Hindawi
Journal of Sensors
Volume 2018, Article ID 6026259, 12 pages
https://doi.org/10.1155/2018/6026259

Research Article

Machine Learning for Estimating Leaf Dust Retention Based on Hyperspectral Measurements

Wenlong Jing,[1,2,3] **Xia Zhou**,[1,2,3] **Chen Zhang**,[1,2,3,4] **Chongyang Wang**,[1,2,3] **and Hao Jiang**[1,2,3]

[1]*Guangzhou Institute of Geography, Guangzhou, China*
[2]*Key Laboratory of Guangdong for Utilization of Remote Sensing and Geographical Information System, Guangzhou, China*
[3]*Guangdong Open Laboratory of Geospatial Information Technology and Application, Guangzhou, China*
[4]*Shandong University of Science and Technology, Shandong, China*

Correspondence should be addressed to Wenlong Jing; jingwl@lreis.ac.cn

Received 13 April 2018; Revised 7 June 2018; Accepted 26 June 2018; Published 6 September 2018

Academic Editor: Zongyao Sha

Hyperspectral sensors provide detailed information for dust retention content (DRC) estimation. However, rich hyperspectral data are not fully utilized by traditional image analysis techniques. We integrated several recently developed machine learning algorithms to estimate DRC on plant leaves using the spectra measured by the ASD FieldSpec 3. The experiments were carried out on three common green plants of southern China. The important hyperspectral variables were first identified by applying the random forest (RF) algorithm. Three estimation models were then developed using the support vector machine (SVM), classification and regression tree (CART), and RF algorithms. The results showed that the increase in dust retention contents on plant leaves enhanced their reflectance in the visible wavelength but weakened their reflectance in the infrared wavelength. Wavelengths in the ranges of 450–500 nm, 550–600 nm, 750–1000 nm, and 1100–1300 nm were identified as important variables using the RF algorithm and were used to estimate the DRC. The comparison of the three machine learning techniques for DRC estimation confirmed that the SVM and RF models performed well because their estimations were similar to the measured DRC. Specifically, the average R^2 for SVM and RF model are 0.85 and 0.88. The technical approach of this study proved to be a successful illustration of using hyperspectral measurements to estimate the DRC on plant leaves. The findings of this study can be applied to monitor the DRC on leaves of other plants and can also be integrated with other types of spectral data to measure the DRC at a regional scale.

1. Introduction

Haze has been an increasing air pollution issue with the rapid industrialization and urbanization of China [1, 2]. Urban green plants play a significant role in improving the urban atmospheric environment by retaining dust [3–5]. Measurement of the dust retention capacity of green plants is helpful for monitoring and controlling urban air pollution [6, 7]. It is also valuable for evaluation of the influence of green plants upon the urban atmospheric environment, further directing the arrangement and management of urban green plants [8, 9].

The mass difference method is the most basic measuring approach of dust retention content (DRC) of plants [10, 11]. However, this method is complicated and time-consuming. In addition, the monitoring of DRC at the regional scale is impossible using the mass difference method. Remotely sensed hyperspectral technology has provided an efficient way for monitoring effects of environmental pollution on plants by measuring plant spectral characteristics [12–15]. Previous studies have investigated the influence of dust upon the spectral curve and characteristics of plants [16, 17]. According to Horler et al. [18], dust retention on leaves leads to a change in the red edge position, but experiments conducted by Xiao [19] showed that dust retention has no impact on the red edge position. The red edge position, according to Wang et al. [20], is not sensitive to the influence of detained dust on leaves, but the red edge slope and the area of the spectra of dust-covered leaves are smaller than those of dust-less leaves. In summary, the spectral characteristics of clean leaves differ from those of dust-covered leaves, which

Table 1: Physical characteristics of the leaf surface.

Plants	Characteristics of the leaf surface				Microscopic feature	Dust retention feature
	Color	Size	Shape	Texture		
GL	Golden yellow or glossy dark green	3–8 cm	Ellipse	Leathery, fleshy and entire leaf, two sides smooth and glabrous	Many depressions in leaf caused by unequal-sized graininess wax, that lots of particle can adhere to them	Massive stratification
LC	Dull-red or green	2–5 cm	Oval or oblate ellipsoid, round root but deflect, asymmetric	Leathery and entire leaf, stellate leaf trichome in two sides	Rough leaf surface with villus	Mosaic particle
CF	Green or purplish red	25–50 cm	Ellipse to elliptical lanceolate	Smooth leaf surface	Cell profile distinct and aligned	Sporadic mosaic particle

makes it possible to estimate the DRC of plant leaves based on remotely sensed spectral information [12, 19, 21, 22].

Spectral information provided by hyperspectral measurement devices or sensors commonly includes hundreds of bands. Selecting the appropriate bands for a DRC estimation model is significant. The normalized difference vegetation index (NDVI), three-edge (blue, red, and yellow) positions, and slopes for use in constructing a DRC estimation model have been assessed through experimentation by some scholars [21, 23–25]. These variables, however, are mostly selected empirically; considerable useful information was excluded. Univariate/multiple linear regression and the partial least squares regression algorithm have been investigated to simulate the relationship between the DRC and spectral characteristics [16]. However, few studies have been conducted using machine learning. Machine learning, as a powerful modeling tool, has successfully improved the estimation and classification accuracy of environmental variables (air pollution, vegetation health condition, soil moisture, land surface temperature, etc.) and land cover types from remotely sensed images [26–32]. In addition, machine learning algorithms are excellent in solving nonlinear problems of variables with very high dimensions. Therefore, this study attempted to investigate the possibilities of constructing an estimation model for DRC using machine learning algorithms.

The objectives of this study were to (1) investigate the effects of the DRC on the spectral characteristics of leaf surfaces, (2) extract important bands from hyper spectra to reduce the very high dimensions of the variables and further evaluate the effectiveness of the selected bands, and (3) construct a DRC estimation model using machine learning algorithms. In this study, we conducted the experiments on three commonly planted green plants in southern China, and the spectra were measured using an ASD FieldSpec 3 device. Three machine learning regression algorithms were used for DRC estimation model construction for comparison purposes.

2. Materials and Methods

2.1. Experimental Plants and Sampling Collection. For the research purpose, three plants were chosen as study objects, including *Ficus microcarpa* L. f. cv Golden leaves (GL), *Loropetalum chinense* (R. Br) Oliv. var. rubrum Yieh (LC), and *Cordyline fruticosa* (L.) A. Cheval (CF). These plants are the most common and typical green plants in southern China; investigation on these plants is beneficial to promoting the simulation of DRC to regional scale in the experiment area. Moreover, as is shown in Table 1, the leaves of three plants have different characteristics on color, size, shape, and surface texture. This is also helpful in detecting the usefulness and possibilities of simulating the DRC by using a machine learning algorithm on different plant leaves. To reduce the influence of extreme rainfall and wind, the plants were planted under an open greenhouse in Guangzhou (Figure 1). Figure 2 shows the three plants and their leaves.

Figure 1: In situ experiment.

Ten experiments were carried out from October 2017 to January 2018. Before the experiments, the plants were sufficiently washed to ensure a dust-free state on the leaves. Later, the leaves of the three plants were collected at an interval of 5–7 days on cloudless days. Leaves of similar health and age conditions were collected at the top, middle, and bottom of the canopy, respectively, and then sealed in sampling bags for the measurement of dust retention. Considering the different sizes of the leaves, the number of leaves collected for each plant was different. Among these, there were 30 pieces for GL, 50 pieces for LC, and 5 pieces for CF. Three replicate measurements were taken each time.

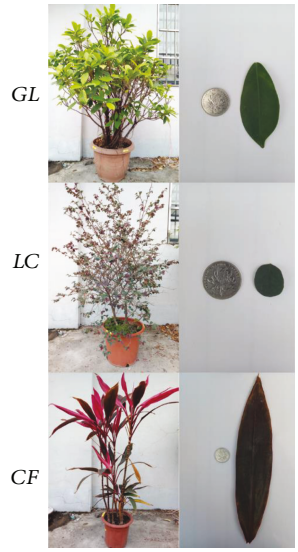

Figure 2: The three plants and their leaves.

2.2. Measurement of Leaf Reflectance Spectra and Dust Content

2.2.1. Measurement of Leaf Reflectance Spectra. Spectral measurements were recorded for each plant before and after dust retention to obtain the precise measurement of dust-related spectral changes. In the study, an ASD FieldSpec 3 spectrograph was used to measure the spectral reflectance of the leaves. The equipment was based on the basic theory of electromagnetic waves. The optical probe collected the electromagnetic waves reflected from the ground objects, which were later transformed into digital signals. The spectral range of this spectrograph was 350–2500 nm and with a sampling interval of 1 nm. The measuring speed was fixed at 0.1 s. The length of the optical probe was 1 m. The front field angle was 25°. The reflectance, radiation, and irradiance of the ground objects were simultaneously collected.

At the same time of leaf collection, the spectral reflectance was measured from 10 : 00 to 14 : 30. The front sides of five leaves were both measured and repeated five times. After measurement, the leaves were sealed in corresponding bags and used to measure the DRC with the collected leaves. The probe should be perpendicular to the leaf, and the height should change according to the leaf size to guarantee spectrometer probe falls within the scope of the leaf. Specifically, the height from the probe to the leaf surface is calculated based on the equation as follows:

$$H \leq \frac{W_L/2}{\tan \theta/2}, \quad (1)$$

where H is the height, W_L is the width of the leaf, and θ is the front field angle. Considering that the leaf size for one plant could also be different, we take the minimum width of each plant type as W_L when calculating the height. For GL and LC, we take 2 cm as W_L, then H is less equal to 4.5 cm; for CF, we take 20 cm as W_L, then H is less equal to 45 cm. In practice, we used H of 4 cm for GL and LC, and 10 cm for CF.

2.2.2. Measurement of Dust Content. In this study, the plant's dust retention ability is represented by the dust content per unit leaf area, which is measured by the ratio of the leaf dust content and unit leaf area within a certain time [6]. The leaf dust content is measured by the weight difference method, and the unit leaf area is measured using a leaf area meter.

The first procedure of the weight difference method is washing the leaves and then drying them. The weights of the two conditions should be measured, and the difference is considered to be the dust content. To avoid other effects on the experiment, the control variate method was used in this research. That means except for the liquid in the beaker, the soak time, dried water, drying temperature, drying time, and cooling time were the same. The detailed method is as follows:

(1) The collected leaves were sealed in a dried and weighted beaker (W1) corresponding to the bags, then soaked in pure water for 5–6 hours. At the same time, the sampling bag was washed using pure water and then poured into a beaker.

(2) The beaker was stirred with a clean glass rod to make the dust dissolve into the water. Then, the leaves were sufficiently washed twice using a tweezer and pure water. After washing, the tweezer and glass rod were washed, and the flushing water was poured into the soak beaker.

(3) The beaker was placed in the drying apparatus, and the temperature was set to 85°C and the drying time set to 25 hours. After drying, the beaker was placed in a drying vessel for 5 hours to cool down.

(4) The leaves were weighted (W2) using a ten-thousandth electronic analytical balance. The difference weight ($\Delta W = W2 - W1$) is the dust retention of the collected sample. The leaf area on the dried clean leaves of each bag was measured using the leaf area meter device. The ratio of dust retention ΔW/leaf area S is the unit leaf dust retention value.

2.3. Spectral Data Process.

Before the spectral characteristic analysis, the noise and nonsensitive wave band needed to be removed from the collected spectral data, and the spectral transformation should be completed for the following research. First, to ensure the accuracy of the spectral measured data, we examined the five-times repeated spectral data of each point and removed the evident deviated curves. Then, the mean value was calculated. Second, to ensure the comparability of the measured data from different times and conditions and eliminate errors caused by the experimental environment, we divided it by the white board reflectance because the spectral reflectance should follow the principle of proximity. Then, we removed the bands that were assimilated by water vapor because water vapor assimilation has a great effect on the spectral curve, and this wavelength band range is meaningless in botany spectral research. To simplify the following data process, the wave bands were removed directly during this research. The result is shown in Figure 3.

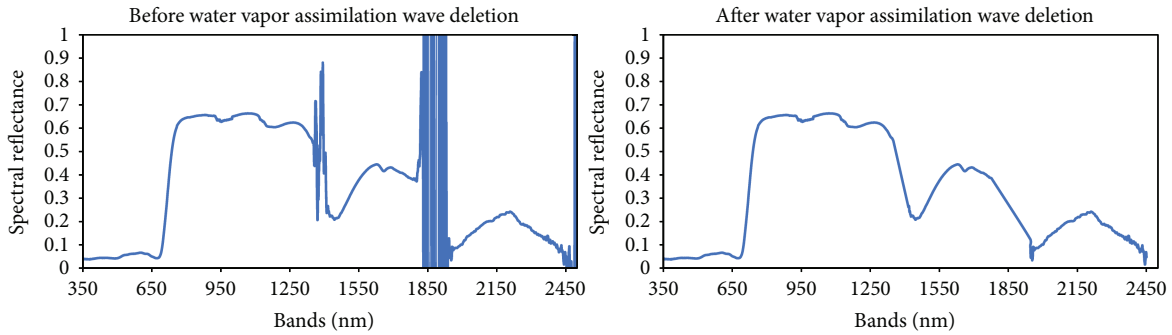

Figure 3: Comparison before and after water vapor assimilation wave deletion.

The first derivative spectra can reduce sublinear and quadratic background noise; it can also reduce the impact of the hyperspectral remote-sensing measurement because of the scattering and absorption of atmosphere to light [33]. It is also able to reduce the impact of multiplicative factors that are produced by changing lighting conditions [22]. Therefore, the first derivative spectra computed from the original measured reflection spectra were implemented to establish the estimation model of the DRC. The equation of the first derivative (dR) is expressed as follows:

$$\mathrm{dR}(\lambda_i) = \frac{R(\lambda_{i+1}) - R(\lambda_{i-1})}{\lambda_{i+1} - \lambda_{i-1}}. \quad (2)$$

λ_{i+1}, λ_i, and λ_{i-1} are the adjacent wavelengths; dR (λ_i) is the first derivative of wavelength λ_i; R (λ_{i+1}), R (λ_i), and R (λ_{i-1}) are the original reflectance of the wavelengths λ_{i+1}, λ_i, and λ_{i-1}, respectively.

2.4. Identifying Feature Importance and Estimating the DRC. In this section, we present our key innovation: integrating machine learning algorithms to identify the important spectral features (bands). We estimated the DRC based on these selected features and compared the effectiveness of the various machine learning algorithms. Our design is depicted in Figure 4.

2.4.1. Identifying Feature Importance. The first derivative spectra were computed from the original measured reflection spectra. Over a thousand independent variables are provided in the original sample datasets. A feature selection needed to be conducted on the very high-dimensional datasets to exclude the redundant bands, further reducing the dimension of the sample sets. In this study, we implemented random forest (RF) to measure the feature importance (FI) of each band [34–36]. The FI values were then sorted from highest to lowest, and the cumulative value of 0.9 was used to derive the important variables.

RF is an outstanding ensemble learning algorithm. The basic concept of the algorithm is to construct numerous tree-based predictors to obtain better performance [37, 38]. A number of subsets are extracted randomly from the total sample with replacement. The remaining samples are out-of-bag data (OOB). Then, a classification and regression tree (CART) was generated at each subset. The prediction was obtained by averaging all the outputs from the prediction of each subset. The feature importance of the bands can be described as follows:

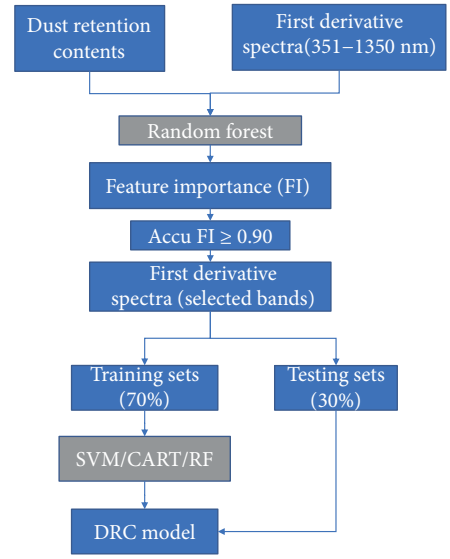

Figure 4: Flowchart of the DRC estimation model.

(1) The OOB of each subset were used to test the estimation of the corresponding tree, and the average errors of the OOB of all trees were calculated.

(2) The order of values of the kth variable was permuted while the other variables remain unchanged.

(3) The OOB error of the permuted set was calculated.

(4) The difference in the OOB error before and after the permutation was calculated for each variable, and the differences of all the variables were then normalized to a range of from 0 to 1.0, which were the FI values of the variables.

2.4.2. Model for DRC Estimation. The DRC estimation model was established based on the selected variables. Three commonly used machine learning algorithms were used to construct the models between DRC and first derivative spectra. SVM, as a supervised learning algorithm, has performed well in many remote-sensing applications [39–42]. CART, as a tree-based algorithm, is easily implemented and has been

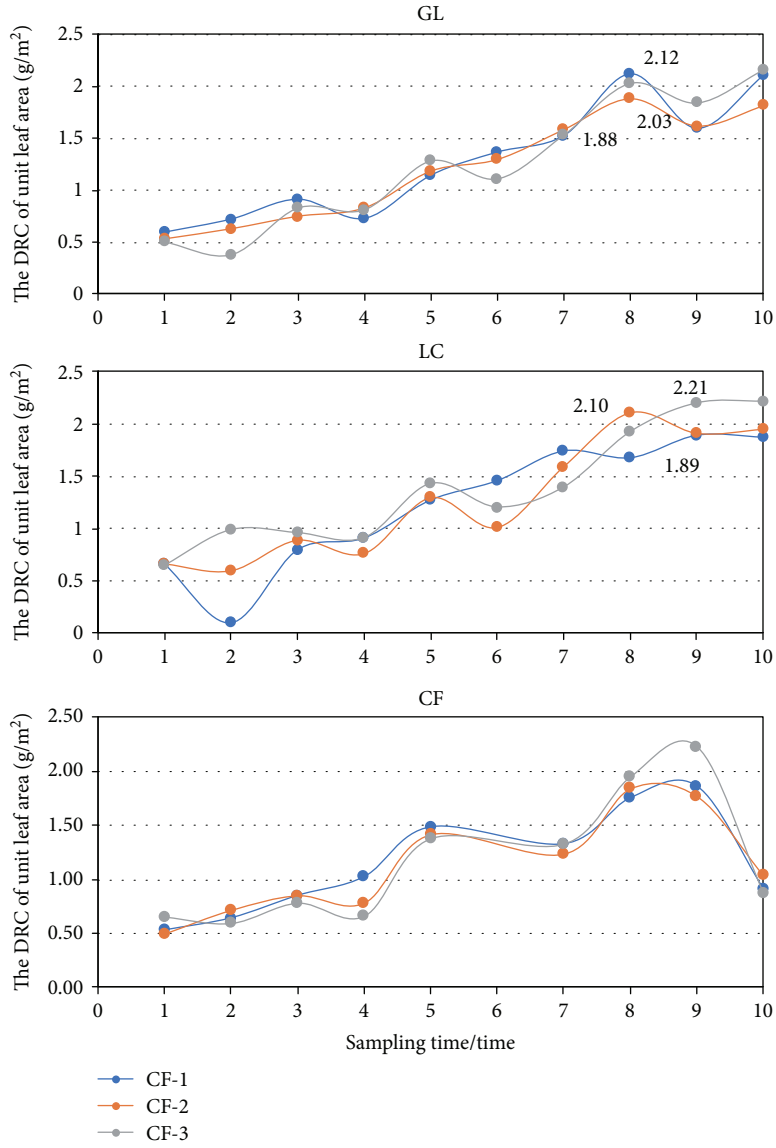

Figure 5: Temporal changes in unit DRC of the three green plants.

widely used in a variety of fields [43–45]. The aforementioned RF algorithm, which has been used for variable selection, was also implemented to model the DRC. These three machine learning algorithms were selected for comparison of their effectiveness in establishing the DRC estimation model because of their extensive use, good performance, and easy implementation. For validation purposes, the total sample set was randomly divided into two subsets, with the 70% part used to train the model and the 30% part set aside for testing.

3. Results and Discussion

3.1. Spectral Characteristics. Figure 5 shows the temporal changes in the unit DRC of the three green plants. It was found that the unit DRC increases with time but tends to remain stable or decrease because of a saturation effect after a long time. In addition, it was also found that the dust retention capability of CF is the most significant among the three green plants, with a mean unit DRC of 2.23 g/m^3, followed by that of LC (2.21 g/m^3). The dust retention capability of GL was the lowest at 2.12 g/m^3. The DRC of plant leaves, as a result, increased temporally and reached saturation after a certain time, without rainfall and wind disturbance.

In general, the reflectance of the three green plants with or without dust on the leaves showed a universal characteristic, which is similar to the reflectance of many other plants. Due to the pigment absorption of the plants, the reflectance of the three green plants in the visible wavelength is low (less than 20%) and they usually have a reflection peak (at approximately 550 nm for GL and LC and approximately 600 nm for CF). The reflectance in the infrared band is high (greater than 60%) as contributed by the structure of the plants' leaves.

For the same plant, the reflectance of the leaves with dust and those that are clean was significantly different, particularly in the visible-infrared wavelength (Figure 6). Compared to the reflectance of clean leaves, that of the dust-covered

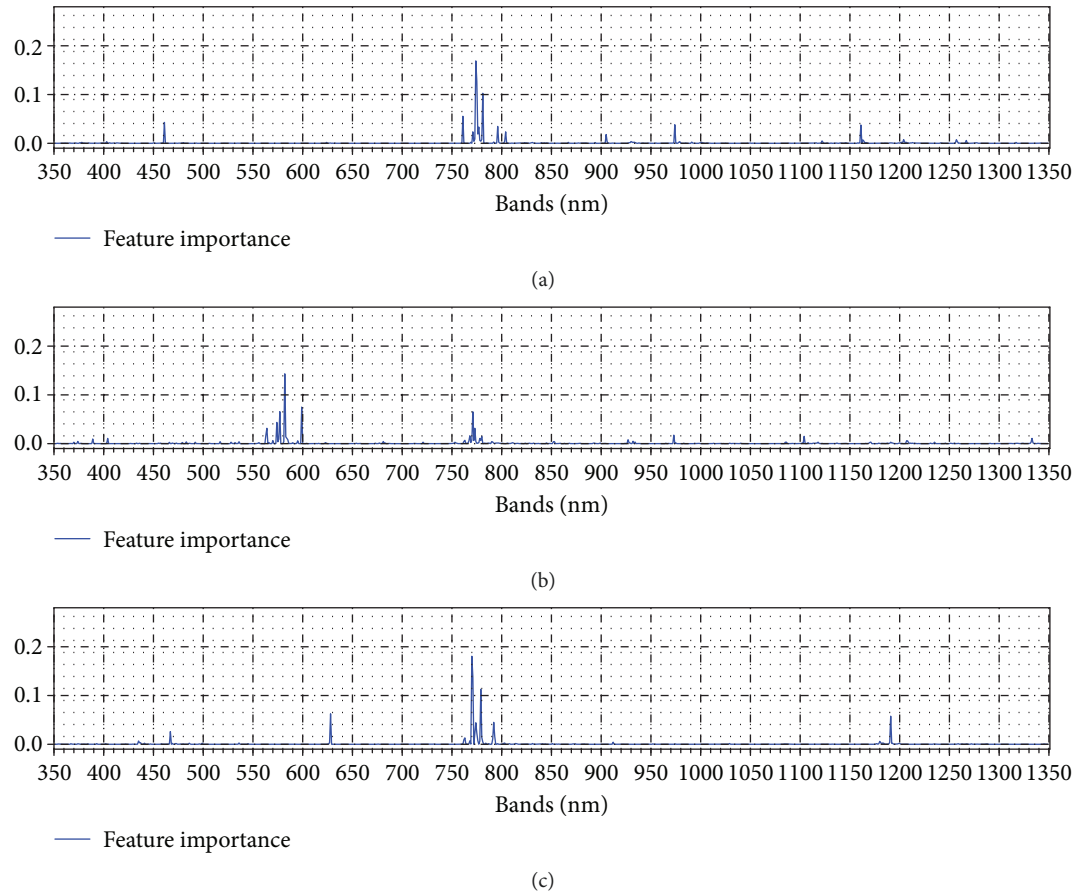

Figure 6: Feature importance of different bands calculated using RF for each plant: (a) GL, (b) LC, and (c) CF.

leaves in the visible wavelength markedly increased by more than 70%. It is well known that the reflectance of plant leaves in the visible wavelength is dependent on the absorption and reflection of the leaves' pigments. The dust retention on the leaves not only influences the pigment absorption but also increases the leaves' reflection, which causes high reflectance in the visible wavelength.

However, the reflectance of dust-covered leaves in the infrared wavelength is significantly lower than that of the clean leaves. The main reason is that dust on the leaves can decrease multiple reflections in the leaves' structure, particularly of CF (Figure 7). The characteristic of the reflectance in the shortwave infrared wavelength of the three green plants with or without dust on the leaves is random, because the reflectance in the shortwave infrared wavelength of the leaves was mainly affected by leaf water. Dust retention had little impact on leaf water (Figure 7).

Based on the aforementioned analysis, it was found that significant correlations between the reflectance in the visible-infrared wavelength of the plant leaves and dust retention exist. Thus, it is possible to establish a model for estimating the DRC on plant leaves by using spectral information.

3.2. Dust Retention Estimation Model

3.2.1. Band Selection Results. As in the aforementioned analysis, the reflectance in the shortwave infrared wavelength of the leaves was mainly affected by leaf water and only slightly influenced by dust retention. Therefore, we only included spectra between the wavelengths of 351 and 1350 nm. There is a total of 1000 bands from the 351 nm to 1350 nm wavelength with an interval of 1 nm. As a result, 1000 independent variables were provided in the sample datasets.

The RF regression algorithm was run for the total samples for each plant, respectively. FI values of bands in a range of 351–1350 nm were then derived from the algorithm outputs. Figure 6 presents the distribution of FI estimated using RF at different bands. It can be seen that FI values for the three plants show a different distribution pattern. Overall, the FI for the three plants all reach peak values between a wavelength of 750 nm and 800 nm, which are the near infrared bands that are sensitive to vegetation health [46]. However, for GL and CF, the high FI values are mainly concentrated between 750 nm and 800 nm, whereas the FI values of LC range from 550 to 600 nm and overwhelm those of the near infrared bands. In addition, FI values also occur at 460 nm and 630 nm for GL and CF, respectively.

The FI values were then sorted from highest to lowest; we then accumulated the FI values from highest to lowest. The accumulating was stopped when the accumulated FI value reached 0.9. Then the bands that have been accumulated were considered as important variables and selected to establish the DRC model. Figure 8 shows the selected bands for each plant based on the threshold of 0.9 for the cumulative

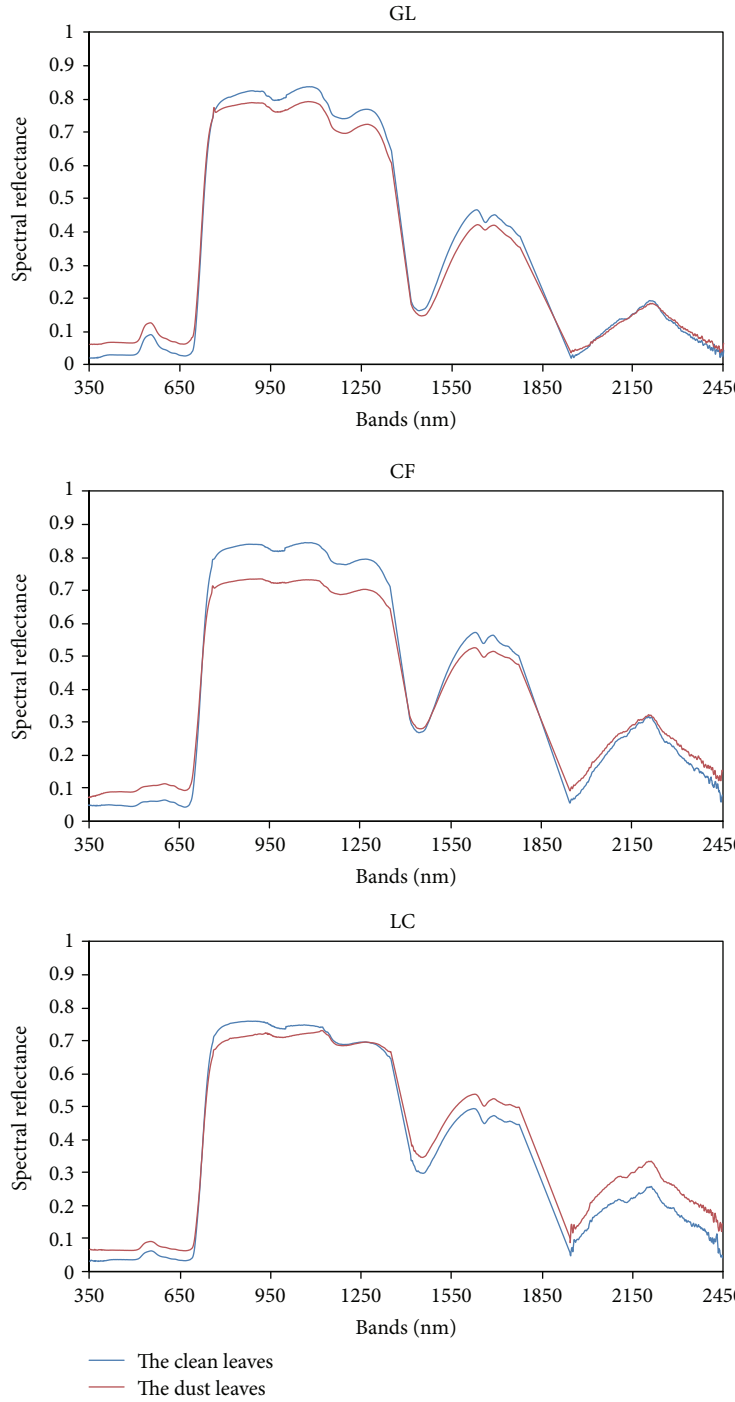

Figure 7: Reflectance of dust-covered and clean leaves.

FI value. The number of selected bands for GL, LC, and CF is 94, 167, and 73, respectively, much less than the total number of samples. Because the distribution of FI values for three plants is different, the number of selected bands is different. For GL and CF, the selected bands are mainly in the ranges of 350–500 nm, 750–900 nm, and 1100–1350 nm. For LC, except for the same ranges of GL and CF, many bands were also selected from 550 to 600 nm and 650 to 700 nm.

In theory, the spectra of the selected bands can provide 90% of the information of the total samples. The selected bands were included as independent variables to establish an estimation model for DRC.

3.2.2. Dust Retention Model. The total sample set was divided into two subsets randomly, with the 70% part used to train the model and the 30% part set aside for testing. Models were established for each plant independently based on the selected bands. Moreover, models were also established using the total bands for comparison purposes. Table 2 summarizes the training and test results for each plant using the CART,

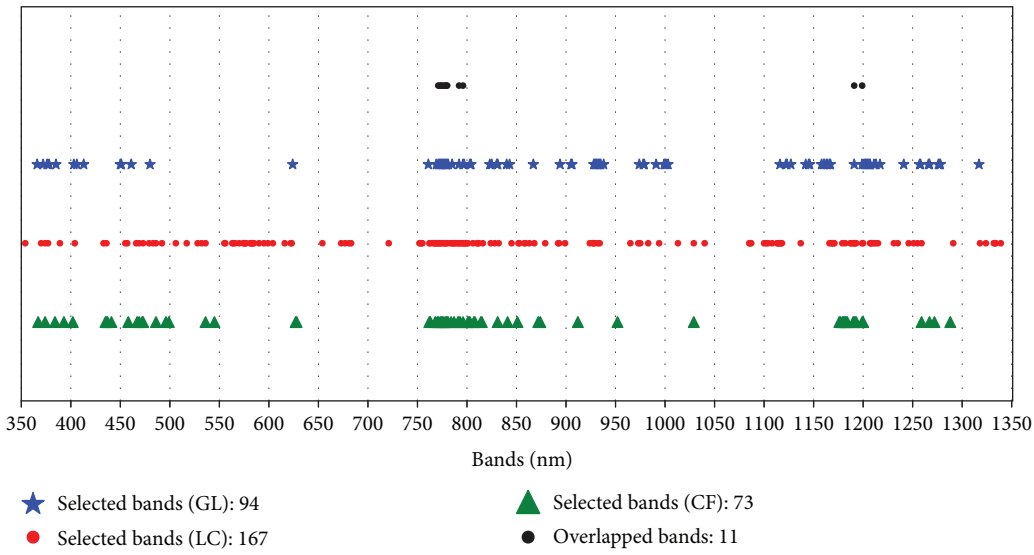

Figure 8: Distribution of selected bands based on the FI values for each plant.

Table 2: Training and testing results for three plants using the different regression algorithms.

Plants	Algorithms	Training						Testing				
		Number of bands	R^2	MAE (g/m^2)	RMSE (g/m^2)	Bias	Number of samples	R^2	MAE (g/m^2)	RMSE (g/m^2)	Bias	Number of samples
GL	CART	1000	0.98	0.04	0.08	0.00	99	0.78	0.17	0.26	0.05	42
		94	0.95	0.08	0.13	0.00		0.76	0.18	0.26	0.04	
	RF	1000	0.98	0.07	0.10	0.00		0.88	0.14	0.19	0.05	
		94	0.98	0.06	0.08	0.00		0.92	0.11	0.15	0.03	
	SVM	1000	0.97	0.08	0.09	−0.01		0.89	0.14	0.19	0.05	
		94	0.92	0.13	0.16	0.00		0.89	0.16	0.19	0.07	
LC	CART	1000	0.97	0.05	0.10	0.00	97	0.50	0.33	0.47	−0.06	41
		167	0.95	0.07	0.12	0.00		0.50	0.29	0.46	−0.03	
	RF	1000	0.97	0.08	0.11	0.00		0.83	0.17	0.23	−0.03	
		167	0.98	0.07	0.10	0.00		0.87	0.14	0.20	−0.02	
	SVM	1000	0.97	0.09	0.10	−0.01		0.80	0.17	0.23	−0.02	
		167	0.95	0.11	0.13	0.00		0.77	0.20	0.25	−0.01	
CF	CART	1000	0.93	0.08	0.13	0.00	94	0.64	0.20	0.31	0.00	41
		73	0.98	0.04	0.07	0.00		0.70	0.17	0.29	0.03	
	RF	1000	0.97	0.06	0.09	0.00		0.80	0.16	0.23	0.02	
		73	0.98	0.05	0.08	0.00		0.84	0.14	0.21	0.02	
	SVM	1000	0.97	0.06	0.08	0.01		0.90	0.13	0.17	0.03	
		73	0.97	0.08	0.08	0.00		0.88	0.14	0.18	−0.01	

SVM, and RF methods, respectively. Figures 9 and 10 show the scatter plots between the estimated and measured dust amount by model using the selected bands and total bands, respectively.

In general, the models established based on selected and total bands both achieved high accuracy, with a similar coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE). The testing results also imply that the models implementing the selected bands produced comparable performance to that of the models considering total bands. The RF-based models for LC and GL using selected bands produced evidently higher R^2 values than those using the total bands. This indicates that dimensionality reduction is necessary, and the selected bands preserve important and useful information that could express the spectral differences of the leaves caused by dust retention.

For LC and GL, the RF-based model performed better than the other two algorithms, with a testing R^2 of 0.87 and 0.92, respectively. The bias values for the models are all less than 0.08. For CF, the SVM-based model outperformed the

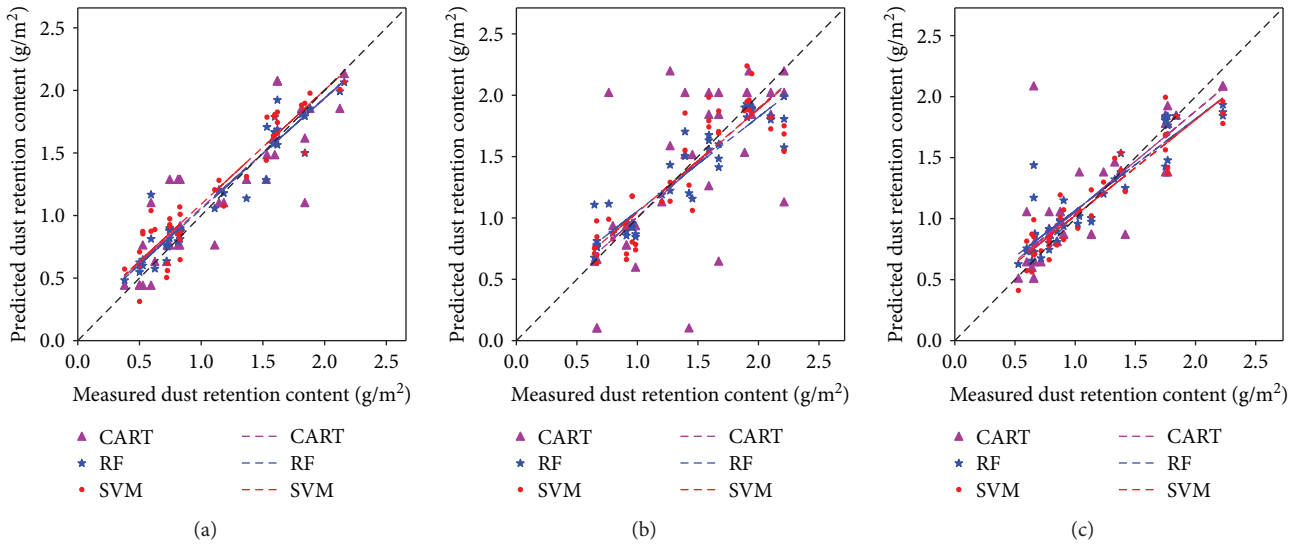

Figure 9: Scatter plots of estimated and measured DRC based on selected bands: (a) GL, (b) LC, and (c) CF.

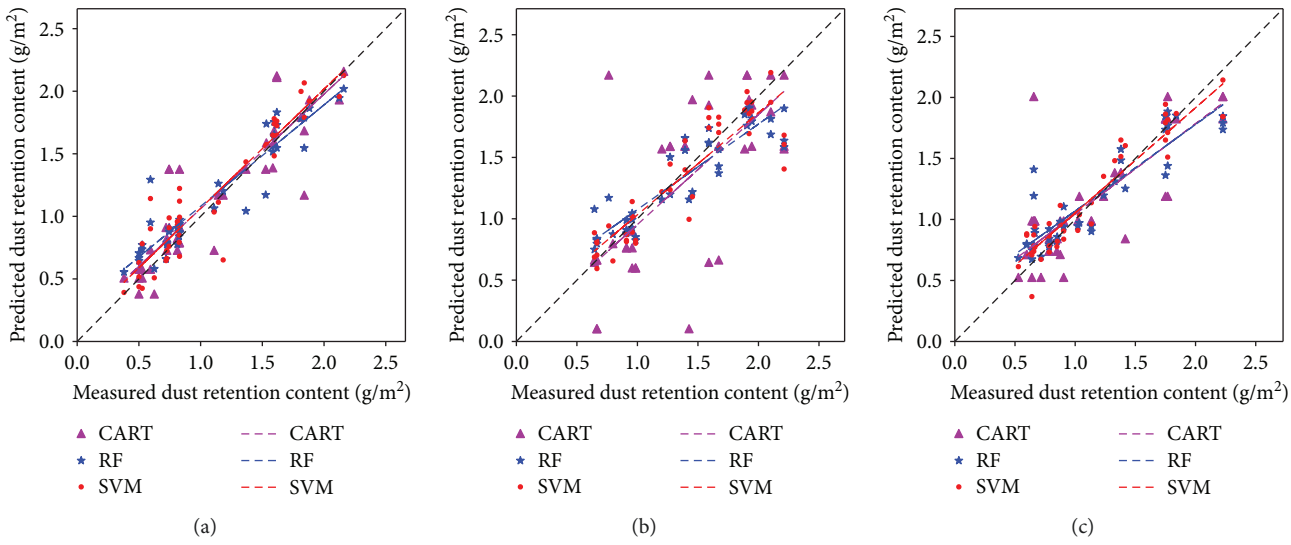

Figure 10: Scatter plots of estimated and measured DRC based on total bands: (a) GL, (b) LC, and (c) CF.

other two algorithms. Overall, the CART-based model produced the least accurate estimation results. Figure 11 shows the Taylor diagrams of testing accuracy for the three plants. The Taylor diagram is helpful for comparative assessment of different models. It integrates four statistics on a diagram to quantify the degree of correspondence between the modeled and observed values. The four statistics are the Pearson correlation coefficient, normalized standard deviation, normalized error standard deviation, and normalized bias. The measured (observed) values serve as the reference. More details can be found in [47]. According to Figure 11, the estimated results produced by the CART-based model show the lowest correlations with the measured values. The results estimated using the SVM and RF models show similar standard deviations and correlations. In addition, the three algorithms all underestimated the measured dust contents of LC.

4. Conclusion

In this study, we investigated the effectiveness of using machine learning to estimate the DRC of leaves based on remotely sensed hyperspectral information. We conducted experiments on three green plants in southern China. The spectra were measured using an ASD FieldSpec 3. A feature selection process was implemented to reduce the high dimensions of the original spectra. Three commonly used machine learning algorithms, SVM, CART, and RF, were used to detect the possible relationships between the DRC and the first derivative spectra. The conclusions can be summarized as follows:

(1) Significant correlations exist between spectral reflection and the DRC of the plant leaves in the visible-infrared wavelength region: the reflectance of the plant leaves increases in the visible wavelength and

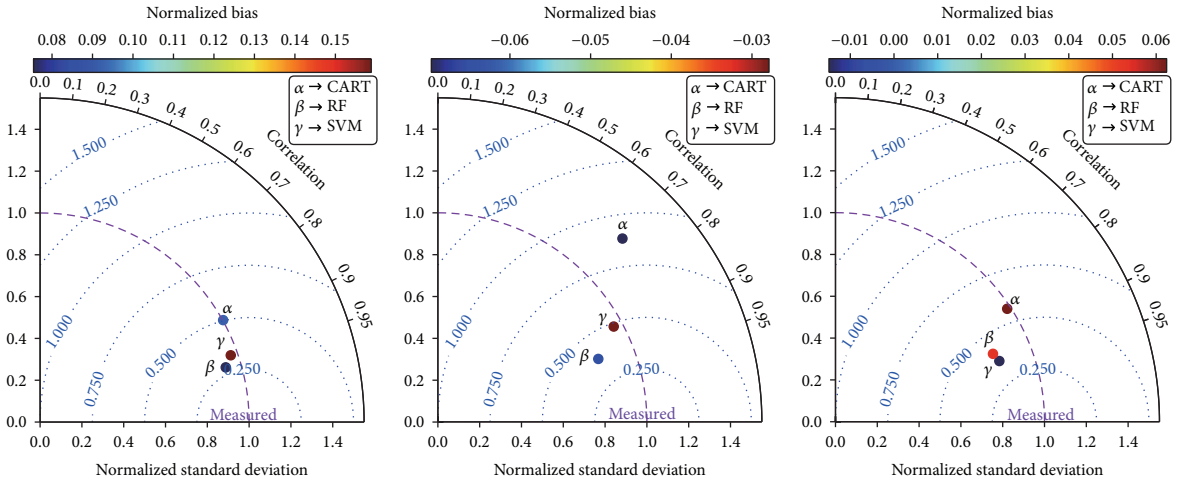

Figure 11: Taylor diagrams showing the correlations between the estimated DRC of the different models and the measured values of (a) GL, (b) LC, and (c) CF.

decreases in the infrared wavelength with increase in DRC. The characteristic of the reflectance in the shortwave infrared wavelength of the three green plants with or without dust on the leaves is random because the reflectance at these wavelengths is mainly affected by leaf water content.

(2) Spectra from 450 to 500 nm, 550 to 600 nm, 750 to 1000 nm, and 1100 to 1300 nm are recommended for detecting and estimating the DRC of leaves. However, appropriate bands differ among plants. The experiments in this study were only conducted on three types of plants; a feature selection process is recommended in any other practical applications.

(3) The RF algorithm reduced the variable dimensions well. Testing results showed that the estimated values using the SVM and RF approaches had good agreement with the measured DRC. SVM and RF, as a result, are recommended for modeling DRC based on hyperspectral data. Optimal algorithms should be determined based on different situations.

This study provided a technical approach for estimating DRC on plant leaves based on hyperspectral measurements. The validation results showed that the machine learning model proposed in this study efficiently reduced the variable dimensions and accurately estimated the DRC of different plants. Consequently, the results of this study can be applied to monitor the DRC on leaves of other plants and further be fused or integrated with other types of spectral data to measure the DRC at a regional scale based on airborne hyperspectral sensors or sensors onboard unmanned aerial vehicles (UAVs).

Acronyms

CART: Classification and regression tree
CF: *Cordyline fruticosa* (L.) A. Cheval
DRC: Dust retention content
FI: Feature importance
GL: *Ficus microcarpa* L. f. cv Golden leaves
LC: *Loropetalum chinense* (R. Br) Oliv. var. rubrum Yieh
MAE: Mean absolute error
NDVI: Normalized difference vegetation index
OOB: Out-of-bag data
R^2: Coefficient of determination
RF: Random forest
RMSE: Root mean square error
SVM: Support vector machine.

Data Availability

The hyperspectral measurements and the dust retention content data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Authors' Contributions

Wenlong Jing and Xia Zhou contributed equally to this work.

Acknowledgments

This study was jointly supported by the National Natural Science Foundation of China (41401430 and 41771380), the Guangdong Innovative and Entrepreneurial Research Team Program (2016ZT06D336), the Guangdong Academy of Sciences' Special Project of Science and Technology Development (2017GDASCX-0101 and 2018GDASCX-0904), the Forest Science and Technology Innovation in Guangdong (2015KJCX047), the Science and Technology Program of Guangzhou (201604016047), and the Science and Technology Program of Guangdong Province (2017B010117008).

References

[1] R. J. Huang, Y. Zhang, C. Bozzetti et al., "High secondary aerosol contribution to particulate pollution during haze events in China," *Nature*, vol. 514, no. 7521, pp. 218–222, 2014.

[2] S. Guo, M. Hu, M. L. Zamora et al., "Elucidating severe urban haze formation in China," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 49, pp. 17373–17378, 2014.

[3] L.-M. Mårtensson, A. Wuolo, A.-M. Fransson, and T. Emilsson, "Plant performance in living wall systems in the Scandinavian climate," *Ecological Engineering*, vol. 71, pp. 610–614, 2014.

[4] H. F. Wang, J. X. Qiu, J. Breuste, C. Ross Friedman, W. Q. Zhou, and X. K. Wang, "Variations of urban greenness across urban structural units in Beijing, China," *Urban Forestry & Urban Greening*, vol. 12, no. 4, pp. 554–561, 2013.

[5] U. Weerakkody, J. W. Dover, P. Mitchell, and K. Reiling, "Particulate matter pollution capture by leaves of seventeen living wall species with special reference to rail-traffic at a metropolitan station," *Urban Forestry & Urban Greening*, vol. 27, pp. 173–186, 2017.

[6] J. Liu, Z. Cao, S. Zou et al., "An investigation of the leaf retention capacity, efficiency and mechanism for atmospheric particulate matter of five greening tree species in Beijing, China," *Science of the Total Environment*, vol. 616-617, pp. 417–426, 2018.

[7] Y. C. Wang, "Carbon sequestration and foliar dust retention by woody plants in the greenbelts along two major Taiwan highways," *Annals of Applied Biology*, vol. 159, no. 2, pp. 244–251, 2011.

[8] A. Baidourela, Ü. Halik, T. Aishan, A. Abliz, and M. Welp, "Dust retention capacities of urban trees and the influencing factors in Aksu, Xinjiang, China," *Journal of Desert Research*, vol. 35, pp. 322–329, 2015.

[9] L. Liu, D. Guan, M. R. Peart, G. Wang, H. Zhang, and Z. Li, "The dust retention capacities of urban vegetation—a case study of Guangzhou, South China," *Environmental Science & Pollution Research*, vol. 20, no. 9, pp. 6601–6610, 2013.

[10] V. M. Kretinin and Z. M. Selyanina, "Dust retention by tree and shrub leaves and its accumulation in light chestnut soils under forest shelterbelts," *Eurasian Soil Science*, vol. 39, no. 3, pp. 334–338, 2006.

[11] J. Yu and R. Yang, "Analysis on dust retention measurement of common plant leaves in Shenyang," in *2016 3rd International Conference on Materials Science and Mechanical Engineering*, Windsor, UK, 2016.

[12] G.-Y. Chi, X.-H. Liu, S.-H. Liu, and Z.-F. Yang, "Spectral characteristics of vegetation in environment pollution monitoring," *Environmental Science & Technology*, 2005.

[13] E. Choe, F. van der Meer, F. van Ruitenbeek, H. van der Werff, B. de Smeth, and K. W. Kim, "Mapping of heavy metal pollution in stream sediments using combined geochemistry, field spectroscopy, and hyperspectral remote sensing: a case study of the Rodalquilar mining area, SE Spain," *Remote Sensing of Environment*, vol. 112, no. 7, pp. 3222–3233, 2008.

[14] E. J. Emengini, G. A. Blackburn, and J. C. Theobald, "Discrimination of plant stress caused by oil pollution and waterlogging using hyperspectral and thermal remote sensing," *Journal of Applied Remote Sensing*, vol. 7, no. 1, article 073476, 2013.

[15] K. Zhao, D. Valle, S. Popescu, X. Zhang, and B. Mallick, "Hyperspectral remote sensing of plant biochemistry using Bayesian model averaging with variable and band selection," *Remote Sensing of Environment*, vol. 132, no. 10, pp. 102–119, 2013.

[16] N. N. Luo, W. J. Zhao, and X. Yan, "Impact of dust-fall on spectral features of plant leaves," *Spectroscopy and Spectral Analysis*, vol. 33, no. 10, pp. 2715–2720, 2013.

[17] S. Shi, Z. Wu, F. Liu, and W. Fan, "Retention of atmospheric particles by local plant leaves in the mount Wutai scenic area, China," *Atmosphere*, vol. 7, no. 8, p. 104, 2016.

[18] D. N. H. Horler, M. Dockray, and J. Barber, "The red edge of plant leaf reflectance," *International Journal of Remote Sensing*, vol. 4, no. 2, pp. 273–288, 1983.

[19] S. L. Xiao and Z. X. Chen, "Assessment of effect of the dust covered the foliage on canopy reflectance," *Chinese Agricultural Science Bulletin*, vol. 23, no. 4, pp. 410–414, 2007.

[20] T. Wang, Y. Liu, H. Y. Wu, and Y. M. Zuo, "Influence of foliar dust on crop reflectance spectrum and nitrogen monitoring," *Spectroscopy and Spectral Analysis*, vol. 32, no. 7, pp. 1895–1898, 2012.

[21] H. F. Wang, N. Fang, X. Yan, F. T. Chen, Q. L. Xiong, and W. J. Zhao, "Retrieving dustfall distribution in Beijing City based on ground spectral data and remote sensing," *Spectroscopy & Spectral Analysis*, vol. 36, no. 9, pp. 2911–2918, 2016.

[22] C. Wu and X. Wang, "Research of foliar dust content estimation by reflectance spectroscopy of Euonymus japonicus Thunb," *Environmental Nanotechnology, Monitoring & Management*, vol. 5, pp. 54–61, 2016.

[23] W. Li, J. Wu, T. Chen, and D. Peng, "Hyperspectral estimation model of dust deposition content on plant leaves," *Transactions of the Chinese Society of Agricultural Engineering*, vol. 32, no. 2, pp. 180–185, 2016.

[24] H.-L. Xiao, X.-P. Chen, Q.-Y. Ling, and Z.-X. Zhou, "Analysis of dust detention capability of landscape plants and the hyperspectral remote sensing quantitative models construction of foliage dust detention," *Resources & Environment in the Yangtze Basin*, vol. S1, pp. 16–19, 2015.

[25] X. Yan, W. Shi, W. Zhao, and N. Luo, "Estimation of atmospheric dust deposition on plant leaves based on spectral features," *Spectroscopy Letters*, vol. 47, no. 7, pp. 536–542, 2014.

[26] S. Ahmad, A. Kalra, and H. Stephen, "Estimating soil moisture using remote sensing data: a machine learning approach," *Advances in Water Resources*, vol. 33, no. 1, pp. 69–80, 2010.

[27] E. Burchfield, J. J. Nay, and J. Gilligan, "Application of machine learning to the prediction of vegetation health," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLI-B2, pp. 465–469, 2016.

[28] K. Hu, A. Rahman, H. Bhrugubanda, and V. Sivaraman, "HazeEst: machine learning based metropolitan air pollution estimation from fixed and mobile sensors," *IEEE Sensors Journal*, vol. 17, no. 11, pp. 3517–3525, 2017.

[29] J. Nay, E. Burchfield, and J. Gilligan, "A machine-learning approach to forecasting remotely sensed vegetation health," *International Journal of Remote Sensing*, vol. 39, no. 6, pp. 1800–1816, 2018.

[30] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmospheric Environment*, vol. 80, no. 6, pp. 426–437, 2013.

[31] S. Heremans and J. Van Orshoven, "Machine learning methods for sub-pixel land-cover classification in the spatially heterogeneous region of Flanders (Belgium): a multi-criteria

comparison," *International Journal of Remote Sensing*, vol. 36, no. 11, pp. 2934–2962, 2015.

[32] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, and D. Roberts, "Mapping land-cover modifications over large areas: a comparison of machine learning algorithms," *Remote Sensing of Environment*, vol. 112, no. 5, pp. 2272–2283, 2008.

[33] A. M. Wahbi and S. Ebel, "The use of the first-derivative curves of absorption spectra in quantitative analysis," *Analytica Chimica Acta*, vol. 70, no. 1, pp. 57–63, 1974.

[34] J. C.-W. Chan and D. Paelinckx, "Evaluation of random forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery," *Remote Sensing of Environment*, vol. 112, no. 6, pp. 2999–3011, 2008.

[35] K. J. Archer and R. V. Kimes, "Empirical characterization of random forest variable importance measures," *Computational Statistics & Data Analysis*, vol. 52, no. 4, pp. 2249–2260, 2008.

[36] C. Strobl, A. L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: illustrations, sources and a solution," *BMC Bioinformatics*, vol. 8, no. 1, p. 25, 2007.

[37] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: a review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[38] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[39] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[40] M. Chi, R. Feng, and L. Bruzzone, "Classification of hyperspectral remote-sensing data with primal SVM for small-sized training dataset problem," *Advances in Space Research*, vol. 41, no. 11, pp. 1793–1799, 2008.

[41] G. M. Foody and A. Mathur, "The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM," *Remote Sensing of Environment*, vol. 103, no. 2, pp. 179–189, 2006.

[42] C. Huang, L. S. Davis, and J. R. G. Townshend, "An assessment of support vector machines for land cover classification," *International Journal of Remote Sensing*, vol. 23, no. 4, pp. 725–749, 2002.

[43] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone, "Classification and regression trees (CART)," *Encyclopedia of Ecology*, vol. 40, no. 3, pp. 582–588, 1984.

[44] M. A. Friedl, D. K. McIver, J. C. F. Hodges et al., "Global land cover mapping from MODIS: algorithms and early results," *Remote Sensing of Environment*, vol. 83, no. 1-2, pp. 287–302, 2002.

[45] L. Giglio, G. R. van der Werf, J. T. Randerson, G. J. Collatz, and P. Kasibhatla, "Global estimation of burned area using MODIS active fire observations," *Atmospheric Chemistry and Physics*, vol. 6, no. 4, pp. 957–974, 2006.

[46] E. Ben-Ze'ev, A. Karnieli, N. Agam, Y. Kaufman, and B. Holben, "Assessing vegetation condition in the presence of biomass burning smoke by applying the aerosol-free vegetation index (AFRI) on MODIS images," *International Journal of Remote Sensing*, vol. 27, no. 15, pp. 3203–3221, 2006.

[47] K. E. Taylor, "Summarizing multiple aspects of model performance in a single diagram," *Journal of Geophysical Research: Atmospheres*, vol. 106, no. D7, pp. 7183–7192, 2001.