

Article

# Influence Analysis and Prediction of ESDD and NSDD Based on Random Forests

Ang Ren <sup>1,2</sup> , Qingquan Li <sup>1,2,\*</sup> and Huaishuo Xiao <sup>1,2</sup>

<sup>1</sup> Department of Electrical Engineering, Shandong University, Jinan 250061, China; renang@mail.sdu.edu.cn (A.R.); 201513014@mail.sdu.edu.cn (H.X.)

<sup>2</sup> Shandong Provincial Key Laboratory of Ultra High Voltage Transmission Technology and Equipments, #17923 Jingshi Road, Jinan 250061, China

\* Correspondence: lqq@sdu.edu.cn; Tel.: +86-53181696115

Received: 4 April 2017; Accepted: 26 June 2017; Published: 30 June 2017

**Abstract:** Equivalent salt deposit density (ESDD) and non-soluble deposit density (NSDD) measurements are a basic requirement of power systems. In order to predict the site pollution severity (SPS) of insulators, a new method based on random forests (RFs) is proposed. Using mutual information (MI) theory and RFs, the weights of factors related to the SPS of insulators are analyzed. The samples of contaminated insulators are extracted from the transmission lines of high voltage alternating current (HVAC) and high voltage direct current transmission (HVDC). The regression models of RFs and support vector machines (SVM) are constructed and compared, which helps to support the lack of information in predicting NSDD in previous works. The results are as follows: according to the mean decrease accuracy (MDA), mean decrease Gini, (MDG), and MI, the types of the insulators (including surface area, surface orientation, and total length) as well as the hydrophobicity are the main factors affecting both ESDD and NSDD. Compared with NSDD, the electrical parameters have a significant effect on ESDD. For the influence factors of ESDD, the weights of the insulator type, hydrophobicity, and meteorological factors are 52.94%, 6.35%, and 21.88%, respectively. For the influence factors of NSDD, the weights of the insulator type, hydrophobicity, and meteorological factors are 55.37%, 11.04%, and 14.26%, respectively. The influence voltage level ( $vl$ ), voltage type ( $vt$ ), polarity/phases ( $pp$ ) exerted on ESDD are 1.5 times, 3 times, and 4.5 times of NSDD, respectively. The influence that distance from the coastline ( $d$ ), wind velocity ( $wv$ ), and rainfall ( $rf$ ) exert on NSDD are 1.5 times, 2 times, and 2.5 times that of ESDD, respectively. Compared with the natural contamination test and the SVM regression model, the RFs regression model can effectively predict the contamination degree of insulators, and the relative error of the predicted ESDD and NSDD is 8.31% and 9.62%, respectively.

**Keywords:** insulators; ESDD; NSDD; random forests; mutual information

## 1. Introduction

Research on insulator natural contamination is a basic requirement of external insulation. The contamination degree of insulators is a result of the actual operating environment, which can reflect the pollution resistance characteristics of insulators under natural conditions [1]. Due to the complex working environment of insulators, the natural contamination characteristics of the insulator are difficult to depict with mathematical expressions.

At present, machine learning algorithms are widely used in the study of the natural contamination characteristics of insulators: Jiao et al. combined particle swarm optimization (PSO) and support vector machine (SVM) to build an insulator contamination on-line monitoring system to predict the contamination degree, and the relative error was less than 10% [2]. Ahmad et al. modeled the relationship between equivalent salt deposit density (ESDD) with temperature, humidity, pressure,

rainfall, and wind velocity, using artificial neural networks (ANN), and the mean absolute error of the model output was found to be 3.6% [3]. Meanwhile, Karamousantas et al. provides the foundation for routine maintenance of insulators according to the same algorithm, and the relative error was less than 17% [4]. Muniraj et al. used the characteristics of leakage current as an input, and predicted ESDD by an adaptive neurofuzzy inference system (ANFIS) model, whose coefficient of determination was 0.998 and root mean square error was just 0.00323 [5]. In recent years, the random forest algorithm has been widely used in the field of power systems. Hannan et al. proposed that the performance of the RFs based space vector pulse width modulation (SVPWM) technique is superior to both ANN-SVM and ANFIS-SVM techniques in terms of damping capability, settling time, steady-state error, and transient response under different operating speeds and load conditions [6]. Samantaray et al. proposed the identification of the fault zone in flexible alternating current transmission systems (FACTS)-based transmission lines using RFs (random forests) with an accuracy and reliability of more than 99% [7]. Shah et al. proposed a RFs-based fault discrimination technique for power transformers, and the fault discrimination accuracy was more than 98% [8]. Meanwhile, Kannan et al. proposed a RFs classifier, and the identification rate of the RF classifier lies above 98% at all pollution conditions [9]. Larivière and Van den Poel found that both random forests and regression forests techniques provide a better fit for the estimation and validation sample compared to ordinary linear regression and logistic regression models, and the prediction accuracy was 95% [10]. Therefore, the study of random forest algorithms is essential in the field of external insulation. However, the following shortcomings exist in the studies of SPS (site pollution severity) prediction: Firstly, the influencing factors are different and there is no weight analysis of the factors related to the natural contamination. Secondly, only ESDD is predicted in all of the above algorithms; NSDD is ignored, and according to IEC60815, NSDD is also an indispensable factor [11]. Lastly, RFs based natural pollution prediction has not been compared with other algorithms, and there is no actual natural contamination test validation.

In this paper, 16 factors that are related to ESDD and NSDD are proposed, and the weights are analyzed using MI (mutual information) and RFs. Based on the most related factors, a new regression method of using RFs as a function estimation is established. Experiments show that the method is greatly superior to the SVM regression model in terms of accuracy. The prediction results of this method have been verified by the actual data from the insulators of the Chinese East Coast transmission line.

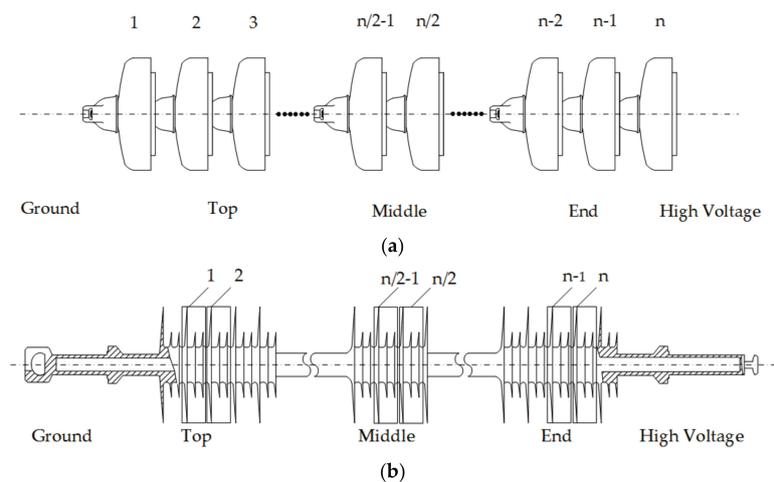
## 2. Onsite Measurements and Research Methodology

### 2.1. Insulator Parameters

In order to study the influence of voltage type and voltage level on the natural accumulation of composite insulators, the surrounding insulators of HVAC transmission lines are sampled. The voltage level of the transmission line and the parameters of the insulators are shown in Table 1. As the contamination degree of the insulators is not uniform [12], the parameters of each insulator sample consists of four values: ESDD of the upper surface, NSDD of the upper surface, ESDD of the lower surface, and NSDD of the lower surface. Since the insulator string of the transmission line is composed of a plurality of insulators, the potential of each insulator is different. So each string of insulators is numbered from 1 to n according to its potential. The sampling diagrams of suspension insulators and the long rod composite insulator are shown in Figure 1.

**Table 1.** Insulator parameters.

Insulator Type	Voltage Level (kV)	Leakage Distance (mm)	Upper Surface Areas (cm <sup>2</sup> )	Lower Surface Areas (cm <sup>2</sup> )
FC160P/C170DC	±660	550 (a piece)	1800	2700
FXBZW ±660/300	±660	9220	224.6/120.2	224/120
FXBZW ±660/160	±660	4680	224.6/120.2	224/120
LXHY4-100	220	450 (a piece)	975	1601
XWP2-160	500	450 (a piece)	1551	1208
FC160P	500	550 (a piece)	1198	2541



**Figure 1.** Insulator sampling diagram. (a) Suspension insulators; (b) long rod composite insulator.

The natural contamination data presented in this paper were sampled in the transmission line of East China. In order to reduce the measurement error, the insulators were removed from the ground, using non-woven cloth (with a small amount of ethanol solution) to wipe off all the contamination. The method introduced in [11] was used to measure the ESDD and NSDD of each insulator.

## 2.2. The Influences of the Natural Contamination

Meteorological factors are evaluated using the average value of one-year, such as days with a wind velocity that is more than 5.5 m/s, days of medium to heavy rain, annual average temperature, and so on.

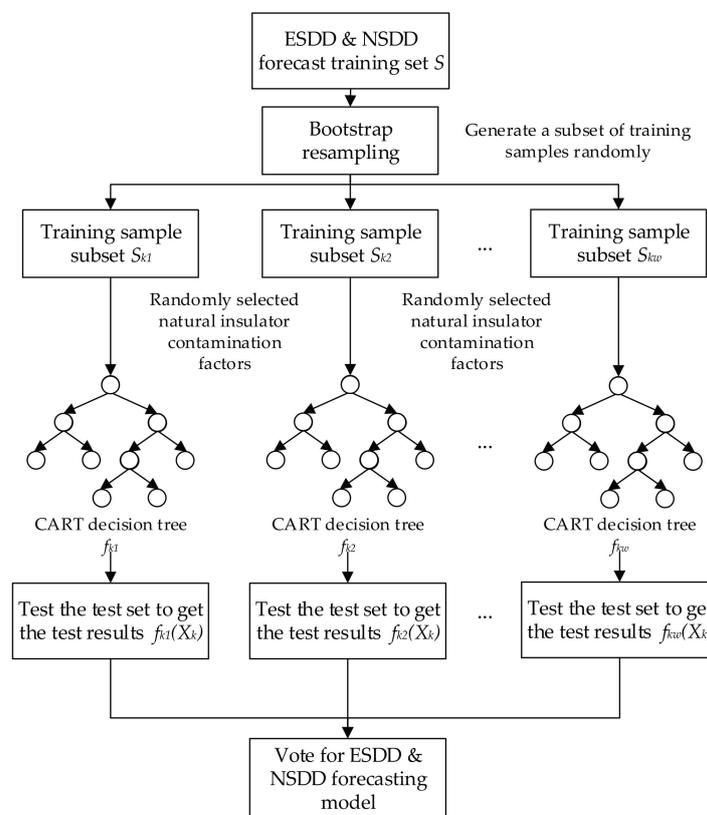
The SPS is determined by ESDD and NSDD under various factors. At present, the difficulty in the modelling of natural contamination is the selection criteria for the related factors. According to references [12–16] and the actual working condition, the following factors are chosen to evaluate the contamination of insulators: (1) Contamination factors: deposition time ( $dt$ ), hydrophobicity ( $HC$ ), particle size ( $ps$ ); (2) Meteorological factors: altitude ( $a$ ), distance from coastline ( $d$ ), rainfall ( $rf$ ), temperature ( $t$ ), and wind velocity ( $wv$ ); (3) Insulator type: material ( $m$ ), position factor ( $pf$ ), surface area ( $sa$ ), surface orientation ( $so$ ), total length ( $n$ ); (4) Electrical factors: voltage type ( $vt$ ), voltage level ( $vl$ ), polarity/phases ( $pp$ ). Considering the factors shown in Table 2 as inputs, RFs was used to calculate the weights.

**Table 2.** Insulator parameters.

Influencing Factors	Note	Symbols	Units
Contamination factors	deposition time	Insulator working time	$dt$ year
	particle size	Contamination particle size	$ps$ $\mu\text{m}$
	hydrophobicity	The hydrophobicity of the contaminated insulator	$HC$ 1-7
Meteorological factors	altitude	Altitude of the tower	$a$ m
	distance from coastline	Distance from the nearest coastline	$d$ m
	rain fall	Medium to heavy rain days	$rf$ day
	temperature	Annual average temperature	$t$ $^{\circ}\text{C}$
	wind velocity	wind speed $\geq 5.5$ m/s-days	$wv$ day
Insulator type	material	1 Coated with RTV; 2 Composite; 3 Porcelain	$m$ -
	position factor	$pf = i/n$ ( $i$ which is the $i$ -th insulator)	$pf$ -
	surface area	Insulator surface area	$sa$ $\text{cm}^2$
	surface orientation	Surface orientation (1 upper; 2 lower)	$so$ -
	total length	The number of chains	$n$ -
Electrical factors	polarity/phases	Polarity/Phases (1+; 2-; 3A; 4B; 5C)	$pp$ -
	voltage level	Insulator operating voltage	$vl$ kV
	voltage type	Voltage type (1 DC; 2 AC)	$vt$ -

### 2.3. Concept of RFs

RFs is a statistical learning algorithm, which adopts the bootstrap resampling method to extract multiple samples from the original sample. Firstly, the decision tree for each bootstrap sample is constructed. Then the predictions of multiple decision trees are combined by voting for the final prediction, and the results are obtained [17]. This method has high prediction accuracy, good tolerance for abnormal values and noise, and it is not easy to over-fit. On the basis of obtaining a variety of influencing factors, the ESDD and NSDD prediction model is carried out by using the random forests regression (RFs-R) model. The modelling process is shown in Figure 2. Firstly, a series of training samples are randomly selected from the original training sample set  $S_k$  ( $k = 1, \dots, n$ ) by the Bootstrap method for the insulator group  $G_k$  ( $k = 1, \dots, n$ ), then we use the test set to test the decision tree, synthesize the test results of multiple decision trees, and obtain the final ESDD and NSDD forecast model by voting.



**Figure 2.** Modelling process of ESDD and NSDD forecasting based on RFs.

There are two methods for calculating the weight of RFs variables: one is based on mean decrease accuracy (MDA); the other is based on Gini impurity, called the mean decrease Gini (MDG). The weight of the corresponding factor will increase with the decrement of the above two parameters [17].

#### 2.3.1. Mean Decrease in Gini

Suppose that  $S$  is a set of  $s$  data samples whose class label attribute has  $m$  different values and defines  $m$  different classes ( $C_i, i = 1, \dots, m$ ). According to the difference of the class label attribute values,

$S$  can be divided into  $m$  ( $S_i, i = 1, \dots, m$ ), and let  $S_i$  be the set of samples belonging to class  $C_i$  and  $S_i$  is the number of samples in set  $S_i$ . The Gini index of set  $S$  is:

$$\text{Gini}(S) = 1 - \sum_{i=1}^m p_i^2 \quad (1)$$

where,  $P_i$  estimated by  $s_i/s$  is the probability of any sample belonging to  $C_i$ . When  $\text{Gini}(S)$  is 0, all the records in the set belong to the same category. When all the samples in set  $S$  are evenly distributed,  $\text{Gini}(S)$  reaches its maximum, indicating that the minimum useful information can be obtained. If the data set  $S$  is divided into  $n$  subsets ( $S_j, j = 1, \dots, n$ ) according to a certain attribute partition, the index of  $\text{Gini}_{\text{split}}$  after splitting is:

$$\text{Gini}_{\text{split}}(S) = \sum_{j=1}^m \frac{s_j}{s} \text{Gini}(S_j) \quad (2)$$

where,  $n$  is the number of child nodes,  $s_j$  is the record number of the child node  $j$ , and  $s$  is the number of records at node  $P$ . When the classification method has traversed all of the attributes, the attribute whose  $\text{Gini}_{\text{split}}$  is the smallest will be selected as the split attribute of the node.

### 2.3.2. Mean Decrease in Accuracy

The order of the eigenvalues of each feature is disrupted and the impact of the order variation on the accuracy of the model is measured. Obviously, the order of disruption has little effect on the accuracy of the model for unimportant variables, but the order of the scramble reduces the accuracy of the model for important variables.

At first, train the RFs model to test the out of bag (OOB) error of each tree in the model using the sample data outside the bag. And then, randomly disrupt the value of the variable  $v$  in the sample data outside the bag and retest the OOB error of each tree. Finally, obtain the weight measure of the single tree to the variable  $v$ , which is the mean value of the difference of the OOB error in the test.

$$\text{MDA}(v) = \frac{1}{n} \sum_{i=1}^n (\text{errOOB}_i - \text{errOOB}'_i) \quad (3)$$

### 2.3.3. RFs-R Model

Using the RFs-R model written in MATLAB [18], and compared with the research of Larivière and Van den Poel on profit evolution estimation in the field of economics, RFs can be trained to predict the SPS of insulators. There are two major parameters of RFs, one is " $n_{\text{tree}}$ " representing the number of trees and the other is " $m_{\text{try}}$ " representing the number of variables used in determining the best split at each node. Leo Breiman [17] explained that the optimal value of  $m_{\text{try}}$  usually lies between  $\log_2 v$  and  $\sqrt{v}$  where  $v$  is the number of features. Larivière and Van den Poel had chosen a total of 30 dependent variables, and set the parameters of RFs-R as  $n_{\text{tree}} = 5000$ ,  $m_{\text{try}} = 6$ . In this paper, there are 16 dependent variables, and the  $m_{\text{try}}$  is set as 4.

The following are the indicators to evaluate the RFs regression model [19], and  $R^2$  is the fitness of the model. Mean percent standard error (MPSE) is an accurate indicator reflecting the fitting effect, where  $n$  represents the total number of insulators involved in the prediction,  $\hat{y}_i$  is the predicted output of the trees in the forest corresponding to the given input  $x_i$  sample, and  $y_i$  is the observed output. The parameter of  $n_{\text{tree}}$  can be set according to the minimum MPSE [17].

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

$$\text{MPSE} = \frac{1}{n} \sum_{i=1}^n \left| (y_i - \hat{y}_i)^2 / \hat{y}_i \right| \times 100\% \quad (5)$$

#### 2.4. Concept of Mutual Information

There are different correlations between the SPS of the insulators and the influencing factors, and the impacting degree of the same influencing factors on the SPS of different insulator strings is also different. Using the theory of mutual information (MI), the high correlation factors with ESDD and NSDD can be selected [20].

The information about the system  $X$  obtained after the known system  $Y$  can be represented by the difference between the unconditional entropy and the conditional entropy, which is defined as the mutual information of  $X$  and  $Y$ .  $X$  is the decision attribute and  $Y$  is the condition attribute.

$$I(X; Y) = H(X) - H(X|Y) \quad (6)$$

For discrete data,

$$H(X) = -\sum_{i=1}^n p(x_i) \log(p(x_i)) \quad (7)$$

$$H(X|Y) = -\sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) \log \left[ \frac{p(x_i y_j)}{p(y_j)} \right] \quad (8)$$

$$I(X; Y) = -\sum_{i=1}^n p(x_i) \log[p(x_i)] - \left\{ -\sum_{i=1}^n \sum_{j=1}^m p(x_i, y_j) \log \left[ \frac{p(x_i, y_j)}{p(y_j)} \right] \right\} \quad (9)$$

$$\sum_{i=1}^n \sum_{j=1}^m p(x_i y_j) = \sum_{j=1}^m p(y_j) \quad (10)$$

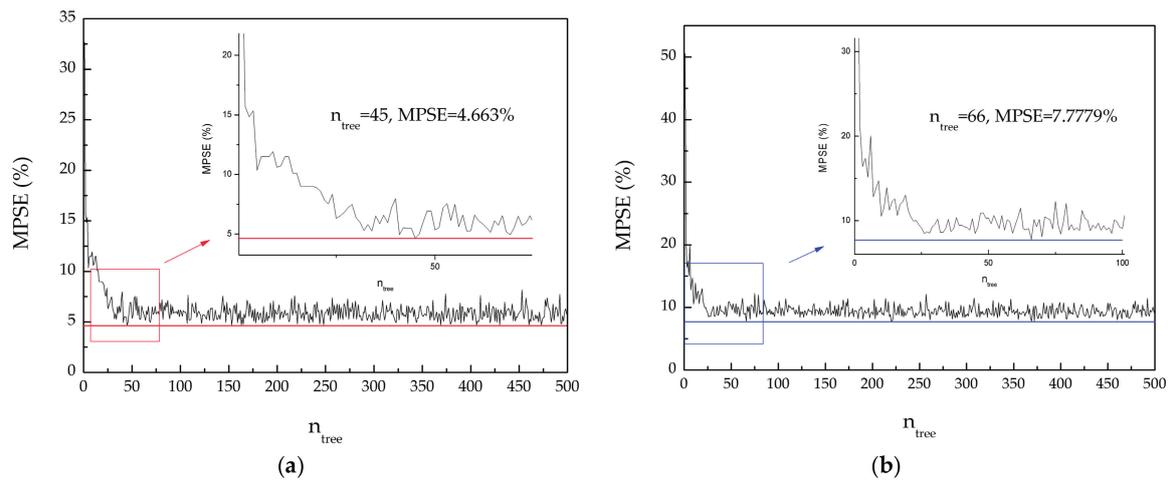
$$I(X; Y) = -\sum_{i=1}^n p(x_i) \log[p(x_i)] - \left\{ -\sum_{j=1}^m p(y_j) \sum_{i=1}^n p(x_i|y_j) \log[p(x_i|y_j)] \right\} \quad (11)$$

where  $p(x_i)$  is the probability of  $X = x_i$ ,  $p(y_j)$  is the probability of  $Y = y_j$ ;  $p(x_i|y_j)$  is the conditional probability of  $X = x_i$  when  $Y = y_j$ . In order to calculate the probabilities, the data of the samples should be discretized. The discretization of the samples can be used to convert the specific value into the interval represented by the probability, and although the discretization will lose the details, the result is more statistically significant. The specific method of discretizing the attribute fields of each condition attribute and decision attribute is to find the maximum and minimum values in each attribute domain respectively, and divide the distance between the maximum and minimum values into  $w$  intervals. Each value is placed in the corresponding interval, and the number of values in each interval is obtained.

### 3. Results

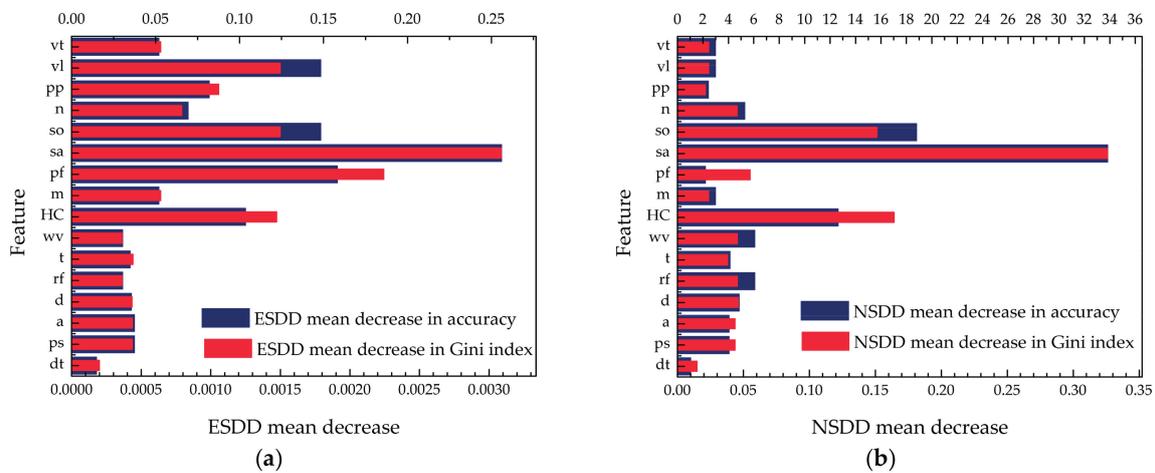
#### 3.1. The Weights of the Related Factors—RFs

Figure 3a,b shows the MPSE between the predicted and actual values based on the RFs when  $n_{tree}$  takes different values. Overall, the MPSE of the prediction model decreases as  $n_{tree}$  increases. However, when the value of  $n_{tree}$  is large, a too fine classification will lead to a rapid increase in the amount of calculations. Considering the modelling speed and the prediction error,  $n_{treeESDD} = 45$  and  $n_{treeNSDD} = 66$  are the best decision tree numbers.



**Figure 3.** Forecasting the OOB error of RFs with different  $n_{tree}$ . (a) The change in MPSE of ESDD when  $n_{tree}$  changes; (b) The change in MPSE of NSDD when  $n_{tree}$  changes.

The RFs overcome the shortcomings of traditional variable selection methods (choosing one or two variables in a group of variables that are equally highly correlated). As can be seen from Figure 4, there are gaps in the importance of different factors. The importance of the factors given by the RFs shows that the dominant factor is the insulator type, and the smallest influencing factor is the contamination deposition time. From the perspective of the insulators' maintenance data during spring, although the meteorological factors affect the natural pollution of the insulators, the effect is not obvious.



**Figure 4.** Importance analysis of influencing factors. (a) The results of ESDD; (b) The results of NSDD.

As can be seen from Figure 4a, the related factors that affect ESDD are: the insulator surface area ( $sa$ ), the position factor of the insulator ( $pf$ ), the insulator voltage level ( $vl$ ), the orientation of the surface ( $so$ ), the hydrophobicity after contamination ( $HC$ ), and the polarity/phase ( $pp$ ) of the line. Compared with ESDD, the factors that affect NSDD are mainly the insulator surface area ( $sa$ ), the orientation of the surface ( $so$ ), and the hydrophobicity after contamination ( $HC$ ), as can be seen from Figure 4b. The electrical factors of the insulators have less of an impact on NSDD.

### 3.2. The Importance of the Related Factors—Mutual Information

For the insulator string  $G_k$  ( $k = 1, \dots, n$ ), it is assumed that the ESDD and NSDD data sequence of the  $p$  insulator samples constitute the data set:

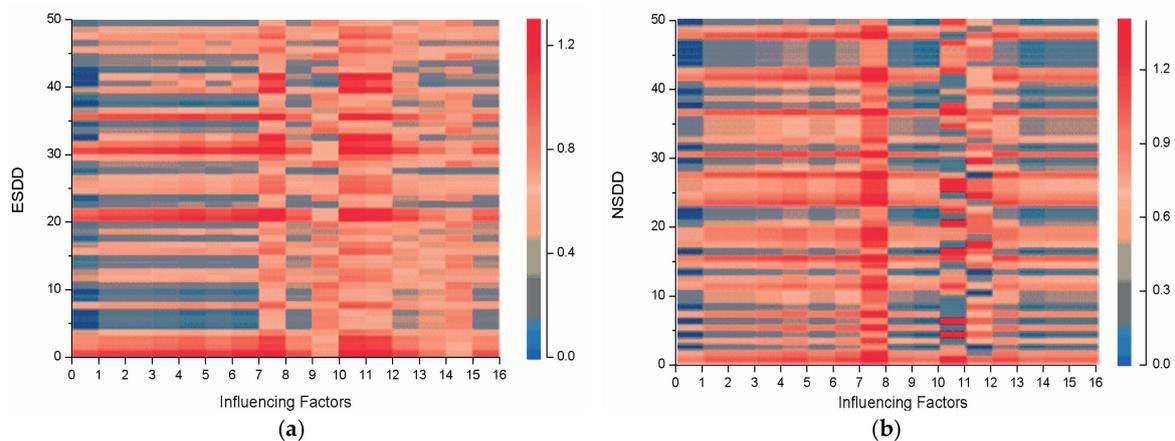
$$X_D = \begin{bmatrix} x_{\text{esdd},1}, x_{\text{esdd},2}, \dots, x_{\text{esdd},p} \\ x_{\text{nsdd},1}, x_{\text{nsdd},2}, \dots, x_{\text{nsdd},p} \end{bmatrix} \quad (12)$$

The data sequence of the  $l$  latent correlation factors constitutes the data set  $Y_D = \{Y_1, Y_2, \dots, Y_l\}$ , and the mutual information between the  $X_D$  and each correlation factor  $Y_D$  can be expressed as:

$$I_{\text{ESDD}} = \begin{bmatrix} I(x_{\text{esdd},1}, y_1) & \cdots & I(x_{\text{esdd},1}, y_j) & \cdots & I(x_{\text{esdd},1}, y_l) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ I(x_{\text{esdd},i}, y_1) & \cdots & I(x_{\text{esdd},i}, y_j) & \cdots & I(x_{\text{esdd},i}, y_l) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ I(x_{\text{esdd},p}, y_1) & \cdots & I(x_{\text{esdd},p}, y_j) & \cdots & I(x_{\text{esdd},p}, y_l) \end{bmatrix} \quad (13)$$

$$I_{\text{NSDD}} = \begin{bmatrix} I(x_{\text{nsdd},1}, y_1) & \cdots & I(x_{\text{nsdd},1}, y_j) & \cdots & I(x_{\text{nsdd},1}, y_l) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ I(x_{\text{nsdd},i}, y_1) & \cdots & I(x_{\text{nsdd},i}, y_j) & \cdots & I(x_{\text{nsdd},i}, y_l) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ I(x_{\text{nsdd},p}, y_1) & \cdots & I(x_{\text{nsdd},p}, y_j) & \cdots & I(x_{\text{nsdd},p}, y_l) \end{bmatrix} \quad (14)$$

where,  $x_{\text{esdd},l}, x_{\text{nsdd},l} \in X_D$ , and  $Y_j \in Y_D$ . Randomly selecting  $p = 50$  and  $j = 16$ , the thermal maps were drawn based on  $I_{\text{ESDD}}$  and  $I_{\text{NSDD}}$ , which are shown in Figure 5.



**Figure 5.** The thermal map. (a) ESDD; (b) NSDD.

In Figure 5, the horizontal axis represents the 16 related factors listed in Table 2, and the vertical axis represents the ESDD and NSDD, respectively. Each colour block represents the mutual trust size of the impacting factor. The greater the mutual information, the stronger the correlation is.

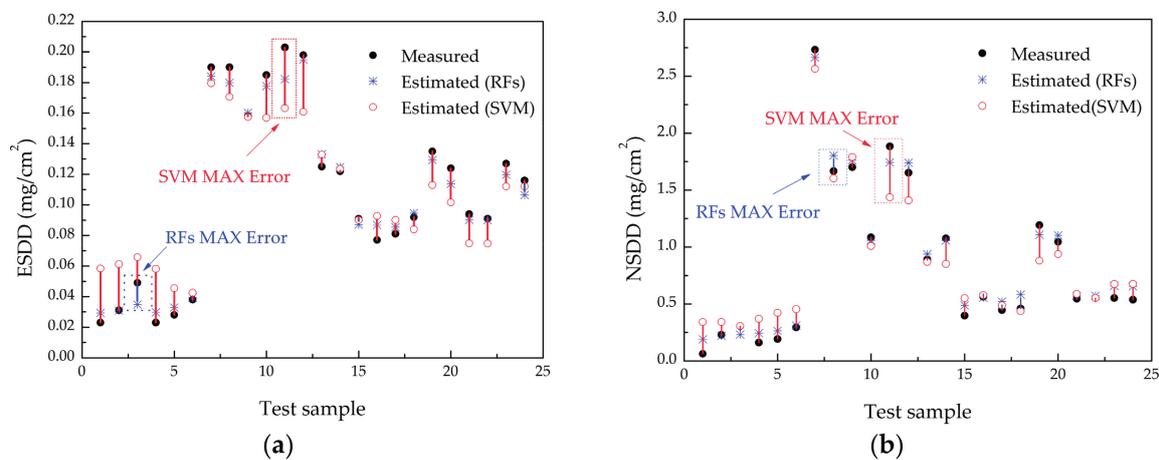
If only the mutual information between the SPS of a single insulator and the correlation factors is analysed, it can be found that there are individual differences in the colour distribution of each row. However, when the results from many insulators are integrated, the overall colour distribution captures the common characteristics of the correlation between SPS and each factor, and the strong correlation factors that affect the SPS of the insulators can be determined.

It can be seen from Figure 5a that the hydrophobicity ( $Y_3$ ), surface area ( $Y_{11}$ ), and surface orientation ( $Y_{12}$ ) have a strong correlation with ESDD. Material ( $Y_9$ ), total length ( $Y_{13}$ ), polarity/phase ( $Y_{14}$ ), and voltage level ( $Y_{15}$ ) have a significant impact on ESDD. It can be seen from Figure 5b that the influencing factors of hydrophobicity ( $Y_3$ ), surface area ( $Y_{11}$ ), surface orientation ( $Y_{12}$ ), and total length ( $Y_{13}$ ) have a strong correlation with NSDD.

### 3.3. Natural Contamination Tests

The regression models of RFs and SVM have been compared in this paper. Two-thirds (2/3) of the data are randomly selected as the RFs-R model training sample set, and we establish the support vector machines regression (SVM-R) model based on the same training sample set. Two regression models are tested with the 1/3 of the remaining data as the test sample set.

Due to the test samples being large, we randomly selected the ESDD and NSDD actual values of the two strings of insulators and compared them with the forecasted values of RFs-R and SVM-R. It can be seen from Figure 6 that the forecasted ESDD and NSDD trends are consistent with the actual ESDD and NSDD. Therefore, it is possible to use RFs-R and SVM for the prediction.



**Figure 6.** Contamination particle size. (a) ESDD; (b) NSDD.

It can be seen from Figure 6a that: using the RFs-R model, the results of the ESDD prediction are  $R^2 = 0.951$  and MPSE = 7.98%. Using the SVM-R model, the results of the ESDD prediction are  $R^2 = 0.86$  and MPSE = 17.65%. The maximum error of the RFs-R model is 31.7% of the maximum error of the SVM-R model. It can be seen from Figure 6b that: using the RFs-R model, the results of the NSDD prediction are  $R^2 = 0.911$  and MPSE = 9.04%. Using the SVM-R model, the results of the NSDD prediction are  $R^2 = 0.88$  and MPSE = 20.82%. The maximum error of the RFs-R model is 48.3% of the maximum error of the SVM-R model. The RFs-R model is feasible in ESDD and NSDD prediction through previous data learning and verification.

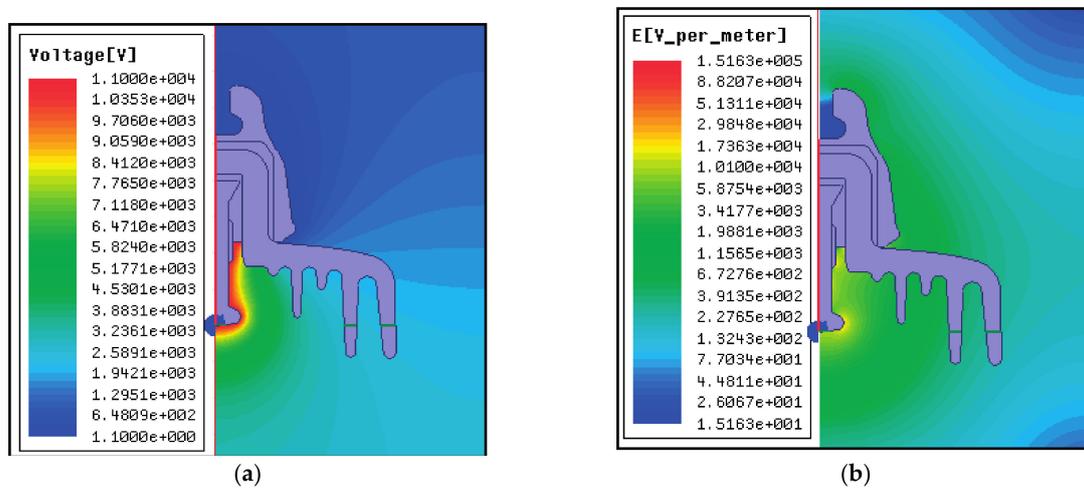
## 4. Discussion

In this chapter, the effects of electric field factors, meteorological factors, and contamination on the SPS of insulators are analysed. At the same time, the RFs classifier is used to quantitatively analyse the weight of the relevant factors.

### 4.1. Potential Distribution of the Insulators

Based on the finite element method, the electric field and the potential distribution of the insulators were simulated. Figure 7 shows the results of two-dimensional simulation of the suspension porcelain insulators, where Figure 7a,b is the voltage distribution and the electric field distribution of the

insulator, respectively. Since there are 60 insulators on the  $\pm 660$  kV DC transmission line, the voltage drop of each insulator is about 11 kV. In the simulation, the excitation voltage of the steel pin is set as 11 kV, and the excitation voltage of the steel cap is set as 0 kV.



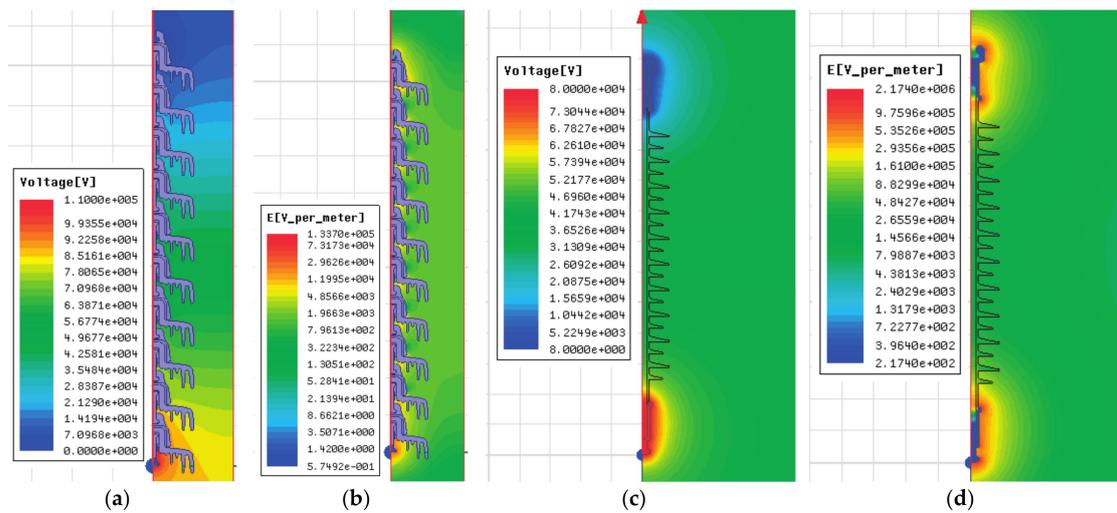
**Figure 7.** Two-dimensional simulation results of the suspension porcelain insulator. (a) The potential distribution of the suspension insulator; (b) The electric field of the suspension insulator.

According to Figure 7, the upper and lower surface of the insulator potential and electric field distribution is significantly different, corresponding to the actual contamination of the suspension insulator as shown in Figure 8, which reflects the importance of *so* (surface orientation).



**Figure 8.** The contaminated insulator. (a) The upper surface; (b) The lower surface.

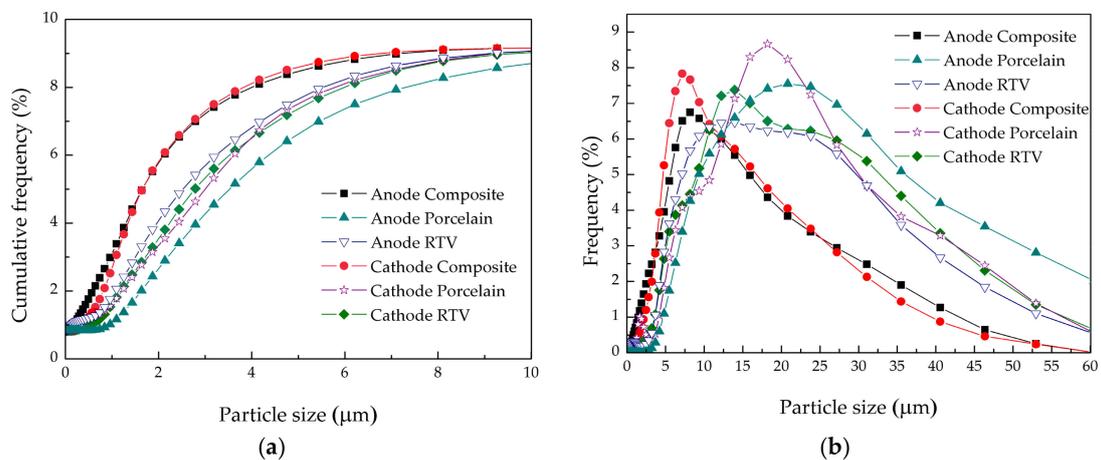
Figure 9 shows the two-dimensional electric field and potential distribution simulation of the suspension insulator string and composite insulators. It can be seen from Figure 9b,d that the electric field strength at both ends of the insulator is high, and the field strength decreases first and then increases along the direction of the insulator string. From Figure 9a,c we can see that the potential difference between the two ends of the insulator is very large, and the potential difference in the middle part of the string is very small. Therefore, the total length (*n*) and the position factor (*pf*) of the insulator string will have a large effect on the electric field strength and potential of the insulator, which indirectly affects ESDD.



**Figure 9.** Insulator electric field and potential distribution. (a) The potential distribution of the suspension insulators; (b) The electric field of the suspension insulators; (c) The potential distribution of the composite insulators; (d) The electric field of the composite insulators.

4.2. Contamination Analysis

From the cumulative frequency point of view, as shown in Figure 10a, the D90 of the composite insulators at the anode is 23.26 and 21.90 at the cathode; the D90 of the porcelain insulator is 41.54 and 33.33 at the cathode; the D90 of the coated room temperature vulcanized silicone rubber (RTV) insulator is 32.62 and 30.99 at cathode, and 90% of the contamination particle size conforms to the following rules: composite insulators < insulators coated with RTV < porcelain insulators.



**Figure 10.** Contamination particle size. (a) The cumulative frequency of the particle size; (b) The frequency of the particle size.

The composite insulator has a stronger ability to adsorb small particles than the porcelain insulators, and the insulator material has a great influence on the particle size of the particles. The insulator string under the cathode more easily adsorbs small particles than the insulator string under the anode. The D90 of the composite insulators at the cathode is lower than that at the anode by 5.85%. The D90 of the porcelain insulators at the cathode are lower than that at the anode by 19.76%. The D90 of the insulators coated with RTV at the cathode is lower than that at the anode by 5.00%, as shown in Figure 10a.

From the frequency point of view, as shown in Figure 10b, the maximum particle size of the composite insulators at the cathode is 7.83  $\mu\text{m}$  and 8.77  $\mu\text{m}$  at anode; the maximum particle size of the RTV-coated insulators at the cathode is 14.2  $\mu\text{m}$  and 12.3  $\mu\text{m}$  at the anode; the maximum particle size of the porcelain insulators at the cathode is 18.1  $\mu\text{m}$  and 22.3  $\mu\text{m}$  at the anode. The maximum particle size of contamination conforms to the same rules as the cumulative frequency, composite insulators < insulators coated with RTV < porcelain insulators. The effect of contamination on the organic materials is greater than that of the inorganic materials, and the particle size of contamination is smaller.

Table 3 shows the particle size distribution of the contamination, where D10 refers to the particle size corresponding to the cumulative probability distribution number of 10%, and P < 3 refers to the cumulative probability of a particle size less than 3  $\mu\text{m}$ .

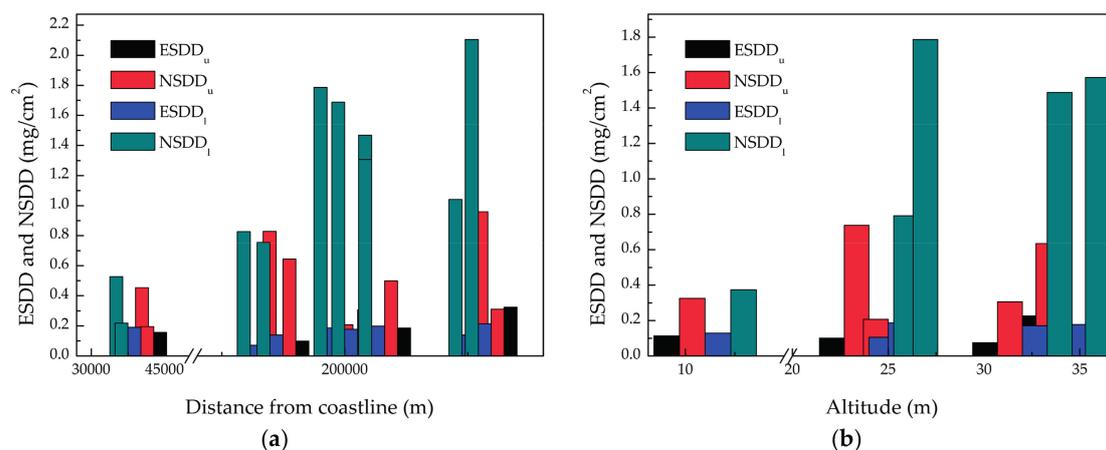
**Table 3.** The particle size distribution of the contamination.

Species	D10 ( $\mu\text{m}$ )	D50 ( $\mu\text{m}$ )	D90 ( $\mu\text{m}$ )	P < 3	P < 10	P < 20	P < 40
Anode composite	2.29	8.24	23.26	14.44%	59.42%	86.05%	98.99%
Anode porcelain	7.19	17.60	41.54	0.15%	21.90%	57.22%	88.91%
Anode RTV	5.16	15.00	32.62	5.07%	30.12%	68.37%	95.30%
Cathode composite	3.51	8.25	21.90	7.15%	59.86%	87.46%	99.23%
Cathode porcelain	5.13	13.89	33.33	2.50%	32.42%	68.13%	94.99%
Cathode RTV	4.59	12.59	30.99	4.94%	38.92%	71.88%	96.06%

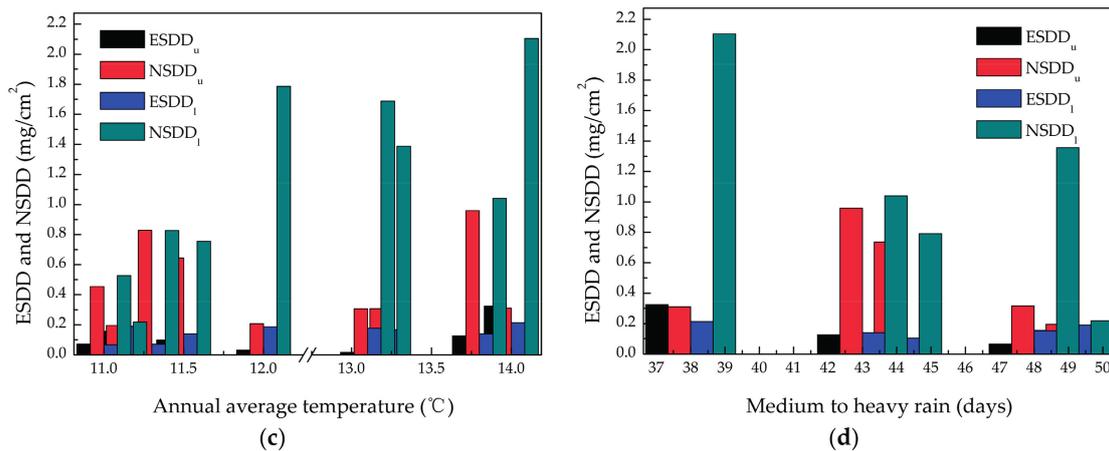
It can be seen from Table 3 and Figure 10 that the particle size distribution of the contamination of the composite insulator is concentrated in the range of small particles below 10  $\mu\text{m}$ , while the other two kinds of insulator particles have a frequency distribution of more than 10  $\mu\text{m}$ . The particle size distribution of the contamination is significantly affected by the insulator type and the suspension insulators have a wider range of particle size distribution than the rod insulators. From the above analysis, the insulator polarity, type, and material affect the contamination particle size distribution, but not obviously.

#### 4.3. Meteorological Factors

The following figures show the impact of related factors on ESDD or NSDD. The trend of the SPS of the insulators is qualitatively analysed according to Figure 11. The ranges of the related factors are shown in Table 4.



**Figure 11.** Cont.



**Figure 11.** Effects of the influencing factors on ESDD and NSDD. (a) Distance from coastline; (b) Altitude; (c) Annual average temperature; (d) Medium to heavy rain.

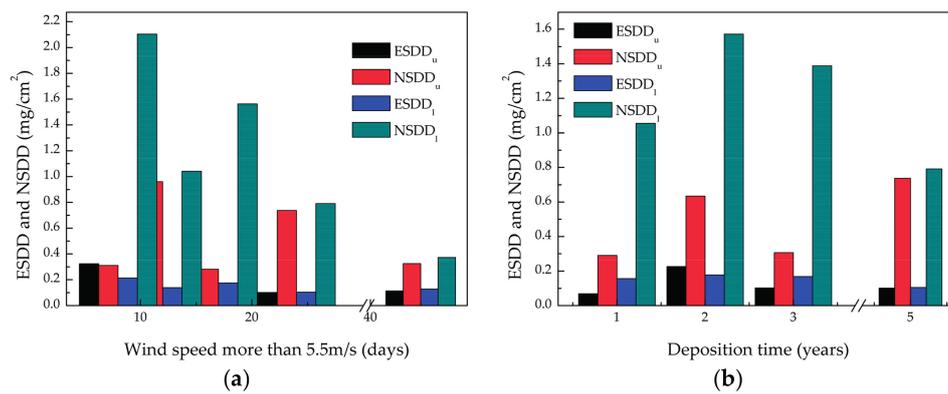
**Table 4.** The particle size distribution of the contamination.

Factors	$d$ (m)	$a$ (m)	$t$ (°C)	$rf$ (days)	$wv$ (days)	$dt$ (years)
Range	40,127–229,735	11–34	11–14	11–14	8–45	1–5

From Figure 11a we can see that the distance factor ( $d$ ) has a significant influence on the NSDD, and little influence on the ESDD. With the increase of the distance factor ( $d$ ), the NSDD of the lower surface shows an increasing trend. The NSDD within 5 km of the coastline is significantly lower than the NSDD in the inland area. The maximum values of ESDD and NSDD in all samples appear on the lower surface of the insulator string farthest from the coastline. Compared with Figure 11a–c, since the distance factor ( $d$ ) has a linear relationship with altitude ( $a$ ) and temperature ( $t$ ), the effect on SPS is basically the same.

As can be seen from Figure 11d, the larger the annual average rainfall ( $rf$ ) is, the greater the impact on ESDD and NSDD is. The ESDD and NSDD in the region of less rainfall are generally higher than that in the heavy rainfall area. However, with the increase of  $rf$ , the ESDD value of the upper surface gradually decreases, indicating that the rainfall plays a role in scouring the upper surface of the insulator. However with the increase of  $rf$ , the ESDD of the lower surface shows a tendency to decrease first and then increase. We believe that the region of  $rf \leq 40$  days is mainly concentrated in the inland areas, and the effect of scouring is obvious, leading to a downward trend of ESDD. With the increase of  $rf$ , the corresponding samples are concentrated in the coastal areas. The residual electrolyte increases on the surface because of the rain. Since it is difficult to scour the lower surface of the insulators by rain, the electrolyte gradually increases, resulting in the lower surface salt density rise.

It can be seen from Figure 12a that the wind has an obvious effect on the ESDD. With the increase in  $wv$ , ESDD showed a decreasing trend. The relationship between NSDD of the upper surface with the wind is complicated. We believe that the contamination can be easily carried by wind, causing uneven accumulation of contamination. As can be seen from Figure 12b, the SPS tends to increase first and then decrease with the increase of  $dt$ . Based on the statistical results of the insulator, the first saturated time of SPS is in the second year.



**Figure 12.** Effects of influencing factors on ESDD and NSDD. (a) Wind speed  $\geq 5.5$  m/s-days; (b) Deposition time.

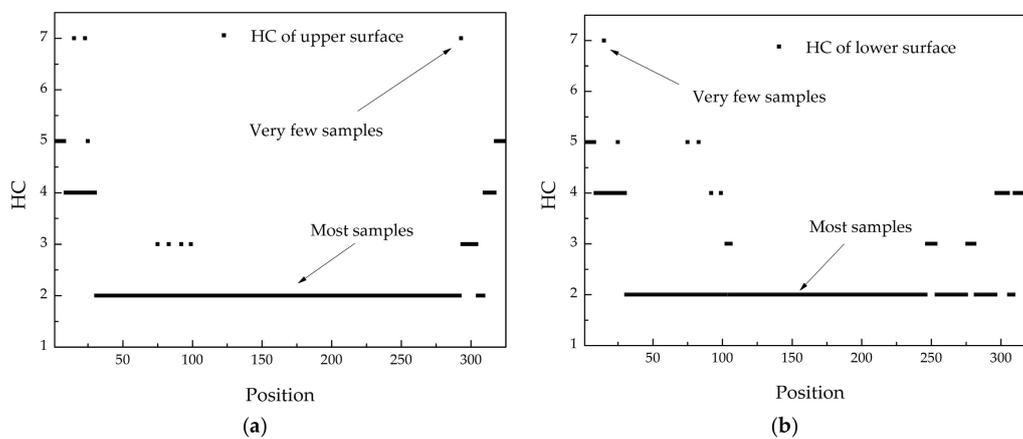
#### 4.4. The Hydrophobicity of the Insulators

Choosing one suspension insulator and one composite insulator as samples, the hydrophobicity classification (*HC*) is measured based on the method of reference [21].

Figure 13 shows the hydrophobicity of the composite insulators. Regardless of the upper surface or the lower surface, the *HC* results of the composite insulators are mostly HC2 and HC3, as can be seen in Figure 14a,b, indicating that the composite insulators have good hydrophobicity. On the whole, the poorly hydrophobic points appear near the ends of the composite insulator, which corresponds with the distribution of the high electric field intensity in Figure 9d.



**Figure 13.** The hydrophobicity of the composite insulators. (a) The upper surface; (b) the lower surface.



**Figure 14.** The hydrophobicity of the composite insulator. (a) The upper surface; (b) the lower surface.

It can be seen in Figure 15a that the upper surface has good hydrophobicity and is concentrated in HC3–HC4. The lower surface of the suspension insulator shown in Figure 15b substantially lacks hydrophobicity. The HC of the entire string of suspended insulators are shown in Figure 16; the HC of the upper surface is maintained at HC4, and the weak hydrophobicity of HC5–HC7 of the insulators is related to the high field strength region of Figure 9b. In Figure 16b, the hydrophobicity of the lower surface is completely lost, whose ESDD is 0.306 mg/cm<sup>2</sup> and NSDD is 3.245 mg/cm<sup>2</sup>, and there is a direct correspondence with the high SPS.



Figure 15. The hydrophobicity of the suspension insulator. (a) The upper surface; (b) the lower surface.

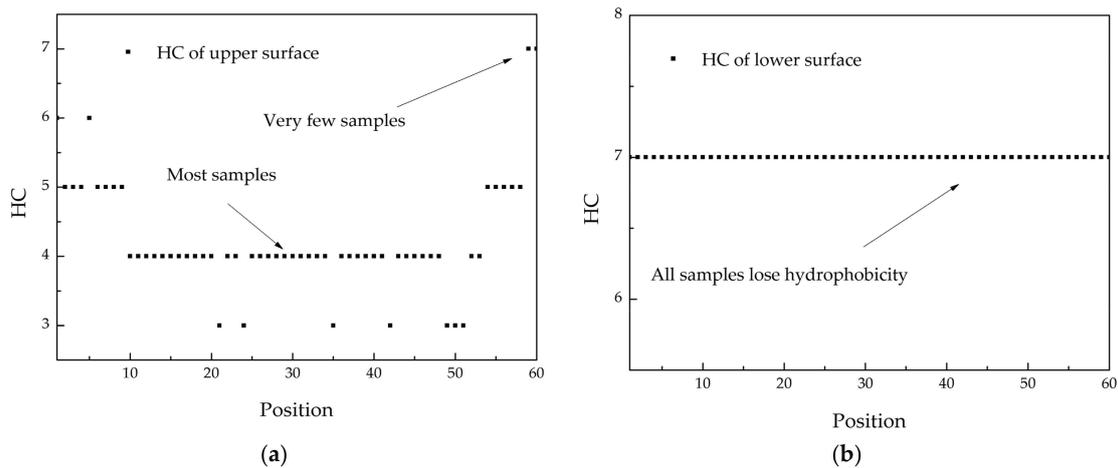


Figure 16. The hydrophobicity of the suspension insulators. (a) The upper surface; (b) the lower surface.

#### 4.5. RFs Analysis and Forecasting

In Table 5,  $p$  represents the  $p$ -th related factor of SPS,  $\alpha$  represents the MDA (or MDG) of each relevant factor in ESDD as a percentage of the total MDA (or MDG).  $\beta$  represents the MDA (or MDG) of each relevant factor in NSDD as a percentage of the total MDA (or MDG). The equations to calculate  $\alpha$  and  $\beta$  are as follows (Equations (15) and (16)).

$$\alpha_{a,p} = \frac{MDA_p}{\sum_{i=1}^{16} MDA_i}; \beta_{a,p} = \frac{MDA_p}{\sum_{j=1}^{16} MDA_j}, \tag{15}$$

$$\alpha_{Gini, p} = \frac{MDG_p}{\sum_{i=1}^{16} MDG_i}; \beta_{Gini, p} = \frac{MDG_p}{\sum_{j=1}^{16} MDG_j} \tag{16}$$

Table 5. The MDA and MDG of related factors.

Feature	ESDD Mean Decrease				NSDD Mean Decrease				$\frac{\alpha_{a,p}}{\beta_{a,p}}$	$\frac{ff_{Gini,p}}{ff_{Gini,p}}$
	Accuracy	$\alpha_a$	Gini Index	$\alpha_{Gini}$	Accuracy	$\beta_a$	Gini Index	$\beta_{Gini}$		
<i>dt</i>	0.00018	1.16%	0.01671	1.29%	0.00992	0.90%	1.549	1.34%	1.28	0.96
<i>ps</i>	0.00024	1.54%	0.01826	1.41%	0.01901	1.73%	2.272	1.97%	0.89	0.72
<i>a</i>	0.00045	2.89%	0.03626	2.80%	0.03902	3.56%	4.544	3.93%	0.81	0.71
<i>d</i>	0.00043	2.76%	0.03616	2.79%	0.04611	4.20%	4.802	4.16%	0.66	0.67
<i>rf</i>	0.00031	1.99%	0.03014	2.33%	0.05835	5.32%	4.701	4.07%	0.37	0.57
<i>t</i>	0.00042	2.70%	0.03668	2.83%	0.03932	3.58%	3.972	3.44%	0.75	0.82
<i>wv</i>	0.00037	2.37%	0.03029	2.34%	0.05881	5.36%	4.722	4.09%	0.44	0.57
<i>HC</i>	0.00125	8.02%	0.12227	9.44%	0.12112	11.04%	17.054	14.77%	0.73	0.64
<i>m</i>	0.00063	4.04%	0.05317	4.10%	0.02811	2.56%	2.462	2.13%	1.58	1.93
<i>pf</i>	0.00191	12.26%	0.18596	14.36%	0.02132	1.94%	5.719	4.95%	6.31	2.90
<i>sa</i>	0.00309	19.83%	0.25613	19.77%	0.32612	29.72%	33.836	29.30%	0.67	0.67
<i>so</i>	0.00178	11.42%	0.12433	9.60%	0.18111	16.50%	15.703	13.60%	0.69	0.71
<i>n</i>	0.00084	5.39%	0.06578	5.08%	0.05104	4.65%	4.713	4.08%	1.16	1.24
<i>pp</i>	0.00099	6.35%	0.08771	6.77%	0.02312	2.11%	2.211	1.91%	3.02	3.54
<i>vl</i>	0.00179	11.49%	0.12433	9.60%	0.02811	2.56%	2.462	2.13%	4.49	4.50
<i>vt</i>	0.00063	4.04%	0.05317	4.10%	0.02873	2.62%	2.462	2.13%	1.54	1.93

The parameters to evaluate the weights are  $\alpha$  for ESDD and  $\beta$  for NSDD [17]. For ESDD, the  $\alpha_a$  ( $\alpha_{Gini}$ ) of the surface factor (*sa*), position factor (*pf*), insulator voltage level (*vl*), surface orientation (*so*), hydrophobicity (*HC*), and the polarity/phases (*pp*) are 19.83% (19.77%), 12.26% (14.36%), 11.49% (9.60%), 11.42% (9.60%), 8.02% (9.44%), and 6.35% (6.77%), respectively. The total impact of the insulator type and electrical factors are 52.94% (52.91%) and 21.88% (20.47%), respectively.

For NSDD, the  $\beta_a$  ( $\beta_{Gini}$ ) of the surface factor (*sa*), surface orientation (*so*), and hydrophobicity (*HC*) are 29.72% (29.30%), 16.50% (13.60%), and 11.04% (14.77%), respectively. The total impact of the insulator type and electrical factors are 55.37% (54.06%) and 7.29% (6.17%), respectively.

According to the ratio of  $\alpha$  to  $\beta$ , the influence of the material (*m*), position factor (*pf*), total length (*n*), polarity/phases (*pp*), voltage level (*vl*), and voltage type (*vt*) on ESDD are greater than that on NSDD, while the other related factors have a larger effect on NSDD. The impact of the voltage level (*vl*), voltage type (*vt*), and polarity/phases (*pp*) on ESDD are 1.5 times, 3 times, and 4.5 times that of NSDD. The impact of the distance from coastline (*d*), wind velocity (*wv*), and rainfall (*rf*) on NSDD are 1.5 times, 2 times, 2.5 times of ESDD. It can be seen that the electrical factors have an obvious impact on ESDD, which corresponds to the results in reference [22].

The smallest influencing factor for ESDD and NSDD is the deposition time (*dt*). It has been considered that the RFs learning data are taken from the annual sampling and the external influencing factors are basically the same in each fixed period, which leads to the least impact on the SPS. On the basis of the SPS measurements, it is believed that the following inputs have high weights for the SPS forecasting of insulators, which include: the surface factor (*sa*), the position factor (*pf*), the insulator voltage level (*vl*), the surface orientation (*so*), hydrophobicity (*HC*), and the polarity/phases (*pp*). When the NSDD is forecasted, the information of the surface factor (*sa*), surface orientation (*so*), and hydrophobicity (*HC*) should be analysed emphatically.

In order to actually detect the validity of the RFs-R trained model, the FC160P/C170DC suspension insulators were suspended on the  $\pm 660$  kV transmission line tower in 2016 for a one-year natural contamination test. A picture is shown in Figure 17. Using the trained RFs-R model, we set the surface factor (*sa*), position factor (*pf*), insulator voltage level (*vl*), the surface orientation (*so*), and the polarity/phases (*pp*) as inputs to predict the ESDD. We set the surface factor (*sa*), surface orientation (*so*), and hydrophobicity (*HC*) as inputs to predict the NSDD. The parameters of the RFs-R model were set as  $n_{treeESDD} = 45$ ,  $n_{treeNSDD} = 66$ , and  $m_{try} = 4$  (Consistent with Section 3.1). The relative error of the predicted ESDD and NSDD and the real ESDD and NSDD is 8.31% and 9.62%, respectively, which is shown in Figure 18.

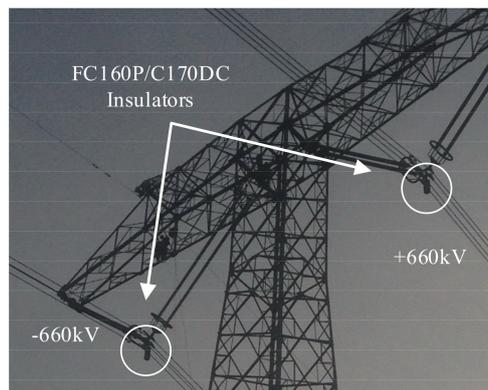


Figure 17. The diagram of the artificial natural pollution test.

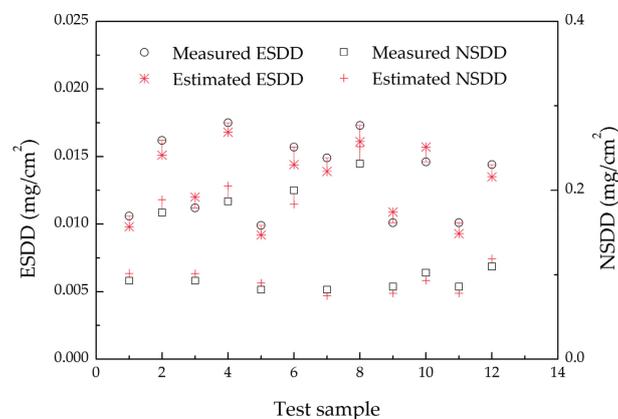


Figure 18. The results of the artificial natural pollution test.

## 5. Conclusions

- (1) The  $R^2$  and MPSE of the trained ESDD RFs-R model are 0.951 and 7.98%, respectively, and the relative error of the predicted ESDD is 8.31%. The  $R^2$  and MPSE of the trained NSDD RFs-R model are 0.911 and 9.04% respectively, and the relative error of the predicted NSDD is 9.62%. Compared with natural contamination test and the SVM regression model, the RFs-R model can effectively predict the natural contamination of insulators.
- (2) According to the MDA (MDG) and MI, the types of the insulators (including surface area, surface orientation, and total length) as well as the hydrophobicity are the main factors affecting both the ESDD and NSDD. Compared with NSDD, the electrical parameters have a significant effect on ESDD. For the influence factors of ESDD, the weights of the insulator type, hydrophobicity, and meteorological factors are 52.94%, 6.35%, and 21.88%, respectively. For the influence factors of NSDD, the weights of the insulator type, hydrophobicity, and meteorological factors are 55.37%, 11.04%, and 14.26%, respectively.
- (3) The effect of electrical parameters on the ESDD is greater than that on NSDD, while other non-electrical parameters have a significant impact on NSDD. The influence that the voltage level ( $vl$ ), voltage type ( $vt$ ), and polarity/phases ( $pp$ ) exert on ESDD are 1.5 times, 3 times, and 4.5 times that of NSDD. The influence that the distance from coastline ( $d$ ), wind velocity ( $wv$ ), rainfall ( $rf$ ) exert on NSDD are 1.5 times, 2 times, and 2.5 times that of ESDD.
- (4) For engineering reasons, SPS has been measured only in a fixed yearly period, and on-line daily ESDD and NSDD measurements are urgently needed, although considerable results have been achieved. More locations and higher accuracy data should be collected and analysed to quantitatively reveal more robust and accurate rules.

**Acknowledgments:** This project was supported by the Science and Technology Project of the State Grid Corporation of China (SGTYHT/15-JS-193).

**Author Contributions:** Ang Ren and Huaishuo Xiao conceived and designed the experiments; Ang Ren wrote the paper and Qingquan Li revised it.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Li, L.; Li, Y.; Lu, M.; Liu, Z.; Wang, C.; Lv, Z. Quantification and comparison of insulator pollution characteristics based on normality of relative contamination values. *IEEE Trans. Dielectr. Electr. Insul.* **2016**, *23*, 965–973. [CrossRef]
- Jiao, S.B.; Liu, D.; Xie, G.; Yi, D. Assessment of contamination condition of insulator based on PSO-SVM. In Proceedings of the 4th IEEE Conference on Industrial Electronics and Applications, Xi'an, China, 25–27 May 2009.
- Ahmad, A.S.; Ghosh, P.S.; Ahmed, S.S.; Ahmad, S.S.; Aljunid, S.A.K. Assessment of ESDD on high-voltage insulators using artificial neural network. *Electr. Power Syst. Res.* **2004**, *72*, 131–136. [CrossRef]
- Karamousantas, D.C.; Chatzarakis, G.E.; Oikonomou, D.S.; Karampelas, P. Effective insulator maintenance scheduling using artificial neural networks. *IET Gener. Transm. Distrib.* **2010**, *4*, 479–484. [CrossRef]
- Muniraj, C.; Chandraseka, S. Adaptive neurofuzzy inference system-based contamination severity prediction of polymeric insulators in power transmission lines. *Adv. Artif. Neural Syst.* **2011**, *2011*, 1–9. [CrossRef]
- Hannan, M.A.; Ali, J.A.; Mohamed, A.; Uddin, M. A random forests regression based space vector PWM inverter controller for the induction motor drive. *IEEE Trans. Ind. Electron.* **2016**, *64*, 2689–2699. [CrossRef]
- Samantaray, S.R. A data-mining model for protection of FACTS-based transmission line. *IEEE Trans. Power Deliv.* **2013**, *28*, 612–618. [CrossRef]
- Shah, A.M.; Bhalja, B.R. Fault discrimination scheme for power transformer using random forests technique. *IET Gener. Transm. Distrib.* **2016**, *10*, 1431–1439. [CrossRef]
- Kannan, K.; Shivakumar, R.; Chandrasekar, S. A random forests model based contamination severity classification scheme of high voltage transmission line insulators. *J. Electr. Eng. Technol.* **2016**, *11*, 951–960. [CrossRef]
- Larivière, B.; Van den Poel, D. Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Syst. Appl.* **2005**, *29*, 472–484. [CrossRef]
- Selection and Dimensioning of High-Voltage Insulators Intended for Use in Polluted Conditions*; IEC: Geneva, Switzerland, 2008.
- Lv, Y.; Li, J.; Zhang, X.; Pang, G.; Liu, Q. Simulation study on pollution accumulation characteristics of XP13-160 porcelain suspension disc insulators. *IEEE Trans. Dielectr. Electr. Insul.* **2016**, *23*, 2196–2206. [CrossRef]
- Jiang, X.; Wang, S.; Zhang, Z.; Hu, J.; Hu, Q. Investigation of flashover voltage and non-uniform contamination correction coefficient of short samples of composite insulator intended for  $\pm 800$  kV UHVDC. *IEEE Trans. Dielectr. Electr. Insul.* **2010**, *17*, 71–80. [CrossRef]
- Gençoğlu, M.T.; Cebeci, M. Review, investigation of contamination flashover on high voltage insulators using artificial neural network. *Expert. Syst. Appl.* **2009**, *36*, 7338–7345. [CrossRef]
- Ahmad, A.S.; Ghosh, P.S.; Aljunid, S.A.K.; Ahmad, H. Modeling of various meteorological effects on contamination level for suspension type of high voltage insulators using ANN. In Proceedings of the 1st International Conference and Exhibition on Transmission and Distribution in the Asia Pacific Region, Yokohama, Japan, 6–10 October 2002.
- Zhao, L.; Li, C.; Xiong, J.; Wang, C. An artificial contamination test on silicone rubber insulators under long-time wetted conditions. In Proceedings of the 2007 Annual Report—Conference on Electrical Insulation and Dielectric Phenomena, Vancouver, Canada, 14–17 October 2007.
- Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
- Package “randomForest”. Available online: <https://cran.r-project.org/web/packages/randomForest/randomForest.pdf> (accessed on 8 June 2017).
- Janitza, S.; Strobl, C.; Boulesteix, A.L. An AUC-based permutation variable importance measure for random forests. *BMC Bioinform.* **2013**, *14*, 1–11. [CrossRef] [PubMed]

20. Cover, T.M.; Thomas, J.A. Elements of information theory. In *Wiley Series in Telecommunications and Signal Processing*; Honig, M., Ed.; Wiley & Sons Inc.: New York, NY, USA, 1990; Volume 294, pp. 155–183.
21. Swedish Transmission Research Institute. Guide 1.92/1 Hydrophobicity Classification Guide. Available online: [http://www.stri.se/wwwpublic/STRI\\_Guide\\_1\\_92\\_1.pdf](http://www.stri.se/wwwpublic/STRI_Guide_1_92_1.pdf) (accessed on 8 June 2017).
22. Takasu, K.; Shindo, T.; Arai, N. Natural contamination test of insulators with DC voltage energization at inland areas. *IEEE Trans. Power Deliv.* **1988**, *3*, 1847–1853. [[CrossRef](#)]



© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).