,(Y)} = \frac{\frac{1}{N}\sum_{j=1}^{N}\left(y_{\boldsymbol{B}}^j\right)^2 - \frac{1}{N}\sum_{j=1}^{N} y_{\boldsymbol{B}}^j y_{\boldsymbol{C_i}}^j}{Var\,(Y)} \tag{10}$$

By comparison, it can be concluded that:

$$S_i^T = \frac{\varepsilon_m^i}{2 \times Var\,(Y)} \tag{11}$$

Thus, the relationship between MDA index of RF importance measure and variance-based global sensitivity indices is explored. ε_m^i can indicate the total impact of variable X_i on output performance. The larger ε_m^i is, the larger S_i^T is, which means that the total contribution of variable on output performance is larger.

2) The main variance-based sensitivity index S_i of single-loop Monte Carlo numerical simulation is equivalent to:

$$S_i = \frac{\frac{1}{N}\sum_{j=1}^{N} y_{\boldsymbol{A}}^j y_{\boldsymbol{C_i}}^j - g_0^2}{Var\,(Y)} - 1 + 1 = 1 - \frac{\frac{1}{N}\sum_{j=1}^{N}\left(y_{\boldsymbol{A}}^j\right)^2 - \frac{1}{N}\sum_{j=1}^{N} y_{\boldsymbol{A}}^j y_{\boldsymbol{C_i}}^j}{Var\,(Y)} \tag{12}$$

By comparison, it can be concluded that:

$$S_i = 1 - \frac{\varepsilon_m^{\sim i}}{2 \times Var\,(Y)} \tag{13}$$

Eq. (13) shows the relationship between $\varepsilon_m^{\sim i}$ and the main variance-based sensitivity index S_i. Index $\varepsilon_m^{\sim i}$ can indicate the main impact of variable X_i on output performance. The larger $\varepsilon_m^{\sim i}$ is, the smaller S_i is, which means that the main contribution of variable on output performance is smaller.

3) The relationship of variance-based sensitivity index of group variables $S_{[i,j]}$ and $\varepsilon_m^{\sim i,j}$ can be expressed as:

$$S_{[i,j]} = 1 - \frac{\varepsilon_m^{\sim i,j}}{2 \times Var\,(Y)} \tag{14}$$

The influence of group variables $\left[X_i, X_j\right]$ on the variance of output $S_{[i,j]}$ is composed of the main sensitivity indices S_i, S_j and second order sensitivity index S_{ij}.

$$S_{[i,j]} = S_i + S_j + S_{ij} \tag{15}$$

Combining Eqs. (13)–(15), the second-order variance sensitivity index can be derived:

$$S_{ij} = \frac{\varepsilon_m^{\sim i} + \varepsilon_m^{\sim j} - \varepsilon_m^{\sim i,j}}{2 \times Var\,(Y)} - 1 \tag{16}$$

So far, the MDA index, single variable index and group variables index are all proposed in the independent variable space.

4) In the correlated variable space, $Var\,(Y) \neq Var\left(y_m^{\sim i}\right) \neq Var\left(y_m^{i}\right)$, Eqs. (11) and (13) should be changed into the following formulas:

$$S_i = 1 - \frac{\varepsilon_m^{\sim i} - E\left(y_m^{\sim i}\right)^2 + E\left(y_m\right)^2}{2 \times Var\,(Y)} \tag{17}$$

$$S_i^T = \frac{\varepsilon_m^{i} - E\left(y_m^{i}\right)^2 + E\left(y_m\right)^2}{2 \times Var\,(Y)} \tag{18}$$

S_i contains the independent contribution of variable X_i and the correlated contribution of Pearson correlation coefficient, while S_i^T consists of the independent contribution by variable itself and interaction contribution with other variables.

5 Examples and Discussion

5.1 Numerical Example 1: Ishigami Function

Ishigami function is considered:

$$Y = \sin\,(X_1) + 7\sin^2\,(X_2) + 0.1X_3^4\sin\,(X_1)$$

where X_i are uniformly distributed on the interval $[-\pi, \pi]$, and the variables are independent. Ishigami function is a highly nonlinear function. For variable X_2, the convergence trends of importance measures with the number of sample points by Monte Carlo simulation and RF are shown in Fig. 4. There are 500 decision trees in the RF model. Tabs. 1 and 2 show the VIM results of single variable and group variables respectively. The analytical results ($S_i^{(\mathrm{Ana})}$, $S_i^{T(\mathrm{Ana})}$ and $S_{ij}^{(\mathrm{Ana})}$) are also presented in Tabs. 1 and 2 for comparison.

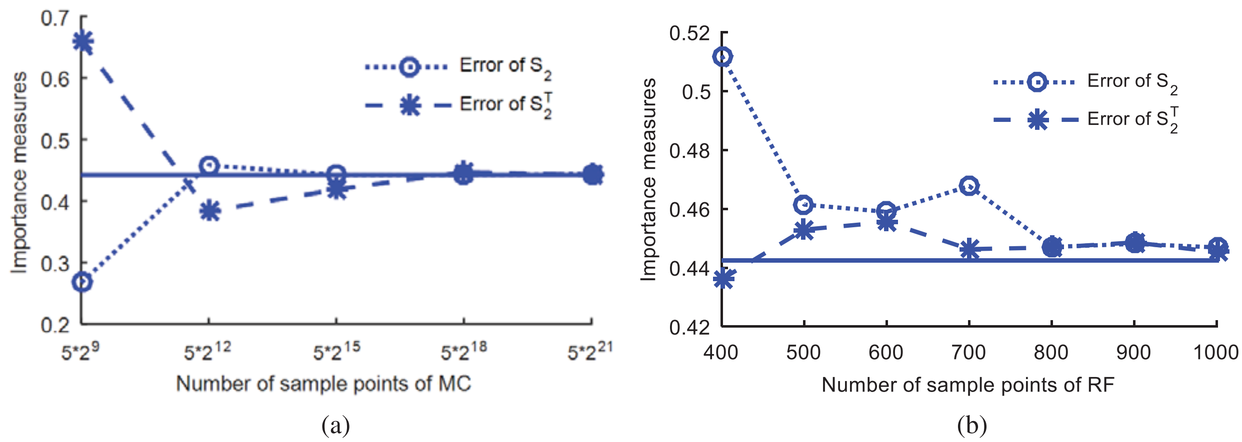


Figure 4: The convergence trends of the important measures with sample size (a) The convergence trend of MC simulation (b) The convergence trend of RF model

Table 1: The single variable VIMs of Ishigami function

	η_i	$\eta_i \Rightarrow S_i$	$S_i^{(\mathrm{Ana})}$	Error	η_i^T	$\eta_i^T \Rightarrow S_i^T$	$S_i^{T(\mathrm{Ana})}$	Error (%)
X_1	18.997	0.314	0.314	–	15.359	0.555	0.558	0.54
X_2	15.316	0.447	0.442	1.13%	12.331	0.445	0.442	0.68
X_3	27.784	0.003	0.000	–	6.690	0.242	0.244	0.82

Table 2: The group variables VIMs of Ishigami function

	η_{ij}	$\eta_{ij} \Rightarrow S_{ij}$	$S_{ij}^{(\mathrm{Ana})}$	Error
X_1X_2	6.698	0.003	0.000	–
X_1X_3	12.413	0.241	0.244	1.23%
X_2X_3	15.364	0.002	0.000	–

In all VIMs results tables, $\eta_i^T \Rightarrow S_i^T$, $\eta_i \Rightarrow S_i$ and $\eta_{ij} \Rightarrow S_{ij}$ mean that importance measures in this column are derived from Eqs. (11), (13) and (16), respectively.

There are 5×10^{20} random samples in single-loop Monte Carlo simulation to achieve the required accuracy, RF model only needs 10^3 samples (seen from Fig. 4). The comparison shows that RF method has faster convergence. The MDA indices of RF can get the variance-based sensitivity indices consistent with the analytical solutions (seen from Tabs. 1 and 2), which suggests the RF model provides high accuracy. For the Ishigami function, the third-order sensitivity index $S_{123} = 0$, so the relationship of the variance-based sensitivity indices is $S_i^T = S_i + \sum_{j \neq i} S_{ij}$, which has a good agreement with the VIM estimators.

5.2 Numerical Example 2: Linear Function with Correlated Variables

A linear model is considered [28]:

$$Y = X_1 + X_2 + X_3$$

where X_i are normally distributed with $\boldsymbol{\mu}_{\boldsymbol{X}} = [0, 0, 0]$ and covariance matrix $\boldsymbol{C}_{\boldsymbol{X}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho\sigma \\ 0 & \rho\sigma & \sigma^2 \end{bmatrix}$.

Analytical solutions for the main and total sensitivity indices can be calculated as:

$$S_1 = \frac{1}{2+\sigma^2+2\rho\sigma}, \quad S_2 = \frac{(1+\rho\sigma)^2}{2+\sigma^2+2\rho\sigma}, \quad S_3 = \frac{(\rho+\sigma)^2}{2+\sigma^2+2\rho\sigma}$$

$$S_1^T = \frac{1}{2+\sigma^2+2\rho\sigma}, \quad S_2^T = \frac{1-\rho^2}{2+\sigma^2+2\rho\sigma}, \quad S_3^T = \frac{\sigma^2\left(1-\rho^2\right)}{2+\sigma^2+2\rho\sigma}$$

There are 500 decision trees and 600 samples used to analyze the importance measures. Fig. 5 shows the importance measures of the correlated input variables with different ρs. Tab. 3 shows the importance measures of independent and correlated variables cases at $\sigma = 2$. Additionally, the analytical solutions are also presented for comparison.

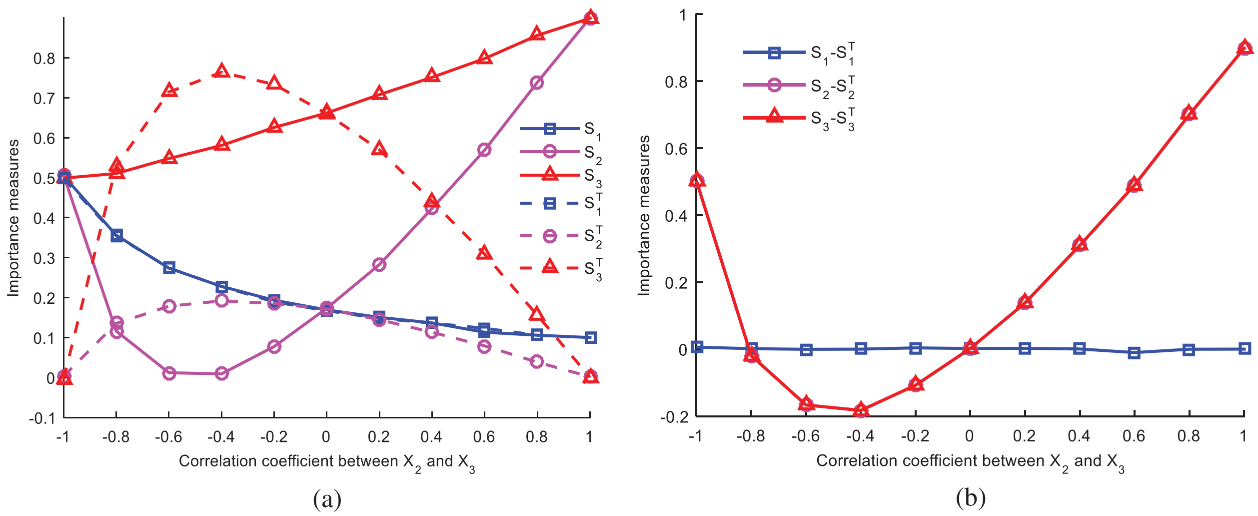


Figure 5: The importance measures of correlated input variables at different correlation coefficients (a) Importance measures *vs.* correlation coefficients (b) $S_i - S_i^T$ *vs.* correlation coefficients

All the importance measures for correlated variables and independent ones are simulated. From the analytical results of main and total sensitivity indices, it can be found that $S_i^T \le S_i$ if $\rho \ge 0$ or $\rho \le -\frac{2\sigma}{\sigma^2+1}$. The interaction sensitivity indices are all equal to zero, so $S_i - S_i^T$ only contain the correlated contribution by the Pearson correlation coefficients. For variable X_1, the main sensitivity index S_1 is equal to total indices S_1^T and $S_1 - S_1^T = 0$, because of the independence of the variable X_1 with other variables. For the variables X_2 and X_3, $S_2 - S_2^T = S_3 - S_3^T$, which suggests that the correlated contribution is generated from Pearson correlation coefficients.

Table 3: The single variable VIMs of Example 5.2

ρ		η_i	$\eta_i \Rightarrow S_i$	$S_i^{(\mathrm{Ana})}$	Error	η_i^T	$\eta_i^T \Rightarrow S_i^T$	$S_i^{T(\mathrm{Ana})}$	Error
0	X_1	9.909	0.163	0.167	2.39%	1.957	0.166	0.167	0.60%
	X_2	9.921	0.162	0.167	2.99%	1.975	0.168	0.167	0.60%
	X_3	3.930	0.667	0.667	–	7.918	0.669	0.667	0.30%
0.5	X_1	14.031	0.124	0.125	0.80%	1.685	0.123	0.125	1.60%
	X_2	8.964	0.498	0.500	0.40%	1.742	0.094	0.094	—
	X_3	3.423	0.781	0.781	–	7.277	0.370	0.375	1.33%
−0.5	X_1	6.440	0.244	0.250	2.40%	1.707	0.252	0.250	0.80%
	X_2	8.927	0.001	0.000	–	1.745	0.190	0.188	1.06%
	X_3	3.444	0.555	0.563	1.42%	7.248	0.754	0.750	0.53%
0.8	X_1	16.527	0.102	0.109	6.42%	1.292	0.106	0.109	2.75%
	X_2	7.330	0.739	0.735	0.54%	1.344	0.039	0.039	—
	X_3	2.624	0.856	0.852	0.47%	6.008	0.150	0.157	4.46%
−0.8	X_1	4.765	0.356	0.357	0.28%	1.298	0.360	0.357	0.84%
	X_2	7.389	0.129	0.129	–	1.355	0.126	0.129	2.33%
	X_3	2.659	0.511	0.514	0.58%	6.012	0.504	0.514	1.95%

5.3 Numerical Example 3: Nonlinear Function with Correlated Variables

Consider a nonlinear model $Y = X_1X_3 + X_2X_4$ [28], where $\boldsymbol{X} \sim N(\boldsymbol{\mu}_{\boldsymbol{X}}, \boldsymbol{C}_{\boldsymbol{X}})$ with $\boldsymbol{\mu}_{\boldsymbol{X}} = [0, 0, \mu_3, \mu_4]$ and covariance matrix $\boldsymbol{C}_{\boldsymbol{X}} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & 0 & 0 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \rho_{34}\sigma_3\sigma_4 \\ 0 & 0 & \rho_{34}\sigma_3\sigma_4 & \sigma_4^2 \end{bmatrix}$.

Analytical values of main and total sensitivity indices are:

$$S_1 = \frac{\sigma_1^2\left(\mu_3 + \mu_4\rho_{12}\frac{\sigma_2}{\sigma_1}\right)^2}{V}, \quad S_2 = \frac{\sigma_2^2\left(\mu_4 + \mu_3\rho_{12}\frac{\sigma_1}{\sigma_2}\right)^2}{V}, \quad S_3 = S_4 = 0$$

$$S_1^T = \frac{\sigma_1^2\left(1-\rho_{12}^2\right)\left(\sigma_3^2+\mu_3^2\right)}{V}, \quad S_2^T = \frac{\sigma_2^2\left(1-\rho_{12}^2\right)\left(\sigma_4^2+\mu_4^2\right)}{V}, \quad S_3^T = \frac{\sigma_1^2\sigma_3^2\left(1-\rho_{34}^2\right)}{V},$$

$$S_4^T = \frac{\sigma_2^2\sigma_4^2\left(1-\rho_{34}^2\right)}{V}$$

where $V = \sigma_1^2\left(\sigma_3^2+\mu_3^2\right) + \sigma_2^2\left(\sigma_4^2+\mu_4^2\right) + 2\rho_{12}\sigma_1\sigma_2\left(\rho_{34}\sigma_3\sigma_4 + \mu_3\mu_4\right)$.

Set $\boldsymbol{\mu}_{\boldsymbol{X}} = [0, 0, 250, 400]$ and standard variance vector $\boldsymbol{\sigma} = [4, 2, 200, 300]$. There are 500 decision trees and 3000 samples to construct the RF model. Tab. 4 shows the VIMs results of group variables for the independent variable. The Pearson correlation coefficients are $\rho_{12} = 0.3$ and $\rho_{34} = -0.3$. Tab. 5 shows the importance measures of single variable in the case of correlated and independent variable space.

Table 4: The group variables VIMs of Example 5.3

	X_1X_2	X_1X_3	X_1X_4	X_2X_3	X_2X_4	X_3X_4
η_{ij}	1.931×10^6	1.975×10^6	3.206×10^6	3.905×10^6	3.207×10^6	5.171×10^6
$\eta_{ij} \Rightarrow S_{ij}$	0.000	0.242	0.002	0.004	0.137	0.008

Table 5: The single variable VIMs of Example 5.3

		η_i	$\eta_i \Rightarrow S_i$	$S_i^{(\text{Ana})}$	Error	η_i^T	$\eta_i^T \Rightarrow S_i^T$	$S_i^{T(\text{Ana})}$	Error
Independent case	X_1	3.205×10^6	0.380	0.379	0.26%	3.223×10^6	0.623	0.621	0.32%
	X_2	3.903×10^6	0.246	0.242	1.65%	1.977×10^6	0.382	0.379	0.79%
	X_3	5.199×10^6	0.004	0.000	–	1.225×10^6	0.237	0.242	2.07%
	X_4	5.188×10^6	0.002	0.000	–	7.063×10^5	0.137	0.136	0.74%
Correlated case	X_1	5.356×10^6	0.492	0.507	2.96%	1.835×10^6	0.490	0.492	0.41%
	X_2	2.473×10^6	0.403	0.399	1.00%	4.319×10^6	0.333	0.300	11.0%
	X_3	6.036×10^6	0.001	0.000	–	1.089×10^6	0.189	0.192	1.56%
	X_4	5.924×10^6	0.000	0.000	–	6.938×10^5	0.108	0.108	–

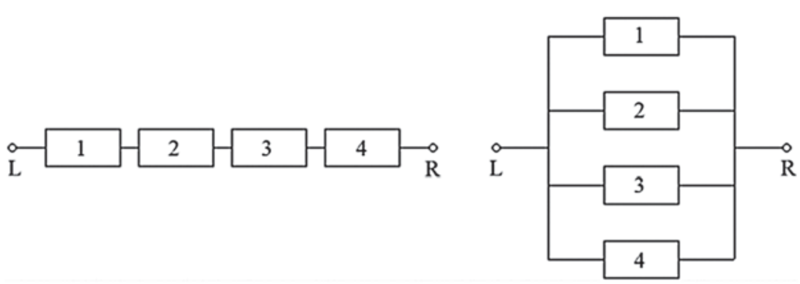
Tabs. 4 and 5 show that analytical values and numerical simulation of VIMs have good consistency. In independent variable space, the third and fourth order sensitivity indices are all equal to zero, so the relationship of important measures of single variable and group variables are also $S_i^T = S_i + \sum_{j \neq i} S_{ij}$.

5.4 Engineering Example 4: Series and Parallel Electronic Models

Since the reliability of an electronic instrument in design stages has attracted much attention. Two simple electronic circuit models from reference [31] are used to get the VIMs. The series and parallel structures (shown in Fig. 6) are all considered in the importance measures. Each of the electronic circuit models contains four elements. The lifetime T_i independently obeys exponential distribution. The failure rate parameters are $\boldsymbol{\lambda} = [1, 1/4.5, 1/9, 1/99]$, and the lifetime T of the models can be respectively expressed as:

Series model: $T = \min(T_1, T_2, T_3, T_4)$

Parallel model: $T = \max(T_1, T_2, T_3, T_4)$



(a) (b)

Figure 6: The series and parallel electronic circuit structures (a) Series model (b) Parallel model

Tabs. 6 and 7 show the computational results of the importance measures by RF model, there are 500 decision trees and 15000 samples in the RF model. Due to the electronic circuit structures are discontinuous, more samples are needed to acquire the precise surrogate model and the importance measures. Additionally, the MC simulation results with 6×2^{25} random samples are presented as approximate exact solutions $S_i^{(MC)}$, $S_i^{T(MC)}$ and $S_{ij}^{(MC)}$ for comparison. From the comparison, the RF importance measures are also appropriate for the discontinuous model. The main sensitivity indices are almost equal to the total indices in the parallel model, while they have a significant difference in the series model (seen from Tab. 6). The second-order indices of series model are not equal to zero (seen from Tab. 7), which causes the VIMs difference between parallel model and series model.

Table 6: The single variable VIMs of electronic models

		η_i	$\eta_i \Rightarrow S_i$	$S_i^{(MC)}$	η_i^T	$\eta_i^T \Rightarrow S_i^T$	$S_i^{T(MC)}$
Series model	T_1	0.429	0.607	0.593	0.942	0.864	0.853
	T_2	0.993	0.090	0.090	0.308	0.282	0.284
	T_3	1.048	0.039	0.043	0.158	0.145	0.153
	T_4	1.090	0.001	0.004	0.005	0.004	0.0149
Parallel model	T_1	1.929×10^4	0.000	0.000	0.000	0.000	0.000
	T_2	1.929×10^4	0.000	0.000	0.000	0.000	0.001
	T_3	1.929×10^4	0.000	0.000	1.929×10^4	0.001	0.001
	T_4	12.232	0.999	0.999	12.217	1.000	1.000

Table 7: The group variables VIMs of series model

	T_1T_2	T_1T_3	T_1T_4	T_2T_3	T_2T_4	T_3T_4
η_{ij}	0.835	0.705	0.602	0.142	0.095	0.047
$\eta_{ij} \Rightarrow S_{ij}$	0.152	0.069	0.006	0.008	0.001	0.000
$S_{ij}^{(MC)}$	0.156	0.069	0.003	0.006	0.003	0.000

5.5 Engineering Example 5: A Cantilever Tube Model

A cantilever tube model (shown in Fig. 7) is used to analyze the variable importance measures. The model is a nonlinear model with six random variables. The input variables are outer diameter d, thickness t, external forces F_1, F_2, P and torsion T, respectively.

The tensile stress σ_x and the torsion stress τ_{zx} can be analyzed:

$$\sigma_x = \frac{P + F_1 \sin\theta_1 + F_2 \sin\theta_2}{A} + \frac{M}{I}, \quad \tau_{zx} = \frac{Td}{4I}$$

where the sectional area A, the bending moment M and the inertia moment I can be calculated by the following formula:

$$A = \frac{\pi}{4}\left[d^2 - (d-2t)^2\right], \quad M = F_1 L_1 \cos\theta_1 + F_2 L_2 \cos\theta_2, \quad I = \frac{\pi}{64}\left[d^4 - (d-2t)^4\right].$$

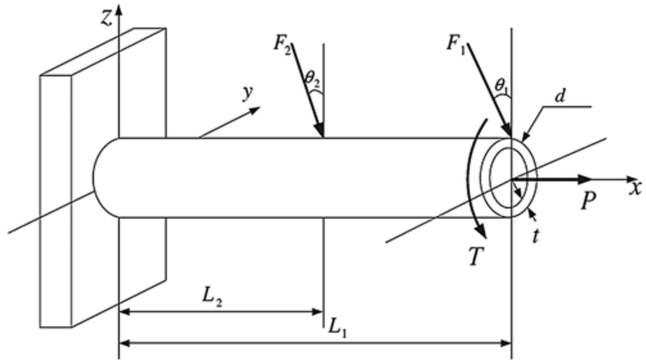


Figure 7: The cantilever tube model

And the maximum stress of the cantilever can be calculated as $\sigma_{\max} = \sqrt{\sigma_x^2 + 3\tau_{zx}^2}$. All input variables t, d, F_1, F_2, P and T are normally distributed with parameters shown in Tab. 8. The Pearson correlation coefficients are $\rho_{td} = 0.3$ and $\rho_{F_1F_2} = 0.5$. There are 500 decision trees and 7000 samples in the RF model. Tab. 9 gives the variable importance measures by RF method and the single-loop Monte Carlo simulation method. The cost of the MC method is 8×2^{23} points for each case.

Table 8: Distribution parameters of input variables

Variable/unit	Mean	Standard variance
t/mm	5	0.1
d/mm	42	0.5
F_1/N	3000	300
F_2/N	3000	300
P/N	12000	1200
T/N·mm	90000	9000

For the independent variables, the main and total sensitivity indices of input variables are very close (seen from Tab. 9), which suggests that the influence of these variables to the output response mainly come from unique variables and the interaction contribution is very small. The external force P is the most important variable in the independent space; the importance of the other input variables has a slight difference.

Furthermore, the importance measures are different in the correlated variable space. For the correlated input variables t, d, F_1 and F_2 the sensitivity indices $S_i > S_i^T$, the influence on the output response mainly originates from the correlated contribution by Pearson correlation coefficients. For the input variables P and T, they are independent with other variables, so the first order indices are almost equal to total sensitivity indices. Therefore, the proposed variable RF importance measure system not only reflects the important variables but also provides useful information to identify the structure of the engineering model, which will provide useful guidance for the engineering design and optimization.

Table 9: The VIMs of cantilever tube model

		t	d	F_1	F_2	P	T
Independent space	η_i	9.690	9.216	9.407	9.937	4.060	9.416
	$\eta_i \Rightarrow S_i$	0.061	0.107	0.089	0.037	0.607	0.088
	$S_i^{(MC)}$	0.060	0.112	0.086	0.038	0.615	0.088
	η_i^T	0.706	1.172	0.906	0.407	6.328	0.934
	$\eta_i^T \Rightarrow S_i^T$	0.068	0.114	0.088	0.039	0.613	0.091
	$S_i^{T(MC)}$	0.060	0.112	0.086	0.038	0.615	0.089
Correlated space	η_i	10.842	9.863	9.730	9.970	4.641	10.335
	$\eta_i \Rightarrow S_i$	0.054	0.140	0.165	0.107	0.590	0.090
	$S_i^{(MC)}$	0.057	0.133	0.151	0.110	0.593	0.085
	η_i^T	0.174	1.180	0.593	0.473	6.747	0.973
	$\eta_i^T \Rightarrow S_i^T$	0.008	0.094	0.064	0.021	0.592	0.086
	$S_i^{T(MC)}$	0.013	0.089	0.065	0.024	0.593	0.086

5.6 Engineering Example 6: Solar Wing Mast of Space Station

The solar wing mast of space station is a truss structure in 3D space based on triangular structure, shown in Fig. 8.

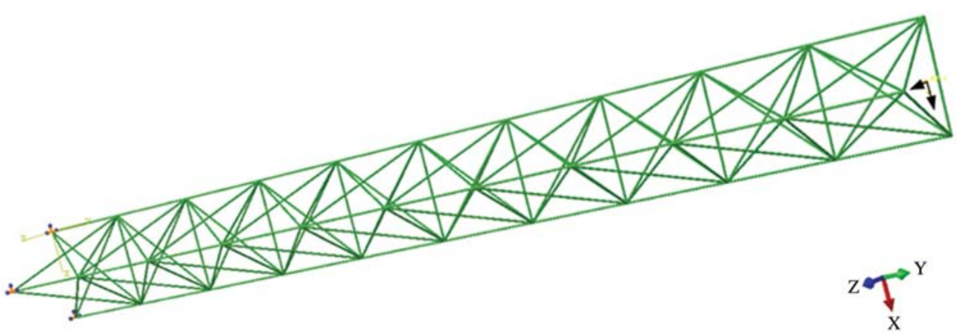


Figure 8: Solar wing mast structure [32]

The solar wing mast is made of titanium alloy. The material properties (including density ρ, Elastic modulus E, Poisson's ration ν), external load (including dynamic load F_1 and static load F_2) and sectional area of truss A are random variables, the corresponding distribution parameters are listed in Tab. 10.

Software CATIA is used to establish the geometry and finite element model, and then taking the maximum stress as the output response, ABAQUS was repeatedly called to analyze the finite element model. And finally 210 samples were obtained. Random forest is used to analyze the variable importance measures, the results of VIMs are listed in Tab. 11.

Table 10: Distribution parameters of input variables

Variable/unit	Mean	Standard variance
ρ/kg·m^{-3}	4300	215
E/GPa	106	5.3
ν	0.3	0.015
A/m^2	0.0001	5×10^{-6}
F_1/N	100	5
F_2/N	100	10

Table 11: The VIMs of solar wing mast

Variable	η_i	$\eta_i \Rightarrow S_i$	η_i^T	$\eta_i^T \Rightarrow S_i^T$
ρ	3.144×10^{12}	0.0106	2.434×10^{12}	0.7586
E	3.133×10^{12}	0.0138	2.454×10^{12}	0.7647
ν	3.179×10^{12}	0.0000	2.692×10^{11}	0.0860
A	2.754×10^{12}	0.1379	1.096×10^{12}	0.3576
F_1	3.161×10^{12}	0.0060	3.225×10^{11}	0.0994
F_2	3.089×10^{12}	0.0309	3.857×10^{11}	0.1301

According to the results of variable importance measures, the main sensitivity index of Poisson's ration ν is almost zero, and the total sensitivity index is also the minimum one. In order to simplify the model, the Poisson's ration ν can be considered as a constant. The sectional area of truss A is the key design variable, since A has the largest main sensitivity to output. There is a large interaction between density ρ and Elastic modulus E, and the interaction sensitivity index can be indirectly solved $S_{\rho E} \approx 0.4623$. For external load, F_1 and F_2 can be regarded as secondary variables. The variable importance measures can give designer reasonable suggestions to allocate optimization spaces of design variables more effectively and reduce the optimization dimension.

6 Conclusions

The Kriging regression model is used as the leaf node model of decision tree to improve the prediction accuracy of RF. The single variable, group variables and correlated variables importance measures based on RF are presented, which constitute the complete RF variable importance measure system. Additionally, a novel approach for solving variance-based global sensitivity indices is presented, and the novel meaning of these VIM indices is also introduced. The results of the numerical and engineering examples testify that the VIM indices of RF can further derive the variance sensitivity indices with higher computational efficiency compared with single-loop MC simulation.

For some incomplete probability information, such as linear correlated non-normal variables, non-linear correlated variables and discrete input-output samples and so on, the proposed importance measure analysis method has some limitations in applicability. In future work, the importance measures under incomplete probability information will be studied based on equivalent transformation or Copula function.

Authors' Contributions: Conceptualization and methodology by Song, S. F., validation and writing by He, R. Y., examples and computation by Shi, Z. Y., examples and writing by Zhang, W. Y.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Lu, Z. Z., Li, L. Y., Song, S. F., Hao, W. R. (2015). *Theory and solution of importance analysis for uncertain structural systems,* pp. 1–5. Beijing: Science Press (in Chinese).
2. Borgonovo, E. (2007). A new uncertainty importance measure. *Reliability Engineering & System Safety, 92(6),* 771–784. DOI 10.1016/j.ress.2006.04.015.
3. Liu, Q., Homma, T. (2009). A new computational method of a moment-independent uncertainty importance measure. *Reliability Engineering & System Safety, 94(7),* 1205–1211. DOI 10.1016/j.ress.2008.10.005.
4. Cui, L. J., Lu, Z. Z., Zhao, X. P. (2010). Moment-independent importance measure of basic random variable and its probability density evolution solution. *Science China Technological Sciences, 53(4),* 1138–1145. DOI 10.1007/s11431-009-0386-8.
5. Saltelli, A., Annon, P., Auini, I. (2010). Variance based sensitivity analysis of model output: Design and estimator for the sensitivity indices. *Computer Physics Communications, 181(2),* 259–270. DOI 10.1016/j.cpc.2009.09.018.
6. Ziehn, T., Tomlin, A. S. (2008). A global sensitivity study of sulphur chemistry in a premixed methane flame model using HDMR. *International Journal of Chemical Kinetics, 40(11),* 742–753. DOI 10.1002/kin.20367.
7. Ratto, M., Pagano, A., Young, P. C. (2007). State dependent parameter meta-modeling and sensitivity analysis. *Computer Physics Communications, 177(11),* 863–876. DOI 10.1016/j.cpc.2007.07.011.
8. Breiman, L. (2001). Random forest. *Machine Learning, 45(1),* 5–32. DOI 10.1023/A:1010933404324.
9. Wang, J. H., Yan, W. Z., Wan, Z. J., Wang, Y., Lv, J. K. et al. (2020). Prediction of permeability using random forest and genetic algorithm model. *Computer Modeling in Engineering & Sciences, 125(3),* 1135–1157. DOI 10.32604/cmes.2020.014313.
10. Yu, B., Chen, F., Chen, H. Y. (2019). NPP estimation using random forest and impact feature variable importance analysis. *Journal of Spatial Science, 64(1),* 173–192. DOI 10.1080/14498596.2017.1367331.
11. Hallett, M. J., Fan, J. J., Su, X. G., Levine, R. A., Nunn, M. E. (2014). Random forest and variable importance rankings for correlated survival data, with applications to tooth loss. *Statistical Modelling, 14(6),* 523–547. DOI 10.1177/1471082X14535517.
12. Cutler, A., Cutler, D. R., Stevens, J. R. (2011). Random forests. *Machine Learning, 45(1),* 157–176. DOI 10.1007/978-1-4419-9326-7_5.
13. Loecher, M. (2020). From unbiased MDI feature importance to explainable AI for trees. https://www.researchgate.net/publication/340224035.
14. Mitchell, M. W. (2011). Bias of the random forest out-of-bag (OOB) error for certain input parameters. *Open Journal of Statistics, 1(3),* 205–211. DOI 10.4236/ojs.2011.13024.
15. Bénard, C., Veiga, S. D., Scornet, E. (2021). MDA for random forests: inconsistency and a practical solution via the Sobol-MDA. http://www.researchgate.net/publication/349682846.
16. Zhang, X. M., Wada, T., Fujiwara, K., Kano, M. (2020). Regression and independence based variable importance measure. *Computers & Chemical Engineering, 135(6),* 106757. DOI 10.1016/j.compchemeng.2020.106757.
17. Fisher, A., Rudin, C., Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire cass of prediction models simultaneously. *Journal of Machine Learning Research, 20(177),* 1–81. http://www.jmlr.org/papers/v20/18-760.html.
18. Song, S. F., He, R. Y. (2021). Importance measure index system based on random forest. *Journal of National University of Defense Technology, 43(2),* 25–32. DOI 10.11887/j.cn.202102004 (in Chinese).

19. Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Mathematics and Computers in Simulation, 55(1),* 271–280. DOI 10.1016/S0378-4754(00)00270-6.
20. Saltelli, A., Tarantola, S. (2002). On the relative importance of input factors in mathematical models: Safety assessment for nuclear waste disposal. *Journal of the American Statistical Association, 97(459),* 702–709. DOI 10.1198/016214502388618447.
21. Saltelli, A. (2002). Sensitivity analysis for importance assessment. *Risk Analysis, 22(3),* 579–590. DOI 10.1111/0272-4332.00040.
22. Abdulkareem, N. M., Abdulazeez, A. M. (2021). Machine learning classification based on radom forest algorithm: A review. *International Journal of Science and Business, 5(2),* 128–142. DOI 10.5281/zenodo.4471118.
23. Athey, S., Tibshirani, J., Wager, S. (2019). Generalized random forests. *The Annals of Statistics, 47(2),* 1179–1203. DOI 10.1214/18-AOS1709.
24. Badih, G., Pierre, M., Laurent, B. (2019). Assessing variable importance in clustering: A new method based on unsupervised binary decision trees. *Computational Statistics, 34(1),* 301–321. DOI 10.1007/s00180-018-0857-0.
25. Behnamian, A., Banks, S., White, L., Millard, K., Pouliot, D. et al. (2019). Dimensionality deduction in the presence of highly correlated variables for Random forests: Wetland case study. *IGARSS 2019–2019 IEEE International Geosciences and Remote Sensing Symposium*, pp. 9839–9842, Yokohama, Japan.
26. Gazzola, G., Jeong, M. K. (2019). Dependence-biased clustering for variable selection with random forests. *Pattern Recognition, 96,* 106980. DOI 10.1016/j.patcog.2019.106980.
27. Mara, T. A., Tarantola, S. (2012). Variance-based sensitivity indices for models with dependent inputs. *Reliability Engineering & System Safety, 107(11),* 115–121. DOI 10.1016/j.ress.2011.08.008.
28. Kucherenko, S., Tarantola, S., Annoni, P. (2012). Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications, 183(4),* 937–946. DOI 10.1016/j.cpc.2011.12.020.
29. Li, L. Y., Lu, Z. Z. (2013). Importance analysis for models with correlated variables and its sparse grid solution. *Reliability Engineering & System Safety, 119,* 207–217. DOI 10.1016/j.ress.2013.06.036.
30. He, X. Q. (2008). *Multivariate statistical analysis,* pp. 9–14. Beijing: Renmin University Press (in Chinese).
31. Song, S. F., Wang, L. (2017). Modified GMDH-NN algorithm and its application for global sensitivity analysis. *Journal of Computational Physics, 348(1),* 534–548. DOI 10.1016/j.jcp.2017.07.027.
32. He, R. Y. (2020). *Variable importance measures based on surrogate model,* pp. 66–69. Xi'an: Northwestern Polytechnical University (in Chinese).