Heliyon 8 (2022) e09795

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

A local model based on environmental variables clustering for estimating foliar phosphorus of rubber trees with vis-NIR spectroscopic data

Peng-Tao Guo [a,b,c,d], A-Xing Zhu [e,f,g,h,*], Zheng-Zao Cha [a,b,c,d], Mao-Fen Li [i,j], Wei Luo [a,b,c,d,**]

[a] Rubber Research Institute, Chinese Academy of Tropical Agriculture Sciences, Haikou, Hainan 571101, China
[b] Key Laboratory of Biology and Genetic Resources of Rubber Tree, Ministry of Agriculture and Rural Affairs, Haikou, Hainan 571101, China
[c] State Key Laboratory Incubation Base for Cultivation & Physiology of Tropical Crops, Haikou, Hainan 571101, China
[d] Soil and Fertilizer Research Center, Chinese Academy of Tropical Agriculture Sciences, Haikou, Hainan 571101, China
[e] School of Geography Science, Nanjing Normal University, 1 Wenyuan Road, Nanjing 210023, China
[f] Department of Geography, University of Wisconsin-Madison, 550 North Park Street, Madison, Wisconsin 53706, USA
[g] State Key Laboratory of Resources and Environmental Information System, Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Science, 11A Datun Road, Beijing 100101, China
[h] Center for Social Sciences, Southern University of Science and Technology, Shenzhen, China
[i] Institute of Scientific and Technical Information, Chinese Academy of Tropical Agriculture Sciences, Haikou, Hainan 571101, China
[j] Hainan Provincial Key Laboratory of Practical Research on Tropical Crops Information Technology, China

ARTICLE INFO

Keywords:
K-means clustering
Partial least squares regression
Hyperspectral reflectance
Regional scale
Environmental factors

ABSTRACT

Existing local models based on multiple environmental variables clustering (LM-MEVC) treat the influences of environmental factors on leaf phosphorus concentration (LPC) of rubber trees (*Hevea brasiliensis*) equally when grouping samples. In fact, the effects that environmental factors assert on LPC are different. So, environmental factors need to be treated differently so that the different effects can be taken into consideration when dividing samples into clusters or groups. According to this basic idea, a local model based on weighted environmental variables clustering (LM-WEVC) was developed. This approach consists of four steps. Firstly, the most important environmental variables that influence LPC were selected. Then, the weights of the selected environmental variables were determined. In the following, the selected environmental variables were weighted and used as clustering variables to group samples. Finally, within each cluster or group of samples, an estimation model was established. In order to verify its effectiveness in predicting LPC of rubber trees, the proposed method was applied to a case study in Hainan Island, China. Rubber tree (cultivar CATAS-7-33-97) leaf samples were collected from three different sampling periods. Spectral reflectance of the collected leaf samples was measured using an ASD spectroradiometer, FieldSpec 3. Leaf samples collected from the three different sampling periods were used separately to test LM-WEVC. Coefficient of determination (R^2), root mean squared error (RMSE), and ratio of prediction deviation (RPD) were employed as evaluation criterion. Performance of LM-WEVC was compared with that of the existing LM-MEVC. Results indicated that for the three sampling periods, the prediction accuracies of LM-WEVC were always higher than those of LM-MEVC. The values of R^2 and RPD for LM-WEVC were increased by 8.15%–36.68%, and by 11.33%–59.40% respectively, while values of RMSE were reduced by 9.09%–37.5%, compared with those for LM-MEVC. These results demonstrate that LM-WEVC was effective in estimating LPC of rubber trees, and also confirmed our hypothesis that environmental factors unequally influenced LPC of rubber trees.

1. Introduction

Rubber trees (*Hevea brasiliensis*) are the main source of natural rubber (van Beilen and Poirier, 2007). In rubber tree, phosphorus is involved in the process of natural rubber synthesis, so phosphorus is closely related to the yield of natural rubber (Guo et al., 2018). Leaf phosphorus concentration (LPC) is a good indicator of phosphorus nutrition status of rubber tree. Therefore, acquiring reliable LPC is the premise for guiding

* Corresponding author.
** Corresponding author.
E-mail addresses: azhu@wisc.edu (A.-X. Zhu), rkylw@163.com (W. Luo).

https://doi.org/10.1016/j.heliyon.2022.e09795
Received 14 December 2021; Received in revised form 12 April 2022; Accepted 21 June 2022

farmers to properly apply phosphate fertilizer to rubber trees, which is important for ensuring the healthy growth of rubber trees and thus maintaining the high yield of natural rubber (Lu and He, 1982). Hyperspectral model has the potential of accurately and rapidly estimating LPC of rubber seedlings cultivated in the greenhouse (Guo et al., 2016) or rubber trees planted at the field scale (Guo et al., 2018). However, this type of models is still limited in predicting LPC of rubber trees grown in large area due to the high variations in LPC and leaf spectra at regional scale (Asner et al., 2014). In order to reduce the variation in LPC and the spectra, a locally modeling approach was developed (Araújo et al., 2014; Gogé et al., 2012; Shi et al., 2014). This approach divides the whole dataset into a few clusters or groups according to the similarity of certain properties of samples. Then, a local model is built for each cluster or group. The prediction accuracy of the local model is generally higher than that of the commonly used global model (GM) at regional scale (Song et al., 2020a). Thus, this approach is receiving more and more attention at present.

The local modeling approach can be classified into three categories. The first one is the local model based on spectral clustering (LM-SC). This method divides the samples into a number of clusters or groups according to the spectral similarity of the target variable. Then, the relationship between target variable and spectra is modeled for each cluster or group (Liu et al., 2019; Ogen et al., 2019; Shi et al., 2014). For example, Shi et al. (2014) employed visible-near infrared spectroscopy (350–2500 nm) as input variables to divide the 1581 soil samples collected from different soil types of China into 5 classes, and within each class a local model was built for predicting soil organic matter content. This method can also be applied to estimate leaf nutrients of plants. However, this method has one limitation that it is opt to classify samples with similar spectral characteristics but samples with markedly different target attributes (such as soil organic matter contents) could be grouped into a same cluster (Castro-Esau et al., 2006; Zhang et al., 2018), which would result in misclassification of samples and consequently reduce the prediction accuracies of local models.

The second one is the local model based on single environment variable clustering (LM-SEVC). This method uses only one influencing environment factor as input variable for clustering, and partitions the samples into several clusters or groups in terms of similarity in the selected environment factor. Then, a local model is built for each cluster or group (Bao et al., 2020; Moura-Bueno et al., 2019). This method overcomes the limitation that exists in LM-SC because environmental factor rather than spectra was used as input variable for clustering. Bao et al. (2020) selected soil types as clustering variable to classify collected soil samples into a number of groups. Finally, within each group, a local model was constructed for predicting soil organic matter content. This method could also be used to predict leaf nutrients of plants. However, this method only considers the influence of a single environment factor on target variable while ignores impacts of other environment variables when clustering samples into clusters or groups. Therefore, this method is still insufficient to classify samples into proper groups, which could adversely impact the predictive abilities of the local models.

The third one is the local model based on multiple or compound environmental variables clustering (LM-MEVC). This method adopts a number of environmental factors as input variables for clustering, and classifies samples into several clusters or groups on basis of similarity in the employed environmental factors. Then, a local model is fitted for each cluster or group (Moura-Bueno et al., 2020; Song et al., 2020b). For example, Moura-Bueno et al. (2020) selected physiographic regions as the input variable for clustering. Physiographic regions are compound environmental variables which are determined by considering the combined effects of climate and parent materials. Soil samples were divided into three groups by physiographic regions, and then a local model of predicting soil organic carbon was established for each group. This method could also be employed to estimate leaf nutrients of plants. However, this method treats the influences of different environmental factors on target variable equally, which does not accord with the actual situation. Therefore, this method still has defect in sample clustering, which would negatively affect the predictive ability of the local model.

The discussion above indicates that LM-MEVC is seemingly the most effective method for estimating leaf nutrients of plants at present. However, this method still has an obvious weakness that it equally considers the influences of different environmental factors on target variable when dividing samples into clusters or groups. In fact, influences of various environmental factors on target variable are different (Said et al., 2021). For example, Asner et al. (2017) found that geologic substrate and elevation were the main factors affecting leaf nutrients of tropical forests, whereas topographic slope, local hydrology, and solar insolation were the secondary factors. Therefore, differences in effects of environmental factors on leaf nutrients should be taken into consideration when using environmental factors as input variables for clustering. Only in this way can samples be accurately clustered, and thus can the optimal local model be obtained.

The aim of this study was to develop a new local modeling approach to estimate LPC of rubber trees at regional scale. This new approach named local model based on weighted environmental variables clustering (LM-WEVC). Compared with the existing LM-MEVC, the novelty of this new approach is that the differences in impacts of various environmental factors on target variable are accounted for when clustering samples into groups. Thus, it would be expected more accurate classification results, which is vital for improving the prediction accuracy of local models. In the next section of this paper, a detailed description of this approach is presented, and then this proposed approach was applied to a case study to evaluate its effectiveness in estimating LPC of rubber trees at regional scale. Performance of this approach was compared to that of LM-MEVC.

2. Methods

2.1. Basic idea and overall design

Environmental factors impose different impacts on target variable (LPC of rubber trees in this study). So, the different effects of environmental factors on target variable should be taken into account when clustering samples into clusters or groups. According to this basic idea, this paper develops and proposes a new approach named local model based on weighted environmental variables clustering (LM-WEVC). This proposed approach mainly consists of four steps: (1) selection of dominant environmental factors influencing target variable; (2) determination of weights of different environmental factors; (3) classification of samples using weighted environmental factors; (4) construction of local model for each cluster or group.

2.2. Selection of dominant environmental factors influencing target variable

In this study, maximal information coefficient (MIC) is used to select the dominant environmental factors that have impacts on the target variable. MIC is a measure of dependence of relationship between two variables (Reshef et al., 2011). The MIC cannot only capture linear relationship but also no-linear relationship between pairwise variables. At the same time, it provides a score that measures the strength of the relationship. The score is roughly equivalent to the coefficient of determination (R^2) of the data relevant to the regression function (Reshef et al., 2011). The formula of MIC is listed as following:

$$\mathrm{MIC}(D) = \max_{XY<B(n)} \frac{I^*(D,X,Y)}{\log(\min X, Y)} \quad (1)$$

where D denotes the sample data domain which is partitioned into X^*Y grids along the pairwise variable x and y; $I^*(D, X, Y)$ indicates the induced mutual information in the domain D with X^*Y grids; $B(n)$ is a function of

sample size n, and it equals $n^{0.6}$. Calculation of MIC was performed using R software with minerva package.

Figure 1 shows the process of how to use MIC to select important environmental factors that influence the target variable. In the first step, MIC between target variable and environmental factors is calculated and a significance test ($p < 0.05$) is carried out for MIC. If MIC passes the significance test, the corresponding environmental factor will be saved to variable set 0 (Set 0); otherwise, the environmental factor will be disregarded. In the second step, the environmental factor (EVs) in Set 0 that has the strongest correlation with the target variable will be selected and saved to the variable selected set (referred to as Selected). In the third step, MIC between the EVs and the other environmental factors is also calculated, respectively. If the value of MIC is smaller than 0.64 (i.e., correlation coefficient (r) is less than 0.8), then it can be assumed that there is no collinearity between the EVs and the corresponding environmental factor (Farrar and Glauber, 1967). So, the corresponding environmental factor will be moved to variable set 1. Otherwise, the environmental factor should be ignored (removed from Set 0). Once Set 0 is empty, move all environmental variables from Set 1 to Set 0, and then repeat the second and the third step until Set 1 is empty. Once the process is completed, the variables in Selected are the dominant variables to be used.

2.3. Determination of weights of different environmental factors

Influencing weights were calculated using random forest (RF). RF is a machine learning algorithm which is developed on basis of ensemble learning (Breiman, 2001). RF employs the sampling with replacement method to derive quantities of sample sets from the original sample set. Then, these sample sets are used to generate a large number of decision trees. Each decision tree votes on the result and the one with the most votes is determined as the final classification or prediction result. RF not only can predict target variables, but also can provide a measure of importance (IncMSE) for auxiliary variables (e.g., environmental factors). The higher the value, the more important the auxiliary variable is. Equation of IncMSE is as follows:

$$\mathrm{IncMSE}_i = \frac{1}{n}\sum_{i=1}^{n}\left(e_i^{'} - e_i\right) \tag{2}$$

where IncMSE_i is the measure of importance for the ith auxiliary variable, e_i indicates the out-of-bag error of the ith single decision tree, $e_i^{'}$ represents the out-of-bag error of the ith decision tree recalculated after adding noise to a certain auxiliary variable, and n denotes the number of decision trees.

IncMSE_i was then used to calculate the weights (W) of environmental factors. Formula of W is listed as below:

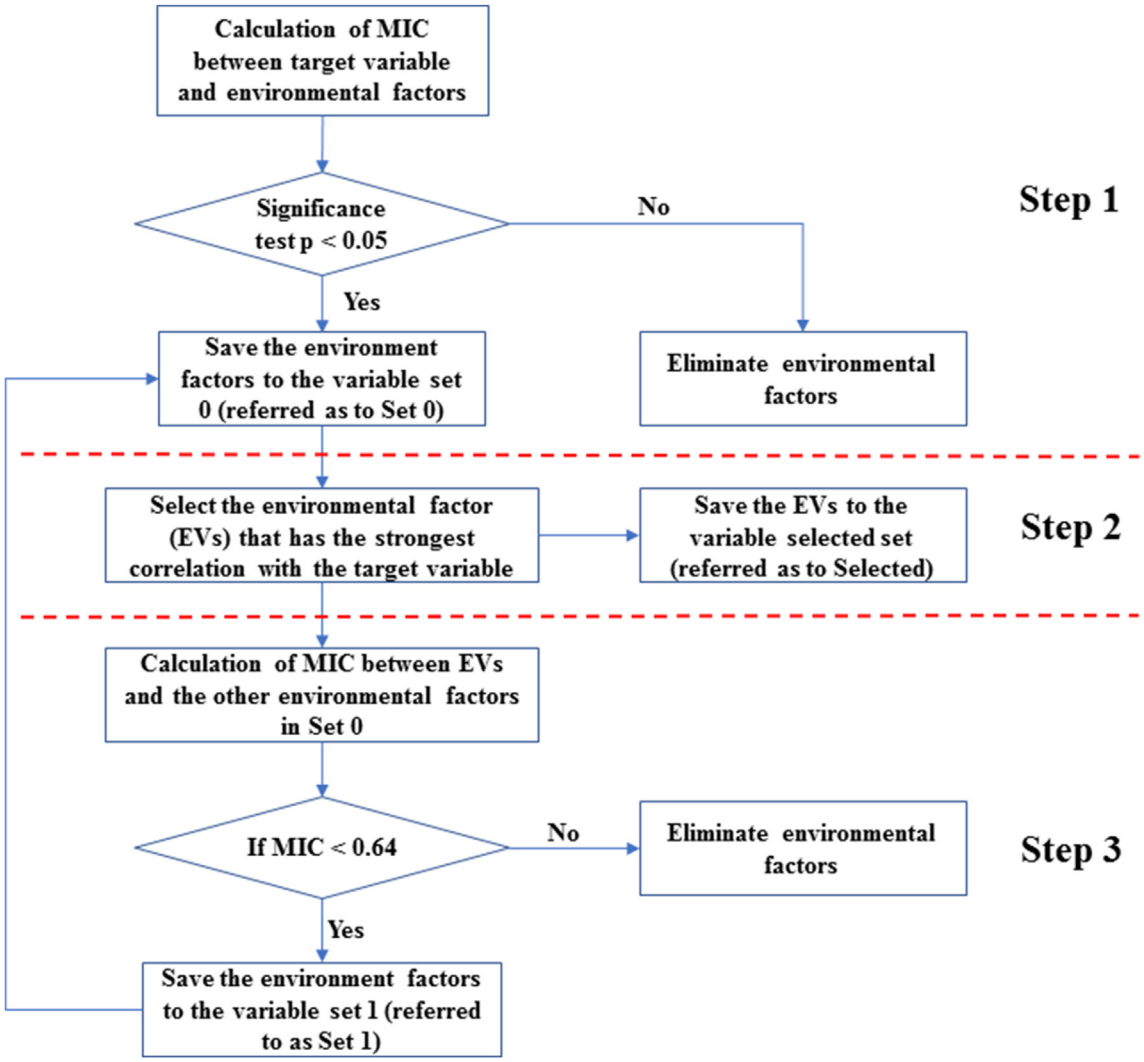

Figure 1. The flowchart of selection of main environmental factors influencing target variable.

$$W_i = \frac{IncMSE_i}{\sum_{i=1}^{n} IncMSE_i} \quad (3)$$

where W_i represents the influencing weight on the target variable for the ith environmental factor and n is the number of environmental factors.

RF was implemented using R software with the randomForest package, and IncMSE was acquired by using Eq. (2).

2.4. Classification of samples using weighted environmental factors

Each environmental factor was weighted using the corresponding weight (W_i), and then the weighted environmental factors were employed as input variables for clustering. In order to perform the clustering and classify samples into a few clusters or groups, the K-means clustering method was adopted. The basic steps of the K-means clustering are that K initial centroids are randomly generated at first, and then each sample is allocated to the cluster represented by the centroid closest to the sample. After the allocation of every sample, the centroid of each cluster is recalculated according to all samples within each cluster. In the following, allocation of samples and recalculation of cluster centroid are repeated and would not stop until changes in cluster centroids are small or a predefined number of iterations is reached (Hartigan, 1975). Distance between sample and cluster centroid is calculated by the Euclidean distance, and the equation is as follows:

$$D(X_i, C_j) = \sqrt{\sum_{t=1}^{m} (X_{it} - C_{jt})^2} \quad (4)$$

where $D(X_i, C_j)$ is the Euclidean distance between the ith sample (X_i) and the jth cluster centroid (C_j), X_{it} represents the tth property of the sample X_i, and C_{jt} indicates the tth property of the cluster centroid C_j.

The cluster centroid C_j is a collection of mean values of all properties of samples within the jth cluster. The formula of C_j is listed as below:

$$C_j = \frac{\sum_{X_i \in S_j} X_i}{|S_j|} \quad (5)$$

where S_j represents the jth cluster, $|S_j|$ denotes the number of samples in cluster S_j, and X_i is the ith sample of the cluster S_j.

The K-means clustering method can cluster the samples into K clusters (or groups) according to the predefined parameter K. However, to date, there are no universal rules about how to determine the optimal value for K. In the current study, the elbow method was used to determine the optimal value for K. This method assumes that the degree of aggregation of each cluster will gradually increase with the increase of the parameter K, while the sum of squared error (SSE) of the distance between each sample and its cluster centroid in all clusters will naturally decrease. Based on this assumption, it can be expected that when K is far less than the optimal value, the increase in K will significantly increase the degree of aggregation of each cluster, and thus the value of SSE would markedly decrease. However, when K arrives at the optimal point, further increase in K would not dramatically increase the degree of aggregation, and thus the corresponding SSE will also decline slowly. Therefore, the relationship between K and SSE is in the shape of an elbow, and the turning point of the elbow is just corresponding to the optimal K value (IBM Corp, 2011). Formula of SSE is listed as follows:

$$SSE = \sum_{j=1}^{K} \sum_{X \in S_j} (X - C_j)^2 \quad (6)$$

where K is the number of clusters, S_j represents the jth cluster, C_j indicates the centroid of S_j, and X denotes the samples belong to S_j.

The K-means clustering algorithm was performed in Matlab 2016a with the kmeans function.

2.5. Construction of local model for each cluster or group

Phosphorus is related to the formation of pigment (Al-Abbas et al., 1974), protein (Zhang et al., 2013), starch (Okita, 1992), cellulose, and lignin (Islam et al., 1999) in leaves of plants. These biochemical substances absorb light of specific wavelengths that can cause variation in leaf reflectance (Curran, 1989; Kumar et al., 2002). Thus, there are close relations between LPC and leaf reflectance. Based on these relations, LPC can be inferred with leaf reflectance. Several studies (Gao et al., 2019; Knox et al. 2011; Mutanga and Kumar, 2007; Ramoelo et al., 2013) reported that the relations between LPC and leaf reflectance were not linear. So, a commonly used non-linear modelling approach back-propagation neural network (BPNN) was employed to model the relations between LPC and leaf reflectance for each cluster or group in this study.

The BPNN model consisted of three layers. They were input layer, hidden layer, and output layer respectively. The input layer contained leaf reflectance as auxiliary variable. The leaf reflectance here referred to those a few bands that contribute most to the explanation of variance in LPC instead of the full spectrum. The reason why used a few important bands as auxiliary variable was that if the full spectrum (2150 bands) was used as auxiliary variable, the structure of the BPNN model would be extremely complex and the model training process could be incredibly time-consuming (Zou et al., 2010). At the same time, the developed BPNN model with full spectrum as auxiliary variable would be inevitably overfitted. These important bands were selected using RF. The RF could provide a measure of importance for each spectral band. According to the measure of importance, the top 10 most important bands were used as auxiliary variable. The hidden layer was composed of a number of neurons which play an important role in controlling the learning ability of the BPNN. If the number of neurons was too small, the developed BPNN would be insufficient to capture the relations between LPC and leaf reflectance. In contrast, if the number was large, the developed BPNN would be overfitted and its generalization ability would be poor (Ito et al., 2008). Therefore, the determination of the number of neurons should be proper. In this study, 5 neurons were determined for the hidden layer according to the result of our previous research (Guo et al., 2018). The output layer contained the estimated values of LPC. The more detailed information about BPNN could be found in Guo et al. (2013). The BPNN was implemented in Matlab 2016a.

After the construction of local model for each cluster or group, a discriminant analysis model was also developed. The discriminant analysis model was used to assign the test samples to one of the clusters or groups. Then, the value of the test sample can be predicted by using the corresponding local model. The discriminant analysis model was established on the basis of the cluster results of the training set and with the classify function in Matlab 2016a.

3. Case study

3.1. Study area

To verify the validity of LM-WEVC in estimating LPC of rubber trees at regional scale, this approach was applied to Hainan Island, China. Within this island, nine sites across environmental gradients were selected (Figure 2) for leaf sample collection. The elevation of these sites ranges from 64 to 226 m, mean annual temperature from 23.7 to 24.1 °C, and precipitation from 925 to 1773 mm. Although soil types of these nine sites were the same (classified as Udic Ferralsol (sub-order) in the World Reference Base for Soil Resource (FAO, 1998)), their properties were significantly different from each other due to the parent materials from which the soils developed were diverse (Guo et al., 2015). These parent materials were granites, basalts, metamorphic rocks, neritic sediment, and sandshale, respectively.

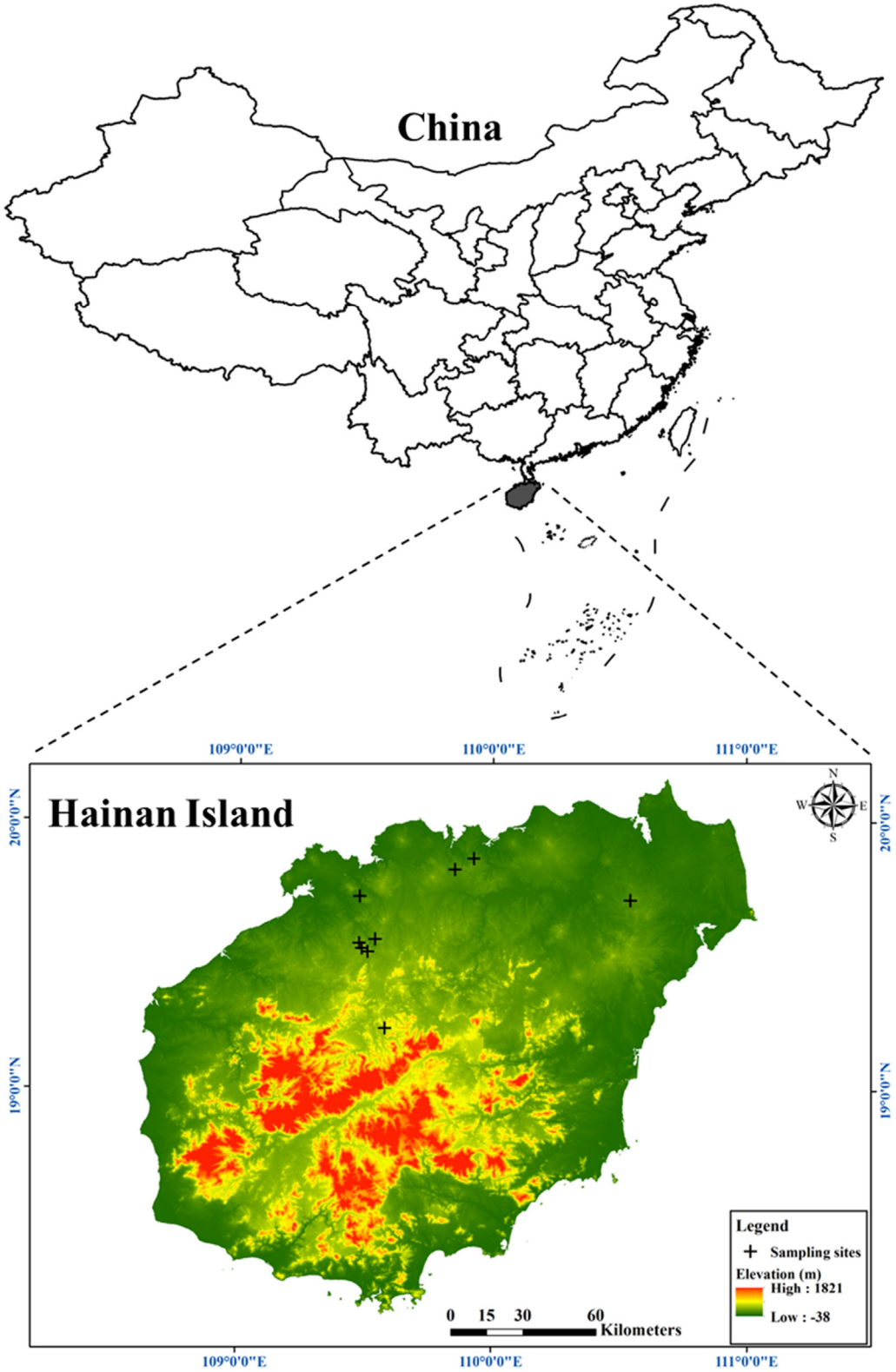

Figure 2. Location of the study area and distribution of sampling sites.

3.2. Data sources

3.2.1. Leaf samples

Rubber trees exhibit obvious seasonal variations in LPC (Guo et al., 2018). From April to June, rubber trees put forth buds and leaves, and a large amount of nutrients is transferred from root and trunk to branches and leaves in order to improve the growth of leaves. So, LPC during this period is the highest of the year. From July to September, the rubber tree leaves have grown up and are in a relatively stable development stage. LPC of rubber trees is in the intermediate level of the year. During the period of October to December, the leaves of the rubber tree gradually age, and nutrients are transferred from the leaves to the trunk and other

parts of the tree. Thus, LPC is the lowest of the year. In order to widen the range of LPC, leaf sample collections were carried out three times in 2018 according to the three periods mentioned above. At each sampling site, the field was divided into 14–20 plots in terms of their own coverage. The size of each plot was 12 m × 21 m (rubber trees were planted with spacing of 3 m × 7 m). Within each plot, five rubber trees were randomly selected and two healthy leaves were collected from the lower crown of each tree. Thus, a total of ten leaves were obtained for each plot and these leaves were mixed together as one composite sample. Each composite leaf sample was placed into a polyethylene bag to maintain moisture. Sample number, name and coordinate of sampling sites, and planting years were recorded on the surface of the bags. Then, these bags were put into one white Styrofoam plastic box which contains ice. In total, 540 composite leaf samples were collected from the 9 sites (one hundred and eighty composite leaf samples were obtained for each sampling period).

Leaf samples were immediately taken back to the dark room for spectral measurement once the sample is obtained in the field. An ASD spectroradiometer, FieldSpec 3 (Analytical Spectral Devices, Boulder, CO, USA) was employed to measure spectral reflectance of leaf samples. Spectral range of this spectroradiometer is from 350 to 2500 nm. Within the range of 350–1000 nm, the sampling interval and the spectral resolution are 1.4 and 3 nm respectively, while in the range of 1000–2500 nm, those are 2 and 10 nm respectively. A vegetation probe with a leaf clip was used to scan surfaces of leaf samples. The vegetation probe was connected to the spectroradiometer by fiber optics. A built-in halogen lamp (3.825 V, 4.05 W) was set in the vegetation probe. This halogen lamp provides illumination for measuring. Prior to each measurement, reflectance spectra were calibrated against a white Spectralon panel. Then, leaves were put into the leaf clip in sequence and the middle left and middle right locations were scanned. Each location was scanned for three times. So, 6 readings were recorded for one leaf, and thus a total of 60 readings were reserved for one composite sample. The 60 readings were averaged to obtain one mean value of the spectral reflectance (SR), and the mean value was used as the final spectral data for the composite leaf sample (Figure 3).

In order to further investigate the impacts of different ways of processing spectral data on LM-WEVC in estimating LPC of rubber trees, the other two commonly used spectral data were also calculated and employed as input variables for the LM-WEVC. These two spectral data were continuum removed reflectance (CR) (Clark and Roush, 1984) and the continuum-removed derivative reflectance (CRDR) (Mutanga et al., 2004). Equation of CR was listed as below:

$$\mathrm{CR} = \frac{R}{R_c} \tag{7}$$

where R and R_c represent the spectral reflectance and the continuum line, respectively. Figure 4 shows the results of CR.

The CRDR was calculated on the basis of the CR. CRDR was obtained by applying the first difference transformation to the CR results. Equation of CRDR was listed as follows:

$$\mathrm{CRDR} = \mathrm{CR}' \tag{8}$$

where $'$ indicates the first difference transformation. Figure 5 presents the result of CRDR.

When the measurement of leaf spectra was completed, leaf samples were taken to the laboratory for chemical analysis. Leaves were put into the oven and dried at 105 °C for 30 min, and then at 70 °C for 8 h. The dried leaves were grinded into powder with a mortar. Then, the powder was passed through a 1 mm screen, and was digested by a mixture of concentrated H_2SO_4 and 30% H_2O_2. Finally, leaf phosphorus concentration (%) was determined using the molybdenum-antimony colorimetric method.

3.2.2. Environmental factors

Climate, parent materials, and topography can impose influences on leaf nutrients of tropical forests (Asner et al., 2009, 2016), so information about these environmental factors were collected in the current study. Climate factors including 19 bioclimatic variables (Fick and Hijmans, 2017) were downloaded from the WorldClim website (https://www.worldclim.org/data/index.html). These bioclimatic variables are in a GRID format with a spatial resolution of 1 km. Parent materials were extracted from an existing digitized geology map of Hainan Island at a cartographic scale of 1: 500,000. There were five parent materials underlain the 9 sampling sites. They were granites, basalts, metamorphic rocks, neritic sediment and sandshale, respectively. Topographic variables including elevation, slope, sine of aspect, and cosine of aspect were also employed in this study. They were calculated from the SRTM DEM with a spatial resolution of 90 m using the easyGC which is available online (http://www.easygeoc.net:8090/) (Zhu et al., 2021). The SRTM DEM was downloaded from the Geospatial Data Cloud website (https://www.gscloud.cn/search). All these environmental factors were listed in Table 1.

3.3. Experimental design

To evaluate the effectiveness of the LM-WEVC in predicting LPC of rubber trees at regional scale, the performance of this approach was compared with that of LM-MEVC. The reason for selecting LM-MEVC for comparison was that LM-MEVC gave equal weights for environmental factors when clustering samples into groups, whereas LM-WEVC put unequal weights. Comparison was carried out for each sampling period separately. This means that leaf samples collected from each sampling period were treated as an independent dataset, respectively. For each dataset, leaf samples were divided into training set (140 leaf samples) and test set (40 leaf samples) using the K–S algorithm (Kennard and Stone, 1969), respectively. Statistical results of the training sets and the test sets of the three datasets are presented in Figure 6.

When building these estimation models, the spectral reflectance (SR) was used as input variable. Prediction accuracies of these models were

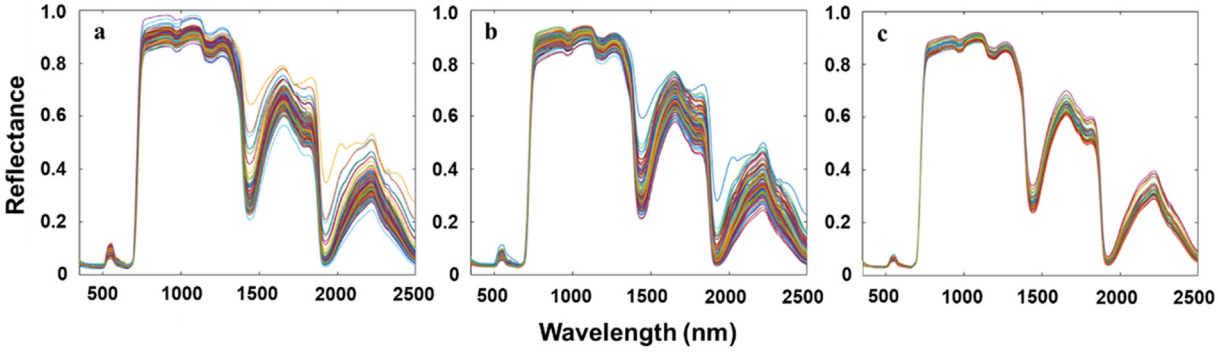

Figure 3. Spectral reflectance of rubber tree leaf samples collected for the three sampling periods: **a** from April to June, **b** from July to September, and **c** from October to December.

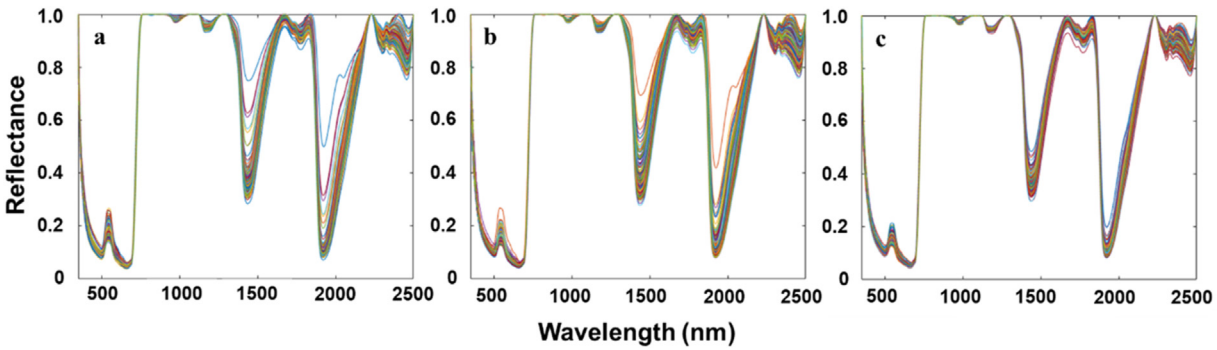

Figure 4. Continuum removed reflectance of rubber tree leaf samples collected for the three sampling periods: **a** from April to June, **b** from July to September, and **c** from October to December.

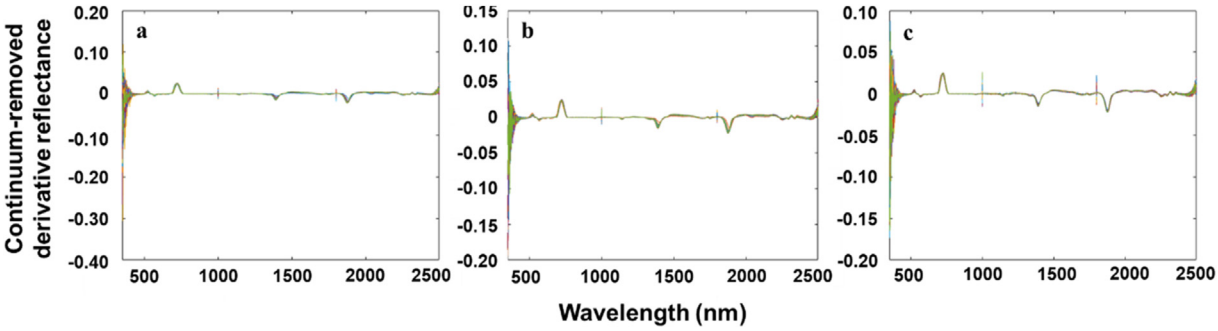

Figure 5. Continuum-removed derivative reflectance of rubber tree leaf samples collected for the three sampling periods: **a** from April to June, **b** from July to September, and **c** from October to December.

evaluated by coefficient of determination (R^2), root mean squared error (RMSE), and ratio of prediction deviation (RPD). Formulas of these indexes are listed as following:

Table 1. Environmental variables used in the current study.

Environmental variables	Abbreviation of variables
Parent Materials	par
Annual Mean Temperature	bio1
Mean Diurnal Range	bio2
Isothermality	bio3
Temperature Seasonality	bio4
Max Temperature of Warmest Month	bio5
Min Temperature of Coldest Month	bio6
Temperature Annual Range	bio7
Mean Temperature of Wettest Quarter	bio8
Mean Temperature of Driest Quarter	bio9
Mean Temperature of Warmest Quarter	bio10
Mean Temperature of Coldest Quarter	bio11
Annual Precipitation	bio12
Precipitation of Wettest Month	bio13
Precipitation of Driest Month	bio14
Precipitation Seasonality	bio15
Precipitation of Wettest Quarter	bio16
Precipitation of Driest Quarter	bio17
Precipitation of Warmest Quarter	bio18
Precipitation of Coldest Quarter	bio19
Elevation	ele
Slope	slo
Sine of aspect	Sinasp
Cosine of aspect	Cosasp

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y})^2} \tag{9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{10}$$

$$RPD = \frac{SD}{RMSE} \tag{11}$$

where n is the number of leaf samples, y_i and $\hat{y}_i$ represent the measured and predicted value of LPC for the ith sample, $\bar{y}$ indicates the mean value of the measured LPC, and SD denotes the standard deviation of the measured LPC.

R^2 indicates the correlation between the predicted and measured LPC. The higher the R^2, the stronger is the correlation. RMSE measures the difference between the predicted and measured LPC. A smaller RMSE indicates the estimation is reliable. RPD assesses the performance of a prediction model. The larger value of RPD, the better is the model performance. If RPD <1.4, the prediction accuracy of the model is unacceptable; if 1.4 < RPD <2.0, the prediction accuracy is acceptable but needs improvement; if RPD >2.0, the prediction accuracy is high (Dong et al., 2022; Li et al., 2018; Wang et al., 2013, 2021). Therefore, a good model should have higher R^2 and RPD, but lower RMSE.

4. Results

4.1. Selected environmental variables and their impacts on LPC

Table 2 lists selected environmental variables and their effects on LPC of rubber trees. It can be seen that parent materials, slope and aspect (sine of aspect and cosine of aspect) were the main factors impacting LPC of rubber trees for the three sampling periods. Among these selected environmental variables, parent materials were the most influencing factor

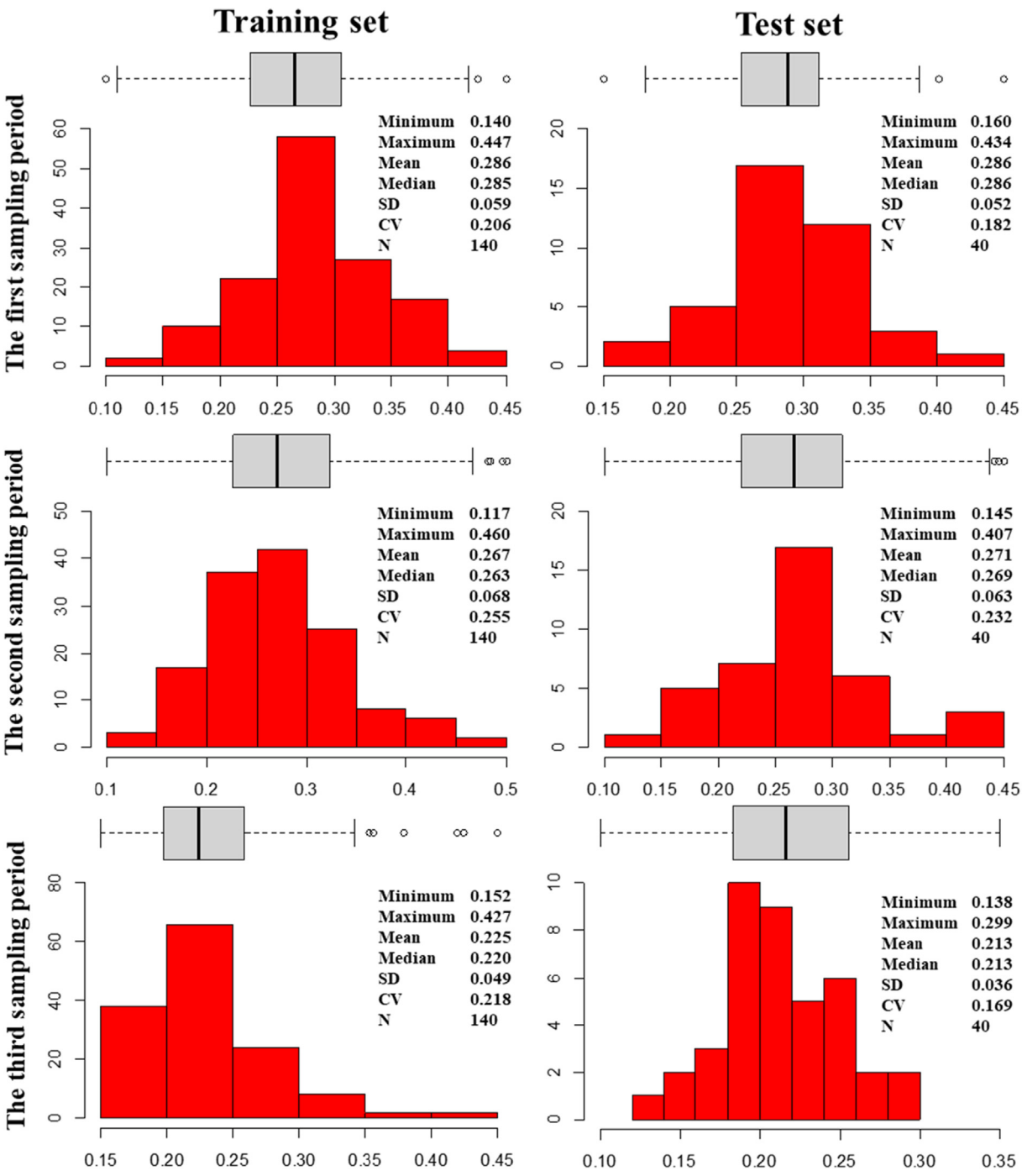

Figure 6. Statistical results of rubber tree leaf phosphorus concentration for leaf samples collected from different sampling periods.

since its weights on LPC were always the largest for the three sampling periods. This finding was consistent with that reported by He et al. (1991). In their study, they found that parent materials significantly influenced soil fertility and the nutrient status of rubber trees. Slope and aspect were the secondary influencing factors. This finding was also consistent with the situation of the study area. In Hainan Island, rubber trees are mostly planted in hilly and mountain area. Within these regions, slope and aspect play an important role in controlling redistribution of soil, water and temperature which could further impact LPC of rubber trees. In addition to these selected environmental variables, bioclimatic variables also exerted some influences on LPC of rubber trees. However, MICs between parent materials and bioclimatic variables were larger than 0.64, which indicated that there were collinearities between parent materials and bioclimatic variables. Thus, bioclimatic factors were disregarded.

4.2. Clustering with the weighted environmental variables

The selected environmental variables were weighted using the calculated weights (Table 2). Then, the weighted environmental variables were used as input variables for the K-mean clustering analysis. The clustering results are shown in Figure 7. It could be seen that 6, 5 and 5 clusters were determined for the leaf samples collected from the first, the second and the third sampling period respectively.

Table 2. Selected environmental variables and their effects on foliar phosphorus of rubber trees.

Leaf sampling periods	Environmental variables	IncMSE (%)	Weight
First sampling period	par	72.38	0.52
	slo	30.60	0.22
	Cosasp	35.21	0.25
Second sampling period	par	77.49	0.58
	slo	25.65	0.19
	Sinasp	31.37	0.23
Third sampling period	par	42.37	0.37
	slo	23.37	0.21
	Cosasp	22.15	0.19
	Sinasp	26.09	0.23

par, slo, Sinasp and Cosasp represent parent materials, slope, sine of aspect and cosine of aspect respectively; IncMSE indicates a measure of importance of environmental factors on foliar phosphorus.

4.3. Important bands of the clusters

RF was employed to select important bands for each cluster. Table 3 lists the selected bands for the clusters of different sampling periods. Numbers in bold denoted bands related to the known absorption feature while those in normal were not associated with known absorption feature. As can be seen, at least one band was related with the known absorption feature except bands selected from SR for the cluster 1 of the second sampling period. These bands were mainly related to chlorophyll (e.g., 416, 500, 545, 556, 674 nm), protein (e.g., 1011, 1728, 2120, 2170, 2303 nm), starch (e.g., 1459, 1535, 1544, 2319, 2254 nm), cellulose (e.g., 1482, 1728, 1819, 2264, 2338 nm), and lignin (e.g., 1117, 1119, 1206 1416, 1691 nm) (Curran, 1989; Kumar et al., 2002). These results were in agreement with findings by Guo et al. (2018) and Ramoelo et al. (2011, 2013). The mechanism involved in the absorption of radiation by chlorophylls is electron transitions while that by protein, starch, lignin, and cellulose is bond vibration. The bond vibration mechanisms associated with proteins are N–H, C–H, and C=O, while with starch, lignin, and cellulose are O–H and C–H (Kumar et al., 2002).

4.4. Comparison results over the different sampling periods

4.4.1. For the first sampling period

Figure 8 presents prediction accuracies of the LM-WEVC and LM-MEVC based on the test set from the first sampling period. It can be seen that no matter which spectral variable was used, data points of the LM-WEVC were always much closer to the 1:1 reference line than those of the LM-MEVC. At the same time, values of R^2 (0.758, 0.849, and 0.820) for LM-WEVC were much higher than those for LM-MEVC (0.574, 0.638, and 0.717), whereas values of RMSE (0.025, 0.020, and 0.022) for LM-WEVC were much smaller than those (0.034, 0.032, and 0.028) for LM-MEVC. Most importantly, values of RPD for LM-WEVC were all larger than 2.0 (indicating the model was with good performance) whereas those for LM-MEVC were all in the range between 1.4 and 2.0 (indicating the performance of the model was acceptable but needs improvement). These results indicated that LM-WEVC performed much better than LM-MEVC in estimating LPC of rubber trees at regional scale for the first sampling period.

4.4.2. For the second sampling period

Figure 9 gives prediction accuracies of the LM-WEVC and LM-MEVC on the basis of the test set from the second sampling period. As can be seen, data points of LM-WEVC and LM-MEVC with high LPC values (higher than 0.30) were both more discrete around the 1:1 line than those with lower LPC values, which indicated that these two models both had limitations in estimating LPC with high values. However, the ability of LM-WEVC was still stronger in estimating LPC with high values than that of LM-MEVC, since the data points of LM-WEVC with high LPC values were closer to the 1:1 line than those of LM-MEVC. At the same time, values of R^2 (0.811, 0.771, and 0.823) for LM-WEVC were higher than those for LM-MEVC (0.724, 0.677, and 0.761), while values of RMSE (0.030, 0.031, and 0.027) for LM-WEVC were smaller than those (0.033, 0.035, and 0.033) for LM-MEVC. Furthermore, values of RPD for LM-WEVC were all larger than 2.0 (indicating the model had good performance) whereas those for LM-MEVC were all in the range between 1.4 and 2.0 (indicating the performance of the model was acceptable but needs improvement). These results demonstrated that LM-WEVC also outperformed LM-MEVC in predicting LPC of rubber trees at regional scale for the second sampling period.

4.4.3. For the third sampling period

Figure 10 displays the prediction accuracies of LM-WEVC and LM-MEVC according to the test set from the third sampling period. It could be seen that data points of LM-WEVC were closer to 1:1 line than those of LM-MEVC. At the same time, values of R^2 (0.585, 0.628, and 0.679) for LM-WEVC were higher than those for LM-MEVC (0.428, 0.524, and 0.523), while values of RMSE (0.023, 0.022, and 0.020) for LM-WEVC were smaller than those (0.027, 0.025, and 0.025) for LM-MEVC. Besides, values of RPD for LM-WEVC were all in the range between 1.40 and 2.00 (indicating the performance of the model was acceptable but needs improvement) whereas those for LM-MEVC were only two (the cases of CR and CRDR) in the same range. In the case of SR, value of RPD for LM-MEVC was lower than 1.40, indicating the prediction accuracy of the model was unacceptable. These results again confirmed that LM-WEVC was superior to LM-MEVC in estimating LPC of rubber trees at regional scale.

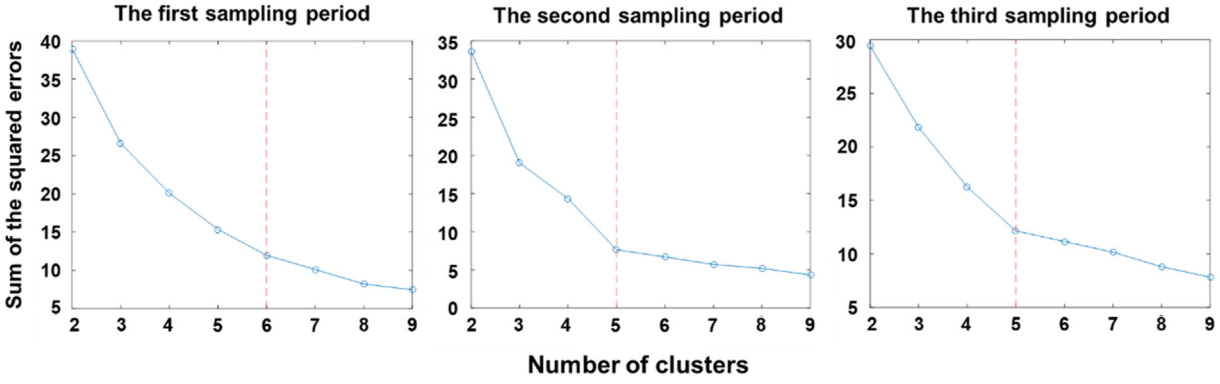

Figure 7. Plot of sum of the squared errors (SSE) versus number of clusters. The clusters are generated by weighted environmental variables clustering. The red dashed line indicates the optimal number of clusters.

Table 3. Important bands for each cluster of different sampling periods.

Sampling periods	Clusters	Bands selected from SR (nm)	Bands selected from CR (nm)	Bands selected from CRDR (nm)
The first sampling period	Cluster1	**545**, 705, **552**, **553**, **556**, 609, 695, **554**, 736, 615	**540**, 730, **545**, 529, 753, 723, **2268**, **546**, **551**, **2239**	**638**, 579, 562, **556**, 1158, 738, 618, 513, 748, **1416**
	Cluster2	**1535**, **1819**, **1544**, **2319**, **1011**, 1265, 901, **1459**, 1186, 1678	**1548**, **2317**, 1650, 1668, **2057**, **2064**, 1221, **2285**, **1513**, **2341**	1355, 1638, 2162, **1927**, 1250, **2272**, 1212, **2172**, 1553, **2005**
	Cluster3	374, **674**, **416**, 1612, 1718, 1793, **446**, 381, 579, **1729**	**2330**, **1730**, **1735**, **1740**, 1712, 1715, **1684**, **1722**, **1720**, 1748	**2081**, 1247, **2238**, 1062, 735, **1481**, 2209, **2057**, **980**, **1482**
	Cluster4	**2292**, **2312**, **1117**, **2303**, **2264**, **2304**, **2315**, **2263**, 568, 758	1714, **1695**, 1757, 1360, 2141, **670**, 1707, **2133**, **1687**, **1543**	**2062**, 2195, **2258**, **667**, **2187**, 2167, 1599, 1611, 1148, 2024
	Cluster5	**2170**, 2117, **2174**, **2120**, 1592, 1676, 366, 1658, 1663, 1677	2216, **1686**, 1648, 2206, 1644, 2204, **2121**, **1694**, **1683**, 2166	1477, 1478, **2065**, **1420**, 1464, 2206, 1466, **2171**, **1460**, 2221
	Cluster6	**500**, 1160, **672**, **2338**, **550**, 1717, **642**, **674**, **455**, **1728**	**2257**, **1206**, 1091, **2254**, 516, 2383, 376, **679**, 1352, 1714	**1818**, 884, 2224, **902**, **681**, **903**, **667**, 708, **2251**, 533
The second sampling period	Cluster1	354, 795, 374, 1004, 580, 751, 1052, 2032, 789, 1168	**1690**, **1692**, 1268, **1685**, 831, 1109, 1717, 625, **1727**, **638**	**2010**, **2278**, **2291**, **1027**, **1780**, 1677, 1845, 1704, 718, **1819**
	Cluster2	**452**, 513, **559**, 1225, **2259**, 728, **1725**, 1311, 516, 849	806, 356, 358, 1215, 1284, 805, **1116**, **644**, 2216, 1566	941, **913**, **971**, 445, 2430, 2481, 2216, **966**, 416, 1282
	Cluster3	**683**, **684**, 956, 574, 1235, **682**, 618, 371, **973**, 1168	897, **2324**, **1129**, 847, **2261**, 2226, 1272, 479, 2143, 2491	**964**, 888, 2414, 2473, 1851, **680**, 1278, 1251, 1597, **1919**
	Cluster4	**1025**, **429**, **664**, 890, **2357**, **470**, **971**, **654**, 384, 514	**923**, 872, 374, **929**, **920**, 625, 409, 1097, 517, **1910**	402, **914**, **2082**, 1164, 1096, 441, **1498**, **910**, 1001, **2299**
	Cluster5	598, **557**, 568, 566, 513, 611, **1482**, **1736**, 526, **435**	810, **546**, 730, **545**, 603, 538, **1193**, **560**, 580, 539	**2352**, **642**, 1185, 571, 562, 864, 598, 611, 586, **1691**
The third sampling period	Cluster1	**2301**, 411, **660**, 442, 1671, 489, 2483, 720, 476, 682	**2288**, **460**, 482, **2273**, **2279**, 1257, 620, 442, 881, **2253**	658, 1137, 1648, **2242**, 1179, 1767, **992**, 445, 411, **1787**
	Cluster2	**1542**, **1039**, **1534**, **560**, 1098, 533, 1556, 1330, 1623, 1380	**975**, **968**, **454**, 747, **1782**, 949, 752, **2098**, **1779**, 843	1620, 2160, 1246, 1330, 1674, 760, 1303, 2218, 2156, **2277**
	Cluster3	357, 359, **1919**, **438**, 1651, **678**, 404, 361, **1691**, 886	1182, 513, 1175, 354, 473, 760, **1582**, 702, **2245**, **631**	1145, 1231, 1597, 1166, 500, 946, 1649, **462**, 745, 746
	Cluster4	**2006**, 760, 781, 789, 1052, **1028**, **983**, 878, 945, **1577**	851, 1653, 849, **992**, 865, **1497**, 1565, **1548**, 850, 741	**1404**, **1414**, **1925**, **1429**, **1440**, **1405**, **1422**, **998**, **1424**, **1418**
	Cluster5	682, 369, 1329, 735, 1760, 680, 402, 514, 1341, **1400**	**2312**, 803, **1734**, **1742**, 1299, **2303**, **2320**, **2289**, 2292, **1724**	**1119**, 2215, **2294**, **2284**, 1706, **2276**, **2277**, **2089**, **2271**, **2280**

The bands for each cluster are sorted in descending order according to their importance in estimating LPC. Numbers in bold mean the bands associated with known absorption features listed by Curran (1989) and Kumar et al. (2002) while those in normal format denote the bands are not related with known absorption.

5. Discussion

5.1. Repeatability and stability of LM-WEVC

In order to test the repeatability and stability of LM-WEVC, the proposed method was applied to datasets from three sampling periods with different spectral data. From Figures.8 and 9 and 10, it can be seen that no matter which spectral data (SR, CR or CRDR) was used as input variable, prediction accuracy of LM-WEVC was always higher than that of LM-MEVC. More importantly, values of RPD for LM-WEVC were larger than 2.00 (indicating the model had good performance) at two seasons while those for LM-MEVC were all smaller than 2.00 at the three seasons. These results indicated that LM-WEVC is superior to LM-MEVC in predicting LPC of rubber trees constantly over the different ways of processing spectral data and over different sampling periods. This demonstrated that LM-WEVC was good in repeatability and stability.

5.1.1. Improvement of LM-WEVC over the LM-MEVC

The main improvement of LM-WEVC over the existing method LM-MEVC was that the different effects of environmental variables on LPC of rubber trees were taken into consideration when classifying samples into clusters or groups. This manipulation is more in conformity with the authentic situations because environment factors undoubtedly can impose influences on the growth and thus the LPC of rubber trees, and the influences of these environmental factors were likely to be different. Results of the current study confirmed this hypothesis. It was found that parent materials, cosine of aspect, and slope were the main factors affecting LPC of rubber trees (taking leaf samples collected from the first sampling period as example), and the weights of these environmental factors on LPC were 0.52, 0.25 and 0.22, respectively (Table 2). Asner et al. (2016) carried out a study throughout Andean and western Amazonian forests and reported a similar finding that elevation and substrate were the dominant factors influencing foliar phosphorus of forests, and the contributions of elevation and substrate to foliar phosphorus were around 15% and 10% respectively. Consideration of diverse influences of environmental factors on target variable would allow the relationships between LPC and spectra within each cluster or group be better characterized. Consequently, model predictive ability would be improved.

5.2. Application conditions and data requirement of LM-WEVC

The LM-WEVC approach was suitable for extensive and complex geographic regions. Within these areas, geographic environment is generally complex and with high heterogeneity (Goodchild, 2004; Zhu et al., 2018). The heterogeneity in geographic environment would give rise to marked variation in LPC and spectra, which could further result in unstable or varied relationships between LPC and spectra over space. From Table 3, it can be seen that important bands of clusters were different from each other. Taking the first sampling period as example, important bands selected from SR for cluster 1 were all located in the visible range (380–780 nm), while those for cluster 2 were mainly distributed in short-wave infrared (SWIR) (1100–2500 nm) and near infrared (780–1100 nm) (NIR) range. This indicated that the relationships between spectral data and LPC were not the same for different clusters. Similarly, Asner et al. (2014) also found that spectral reflectance

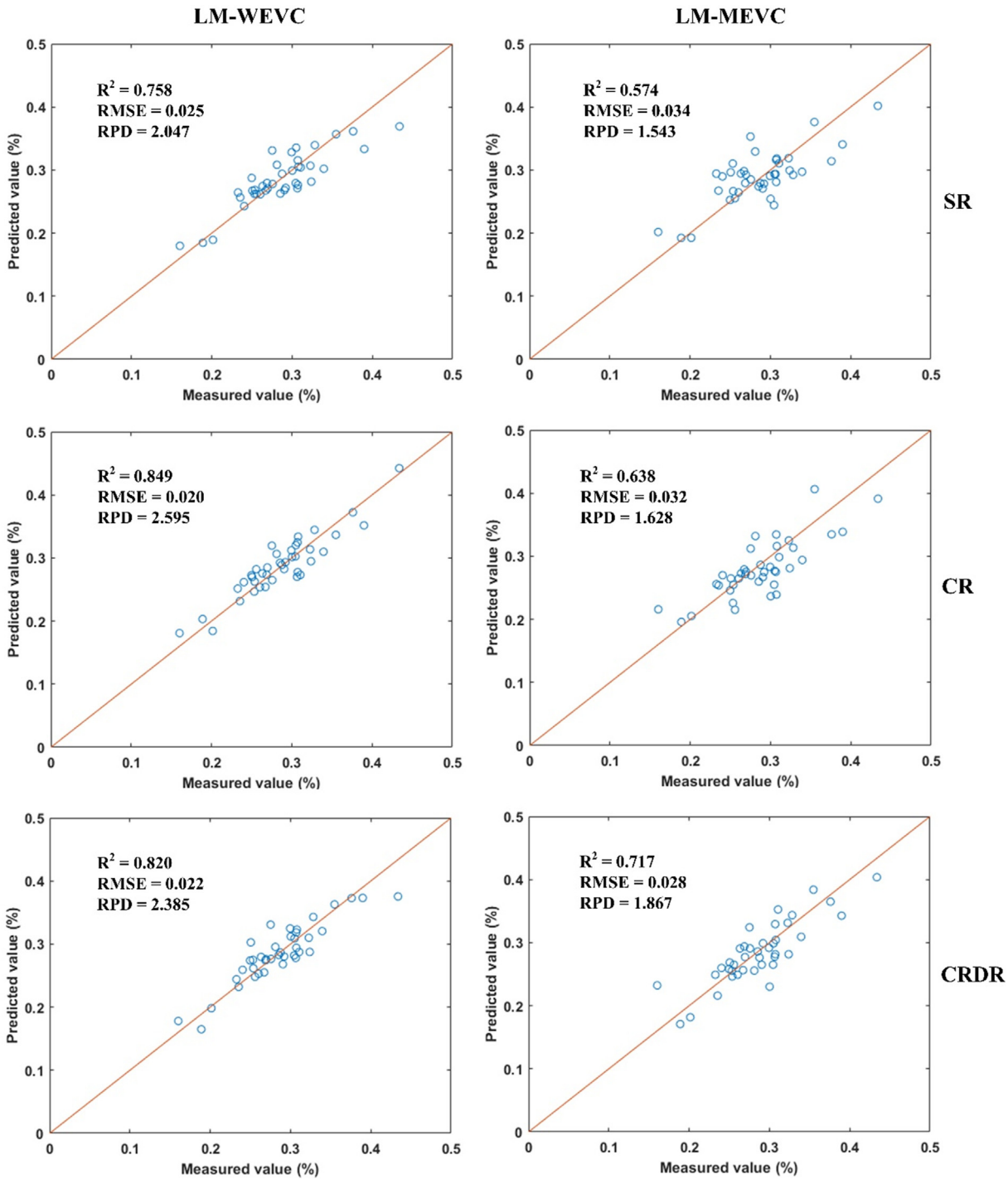

Figure 8. Prediction results of LM-WEVC and LM-MEVC in test set from the first sampling period. LM-WEVC and LM-MEVC indicates local model based on weighted environmental variables clustering, and local model based on multiple environmental variables clustering, respectively. SR, CR, and CRDR represent the spectral reflectance, the continuum removed reflectance, and the continuum-removed derivative reflectance, respectively.

of tropical forests was highest within the NIR region but lowest within the SWIR at the summit of the mountain where the ratio of nitrogen to phosphorus (N: P) was the lowest. However, relationships between N: P ratio and spectral reflectance within NIR and SWIR were on the contrary at the lower elevation sites. The unstable or varied relationships between target variable and spectra impose challenges on estimating plant traits using hyperspectral technique at large scale. Up to now, almost all of studies used GM method to estimate plant traits with hyperspectral reflectance. The commonly used GM method assumed that the relationships between target variable and spectral data were stable over space. This assumption might be valid within small area or region with homogeneous environmental settings, but would be invalid in large area with complex environmental conditions over which the LM-WEVC is more appropriate.

Application of LM-WEVC requires a large number of samples. Within each cluster or group, sufficient samples are acquired to

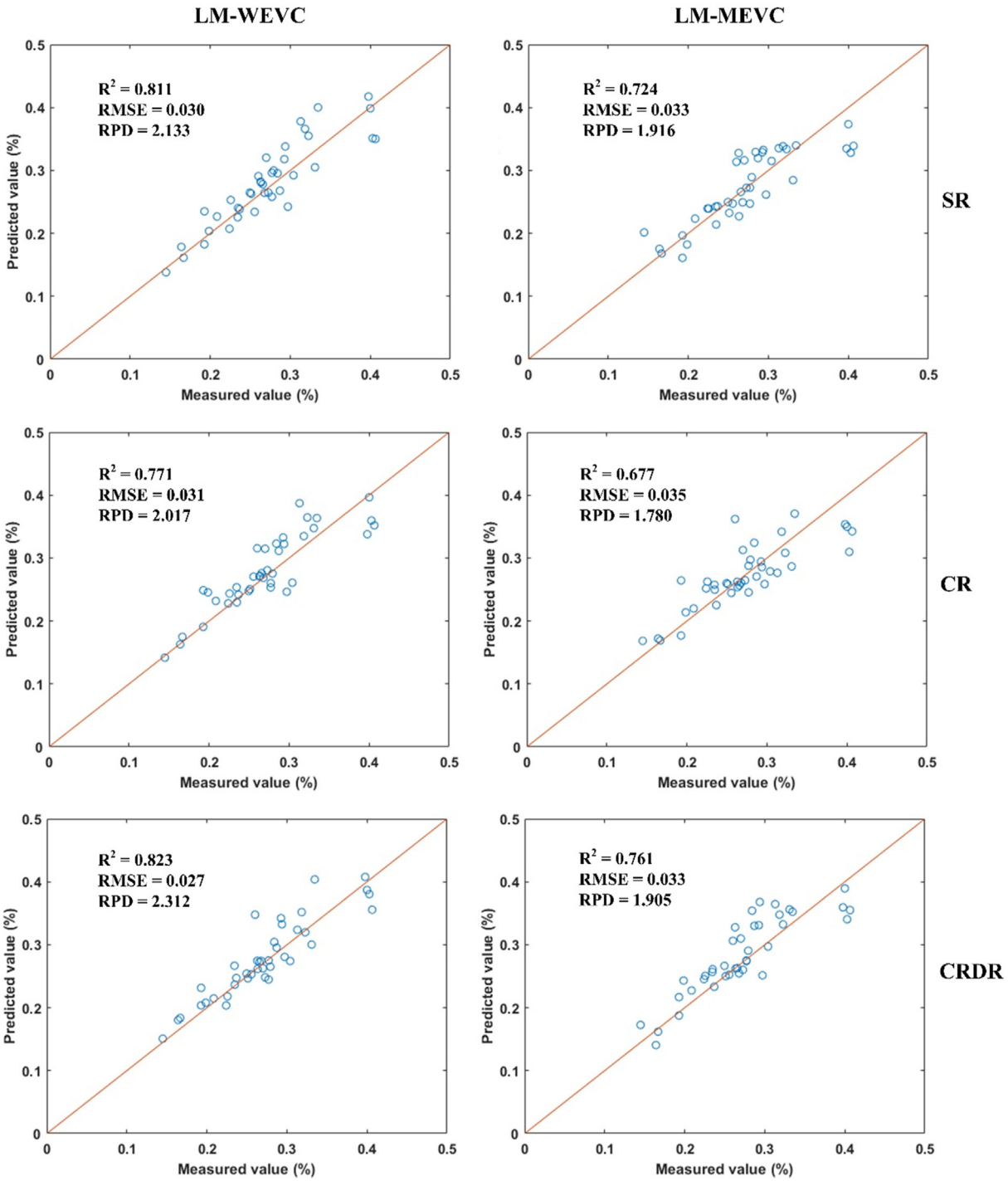

Figure 9. Prediction results of LM-WEVC and LM-MEVC in test set from the second sampling period. LM-WEVC, and LM-MEVC indicates local model based on weighted environmental variables clustering, and local model based on multiple environmental variables clustering, respectively. SR, CR, and CRDR represent the spectral reflectance, the continuum removed reflectance, and the continuum-removed derivative reflectance, respectively.

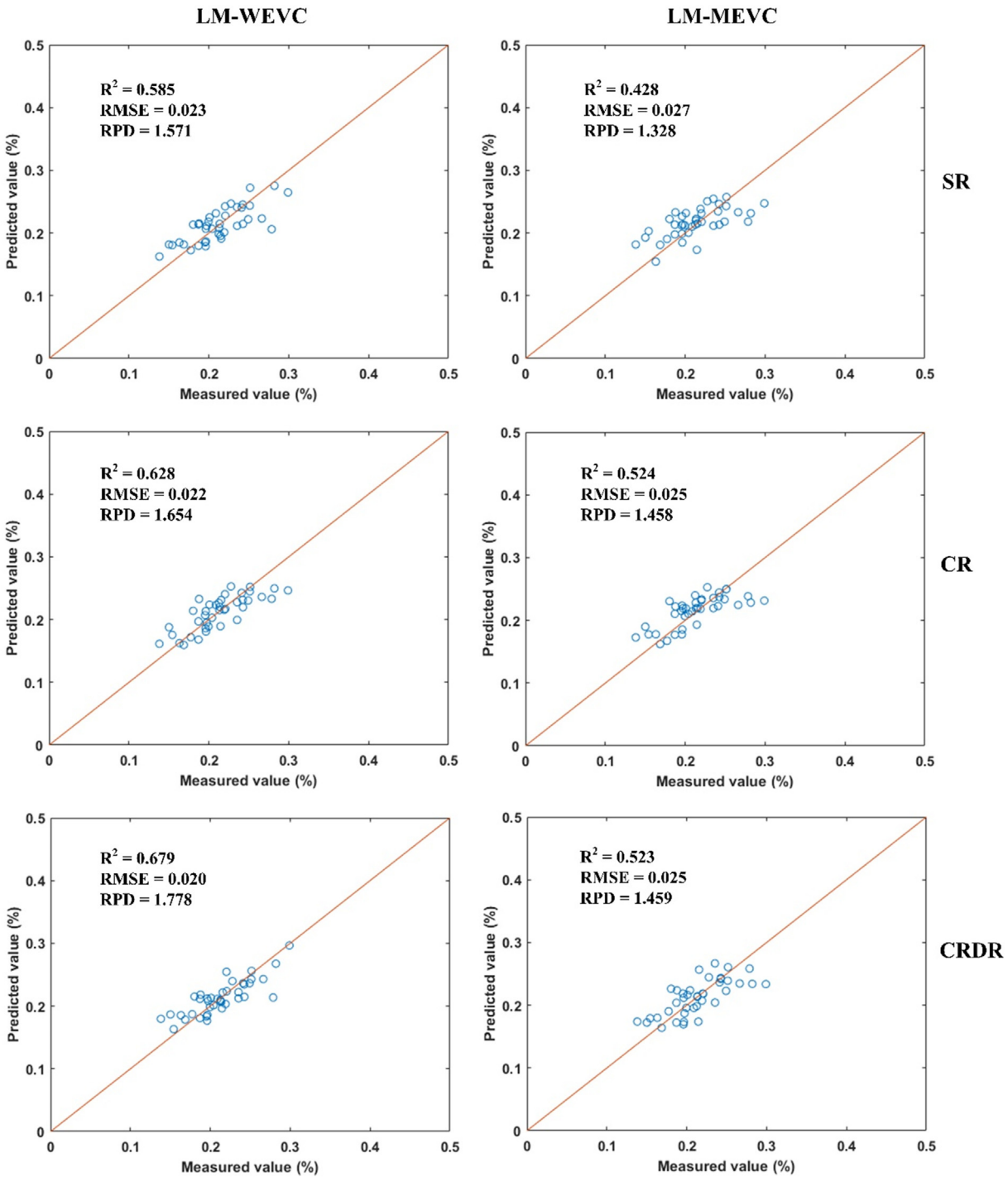

Figure 10. Prediction results of LM-WEVC and LM-MEVC in test set from the third sampling period. LM-WEVC and LM-MEVC indicates local model based on weighted environmental variables clustering and local model based on multiple environmental variables clustering, respectively. SR, CR, and CRDR represent the spectral reflectance, the continuum removed reflectance, and the continuum-removed derivative reflectance, respectively.

effectively capture the relationship between target variable and spectra. If the samples are too small, the relationship between target variable and spectra could not be characterized well, and the predictive ability of the estimation model would be reduced. Moura-Bueno et al. (2020) divided soil samples into three clusters on basis of physiographic regions. Within cluster 2, there were only a small number of soil samples. These limited soil samples leaded to insufficient capture of variations in soil organic carbon and spectra, which further brought about large prediction errors.

6. Conclusions

This paper presents a new local modeling approach, LM-WEVC, for estimating LPC of rubber trees at regional scale. This proposed approach takes differences in impacts of environmental factors on LPC into consideration when using environmental factors as classification variables to divide leaf samples into clusters or groups. The case study showed that LM-WEVC outperformed the existing LM-MEVC method which does not consider the different influences of environmental

variables on LPC when classify leaf samples into clusters or groups. This demonstrated that consideration of the differences in impacts of environmental variables on LPC when classify leaf samples into clusters or groups can improve the predictive ability of local models. Therefore, the main contribution of this study is that the existing LM-MEVC is improved by taking the different influences of environmental variables on LPC into consideration when divide leaf samples into clusters or groups. The LM-WEVC is appropriate for estimating LPC of rubber trees at large scale especially with complex environmental conditions. Nevertheless, application of LM-WEVC requires a large quantity of samples to effectively characterize the relationships between LPC and environmental factors within each cluster or group.

Declarations

Author contribution statement

Peng-Tao Guo: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

A-Xing Zhu: Conceived and designed the experiments; Wrote the paper.

Zheng-Zao Cha: Performed the experiments; Contributed reagents, materials, analysis tools or data.

Mao-Fen Li: Contributed reagents, materials, analysis tools or data.

Wei Luo: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data.

Funding statement

Dr. Peng-Tao Guo was supported by Hainan Provincial Natural Science Foundation of China [321RC656].

Prof. A-Xing Zhu was supported by National Natural Science Foundation of China [41871300].

Dr Mao-Fen Li was supported by Opening Project Fund of Key Laboratory of Rubber Biology and Genetic Resource Utilization, Ministry of Agriculture and Rural Affairs [RRI-KLOF201803].

Prof. Zheng-Zao Cha was supported by the National Technical System of Natural Rubber Industry [CARS-33-ZP-2].

Data availability statement

Data will be made available on request.

Declaration of interest's statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

Acknowledgements

The authors thank Dr. Marian-Daniel Iordache (Flemish Institute for Technological Research, Mol, Flanders, Belgium) for his help in providing Matlab codes for calculating CR and CRDR. The authors also thank Mei-Rong Bei and Hong-Zhu Yang for their help in sample collection and chemical analysis.

References

Al-Abbas, A.H., Barr, R., Hall, J.D., Crane, F.L., Baumgardner, M.F., 1974. Spectra of normal and nutrient deficient maize leaves. Agron. J. 66 (1), 16–20.

Araújo, S.R., Wetterlind, J., Demattê, J.A.M., Stenberg, B., 2014. Improving the prediction performance of a large tropical vis-NIR spectroscopic soil library from Brazil by clustering into smaller subsets or use of data mining calibration techniques. Eur. J. Soil Sci. 65, 718–729.

Asner, G.P., Knapp, D.E., Anderson, C.B., Martin, R.E., Vaughn, N., 2016. Large-scale climatic and geophysical controls on the leaf economics spectrum. P. Natl. Acad. Sci. USA 113 (28), 201604863.

Asner, G.P., Martin, R.E., Carranza-Jiménez, L., Sinca, F., Tupayachi, R., Anderson, C.B., Martinez, P., 2014. Functional and biological diversity of foliar spectra in tree canopies throughout the Andes to Amazon region. New Phytol. 204, 127–139.

Asner, G.P., Martin, R.E., Ford, A.J., Metcalfe, D.J., Liddell, M.J., 2009. Leaf chemical and spectral diversity in Australian tropical forests. Ecol. Appl. 19 (1), 236–253.

Asner, G.P., Martin, R.E., Knapp, D.E., Tupayachi, R., Anderson, C.B., Sinca, F., Vaughn, N.R., Llactayo, W., 2017. Airborne laser-guided imaging spectroscopy to map forest trait diversity and guide conservation. Science 355, 385–389.

Bao, Y., Meng, X., Ustin, S., Wang, X., Zhang, X., Liu, H., Tang, H., 2020. Vis-SWIR spectral prediction model for soil organic matter with different grouping strategies. Catena 195, 104703.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Castro-Esau, K., Sánchez-Azofeifa, G.A., Rivard, B., Wright, S.J., Quesada, M., 2006. Variability in leaf optical properties of Mesoamerican trees and the potential for species classification. Am. J. Bot. 93 (4), 517–530.

Clark, R.N., Roush, T.L., 1984. Reflectance spectroscopy: quantitative analysis techniques for remote sensing applications. J. Geophys. Res. 89 (B7), 6329–6340.

Curran, P.J., 1989. Remote sensing of foliar chemistry. Remote Sens. Environ. 30, 271–278.

Dong, C., An, T., Yang, M., Yang, C., Liu, Z., Li, Y., Duan, D., Fan, S., 2022. Quantitative prediction and visual detection of the moisture content of withering leaves in black tea (Camellia sinensis) with hyperspectral image. Infrared Phys. Technol.

FAO, 1998. World Reference Base for Soil Resources. World Soil Resources Report No. 84FAO, Rome.

Farrar, D.E., Glauber, R.R., 1967. Multicollinearity in regression analysis: the problem revisited. Rev. Econ. Stat. 49 (1), 92–107.

Fick, S.E., Hijmans, R.J., 2017. WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas. Int. J. Climatol. 37 (12), 4302–4315.

Gao, J., Meng, B., Liang, T., Feng, Q., Ge, J., Yin, J., Wu, C., Cui, X., Hou, M., Liu, J., Xie, H., 2019. Modeling alpine grassland forage phosphorus based on hyperspectral remote sensing and multi-factor machine learning algorithm in the east of Tibetan Plateau, China. ISPRS J Photogramm 147, 104–117.

Gogé, F., Joffre, R., Jolivet, C., Ross, I., Ranjard, L., 2012. Optimization criteria in sample selection step of local regression for quantitative analysis of large soil NIRS database. Chemometr. Intell. Lab. 110, 168–176.

Goodchild, M.F., 2004. The validity and usefulness of laws in geographic information science and geography. Ann. Assoc. Am. Geogr. 94, 300–303.

Guo, P.T., Li, M.F., Luo, W., Tang, Q.F., Liu, Z.W., Lin, Z.M., 2015. Digital mapping of soil organic matter for rubber plantation at regional scale: an application of random forest plus residuals kriging approach. Geoderma 237–238, 49–59.

Guo, P.T., Shi, Z., Li, M.F., Luo, W., Cha, Z.Z., 2018. A robust method to estimate foliar phosphorus of rubber trees with hyperspectral reflectance. Ind. Crop. Prod. 126, 1–12.

Guo, P.T., Su, Y., Cha, Z.Z., Lin, Q.H., Luo, W., Lin, Z.M., 2016. Prediction of leaf phosphorus contents for rubber seedlings based on hyperspectral sensitive bands and back propagation artificial neural network. J. Clin. Transl. Endocrinol. Case Rep. 32 (Supp.1), 177–183 (In Chinese with English abstract).

Guo, P.T., Wu, W., Sheng, Q.K., Li, M.F., Liu, H.B., Wang, Z.Y., 2013. Prediction of soil organic matter using artificial neural network and topographic indicators in hilly areas. Nutrient Cycl. Agroecosyst. 95, 333–344.

Hartigan, J.A., 1975. Clustering Algorithms. Wiley, New York, p. 209.

He, X.D., Lu, X.Z., Wu, X.P., 1991. Soil fertility zoning of rubber plantation and its application in Hainan Island. Chin. J. Trop. Agric. 1, 40–48 (in Chinese).

IBM Corp, 2011. IBM SPSS Statistics for Windows, Version 20.0. IBM Corp., Armonk, NY.

Islam, M., Ahmad, S., Aslam, S., Athar, M., 1999. Mineral content and nutritive value of native grasses and the response to added phosphorus in a Pilbara rangeland. Trop. Grassl. 33 (4), 193–200.

Ito, E., Ono, K., Ito, Y.M., Araki, M., 2008. A neural network approach to simple prediction of soil nitrification potential: a case study in Japanese temperate forests. Ecol. Model. 219, 200–211.

Kennard, R.W., Stone, L.A., 1969. Computer-aided design of experiments. Technometrics 11, 137–148.

Knox, N.M., Skidmore, A.K., Prins, H.H.T., Asner, G.P., van der Werff, H.M.A., de Boer, W.F., van der Waal, C., de Knegt, H.J., Kohi, E.M., Slotow, R., Grant, R.C., 2011. Dry season mapping of savanna forage quality, using the hyperspectral Carnegie Airborne Observatory sensor. Remote Sens. Environ. 115, 1478–1488.

Kumar, L., Schmidt, K., Dury, S., Skidmore, A., 2002. Imaging spectrometry and vegetation science. In: Meer, F.D.v., Jong, S.M.D. (Eds.), Imaging Spectrometry, Remote Sensing and Digital Image Processing, 4. Springer, Dordrecht, the Netherlands, pp. 130–133.

Li, L., Jákli, B., Lu, P., Ren, T., Ming, J., Liu, S., Wang, S., Lu, J., 2018. Assessing leaf nitrogen concentration of winter oilseed rape with canopy hyperspectral technique considering a non-uniform vertical nitrogen distribution. Ind. Crop. Prod. 116, 1–14.

Liu, S., Shen, H., Chen, S., Zhao, X., Biswas, A., Jia, X., Shi, Z., Fang, J., 2019. Estimating forest soil organic carbon content using vis-NIR spectroscopy: implications for large-scale soil carbon spectroscopic assessment. Geoderma 348, 37–44.

Lu, X., He, X., 1982. Fertilizer application based on nutrient diagnosis of rubber trees. Chinese J. Trop. Crop. 3 (1), 27–39 (In Chinese with English abstract).

Moura-Bueno, J.M., Dalmolin, R.S.D., Horst-Heinen, T.Z., ten Caten, A., Vasques, G.M., Dotto, A.C., Grunwald, S., 2020. When does stratification of a subtropical soil spectral library improve predictions of soil organic carbon content? Sci. Total Environ. 737, 139895.

Moura-Bueno, J.M., Dalmolin, R.S.D., ten Caten, A., Dotto, A.C., Demattê, J.A.M., 2019. Stratification of a local VIS-NIR-SWIR spectral library by homogeneity criteria yields more accurate soil organic carbon predictions. Geoderma 337, 565–581.

Mutanga, O., Kumar, L., 2007. Estimating and mapping grass phosphorus concentration in an African savanna using hyperspectral image data. Int. J. Rem. Sens. 28 (21), 4897–4911.

Mutanga, O., Skidmore, A.K., Prins, H.H.T., 2004. Predicting in situ pasture quality in the Kruger National Park, South Africa, using continuum-removed absorption features. Remote Sens. Environ. 89, 393–408.

Ogen, Y., Zaluda, J., Francos, N., Goldshleger, N., Ben-Dor, E., 2019. Cluster-based spectral models for a robust assessment of soil properties. Geoderma 340, 175–184.

Okita, T.W., 1992. Is there an alternative pathway for starch synthesis? Plant Physiol 100 (2), 560–564.

Ramoelo, A., Skidmore, A.K., Cho, M.A., Mathieu, R., Heitkönig, I.M.A., Dudeni-Tlhone, N., Schlerf, M., Prins, H.H.T., 2013. Non-linear partial least square regression increases the estimation accuracy of grass nitrogen and phosphorus using in situ hyperspectral and environmental data. ISPRS J. Photogramm. 82, 27–40.

Ramoelo, A., Skidmore, A.K., Schlerf, M., Mathieu, R., Heitkönig, 2011. Water-removed spectra increase the retrieval accuracy when estimating savanna grass nitrogen and phosphorus concentration. ISPRS J Photogramm 66, 408–417.

Reshef, D.N., Reshef, Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M., Sabeti, P.C., 2011. Detecting novel associations in large data sets. Science 334, 1518.

Said, M., Hyandye, C., Komakech, H.C., Mjemah, I.C., Munishi, L.K., 2021. Predicting land use/cover changes and its association to agricultural production on the slopes of Mount Kilimanjaro, Tanzania. Spatial Sci. 27 (2), 189–209.

Shi, Z., Wang, Q.L., Peng, J., Ji, W.J., Liu, H.J., Li, X., 2014. Development of a national VNIR soil-spectral library for soil classification and prediction of organic matter concentrations. Sci. China Earth Sci. 44, 978–988 (In Chinese with English abstract).

Song, C., Shi, X., Wang, J., 2020a. Spatiotemporally varying coefficients (STVC) model: a Bayesian local regression to detect spatial and temporal nonstationarity in variables relationships. Spatial Sci. 26 (3), 277–291.

Song, X., Wu, H., Ju, B., Liu, F., Yang, F., Li, D., Zhao, Y., Yang, J., Zhang, G., 2020b. Pedoclimatic zone-based three-dimensional soil organic carbon mapping in China. Geoderma 363, 114145.

van Beilen, J.B., Poirier, Y., 2007. Establishment of new crops for the production of natural rubber. Trends Biotechnol. 25 (11), 522–529.

Wang, J., Tian, T., Wang, H., Cui, J., Zhu, Y., Zhang, W., Tong, X., Zhou, T., Yang, Z., Sun, J., 2021. Estimating cotton leaf nitrogen by combining the bands sensitive to nitrogen concentration and oxidase activities using hyperspectral imaging. Comput. Electron. Agric. 189, 106390.

Wang, S.Q., Li, W.D., Li, J., Liu, X.S., 2013. Prediction of soil texture using FT-NIR spectroscopy and PXRF spectrometry with data fusion. Soil Sci. 178, 626–638.

Zhang, Z., Liu, F., He, Y., Gong, X., 2013. Detecting macronutrients contents and distribution in oilseed rape leaves based on hyperspectral imaging. Biosyst. Eng. 115, 56–65.

Zhang, X., Liu, H., Zhang, X., Yu, S., Dou, X., Xie, Y., Wang, N., 2018. Allocate soil individuals to soil classes with topsoil spectral characteristics and decision trees. Geoderma 320, 12–22.

Zhu, A.X., Lu, G., Qin, C.Z., Zhou, C., 2018. Spatial prediction based on third law of geography. Spatial Sci. 24 (4), 225–240.

Zhu, A.X., Zhao, F.H., Liang, P., Qin, C.Z., 2021. Next generation of GIS: must be easy. Spatial Sci. 27 (1), 71–86.

Zou, X., Zhao, J., Povey, M.J.W., Holmes, M., Hanpin, M., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667, 14–32.