Mapping large area tea plantations using progressive random forest and Google Earth Engine

Le'an Qu,[a,b,c] Manchun Li,[a,c] Zhenjie Chen,[a,c,*] Wangbing Liu,[d] Junjun Zhi,[b] and Lechun Zhang[b]

[a]Nanjing University, School of Geography and Ocean Science, Nanjing, China
[b]Anhui Normal University, School of Geography and Tourism, Wuhu, China
[c]Nanjing University, Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing, China
[d]Ministry of Natural Resources, Anhui Province Institute of Land Surveying and Planning, Key Laboratory of JiangHuai Arable Land Resources Protection and Eco-Restoration, Hefei, China

Abstract. A timely and accurate understanding of the spatial distribution of tea plantations is beneficial for agricultural management and regional sustainable development. However, obtaining detailed distribution data on large-area tea plantations remains challenging owing to limitations in computational capabilities, training data, and workflow design. Utilizing the Google Earth Engine, which provides a catalog of multisource data in a cloud-based environment, we developed a methodology to generate a highly accurate tea plantation map, with a 10-m resolution, for Anhui Province, China, by integrating a random forest model with a progressive model. Our major contribution lies in this hybrid approach, which comprises two major components: (1) an optimal classification band combination derived from Sentinel-2 products and the digital elevation model filtered by the J-M distance model and (2) a progressive random forest method introduced for tea plantation classification. The experimental results show that our proposed workflow achieved an average classification accuracy of 89.27% for the entire Anhui Province. In addition, this approach is semiautomatic and can effectively reduce the labor required during the generation of training data compared with traditional classification approaches. These findings demonstrate the potential of integrating machine learning and progressive models to produce high-precision remote sensing classification maps. © *The Authors. Published by SPIE under a Creative Commons Attribution 4.0 International License. Distribution or reproduction of this work in whole or in part requires full attribution of the original publication, including its DOI.* [DOI: 10.1117/1.JRS.16.024509]

1 Introduction

Tea, which originated in China, has become one of the three most popular beverages worldwide.[1] In China, the global leader in tea production, tea is primarily grown in tropical and subtropical regions, such as Anhui Province,[2] where tea cultivation plays a vital role in the agricultural economy and rural development.[3] According to the Anhui Provincial Bureau of Statistics, the area of tea plantations in Anhui in 2019 was 1870.58 km^2, an increase of 57.38% from 1990.[4] Although the expansion of tea plantations has promoted local economic development, it has also caused a series of environmental problems, such as decreased soil fertility and soil erosion.[5] Therefore, a timely and accurate understanding of the spatial distribution of tea plantations is conducive to governmental environmental protection and agricultural management.[6]

Information on the distribution of tea plantations can generally be obtained through manual reporting and remote sensing image classification.[7] Utilizing remote sensing data to identify tea plantations saves time and labor costs; hence, it is more popular than manual reporting.[8] The biggest difficulty in using remote sensing data is that tea trees are perennial evergreen woody

*Address all correspondence to Zhenjie Chen, chenzj@nju.edu.cn

plants, and their spectral characteristics are easily confused with similar woody plants, such as natural forests.[9] Thus, it is difficult to achieve accurate tea plantation classifications using only spectral bands.[1] Hence, previous studies primarily used hyperspectral and high-resolution images, fusions of multiple types of remote sensing data, and classifications based on phenological features to identify tea plantations.[1,2,9–11] The classification accuracies using the former three methods are very high, but often the data are expensive and the study area is small, which makes it difficult to identify tea plantations over large areas.[12] In this situation, phenological feature-based classification using Sentinel-2 data is among best choices for identifying the distribution of tea plantations over large areas.[13] This is because Sentinel-2 data have three distinct advantages over other satellite images: worldwide coverage, high temporal and spatial resolution, and a variety of red-edge bands that are sensitive to vegetation.[14]

Nevertheless, there are some challenges in applying this method to identifying tea plantations over a large area.[15] First, determining the optimal combination of bands for classification is difficult.[12] Too few classification bands may lead to low classification accuracy, whereas too many bands may lead to low classification efficiency, over-fitting, and local optima.[16] Second, the classification of tea plantations over a large area requires many accurate training samples.[17] If the number of training samples is small, the classification accuracy is usually low and has a large uncertainty.[18] However, obtaining multiple training samples is costly and time-consuming.[19] Therefore, it is essential to balance the number of training samples with the required labor. Finally, for large-area time-series classifications, a high storage capacity is necessary to store the remote sensing data.[20] Furthermore, strong computing power is required to achieve high classification efficiency.[21]

Web-based remote-sensing cloud platforms such as Google Earth Engine (GEE) can greatly reduce computing time and enable the classification of large areas using remote sensing time series data.[22] GEE contains a variety of remote sensing images from different sensors, which reduces personal data storage requirements.[23] In addition, GEE provides high computing power that greatly improves classification efficiency, and it can effectively perform repeated comparison experiments to determine the optimal classification bands.[24] Moreover, GEE provides a visual interactive platform that can dynamically add training samples, providing a balance between training sample acquisition and labor.[25] In this study, we used the Jeffries–Matusita distance (J-M distance) model to select the optimal classification features and utilized the progressive random forest method to classify tea plantations. These processes were conducted on the GEE platform, allowing our workflow to achieve high classification accuracy for tea plantations over a large study area with minimal time and labor costs.

The goals of this study were to classify Sentinel-2 time series data to obtain a high-precision tea plantation distribution map for Anhui Province during 2020 and determine which classification features and methods effectively perform high-precision tea plantation classification over a large area with complex terrain. Specifically, we addressed the following two questions: (1) how can we select the optimal combination of classification features in GEE, so it can classify tea plantations with high classification accuracy and efficiency? (2) How can a suitable classification method be constructed in GEE to ensure that tea plantations can be identified with high accuracy over a large area?

2 Study Area and Datasets

2.1 *Study Area*

Anhui Province is in eastern China, between 114°54E and 119°37E and 29°41N and 34°38N. It spans 450 km from east to west and 570 km from north to south, with a total area of 140, 100 km^2, accounting for 1.45% of the total land area of China.[26,27] Anhui is in the middle to lower reaches of the Yangtze River and the Huaihe River. It borders Jiangsu to the east, Zhejiang and Jiangxi to the south, Henan and Hubei to the west, and Shandong to the north.[28] As of July 2020, there were 16 cities in the administrative regions in Anhui Province: Hefei, Wuhu, Bengbu, Huainan, Ma'anshan, Huaibei, Tongling, Anqing, Huangshan, Fuyang, Suzhou, Chuzhou, Lu'an, Xuancheng, Chizhou, and Bozhou (Fig. 1).
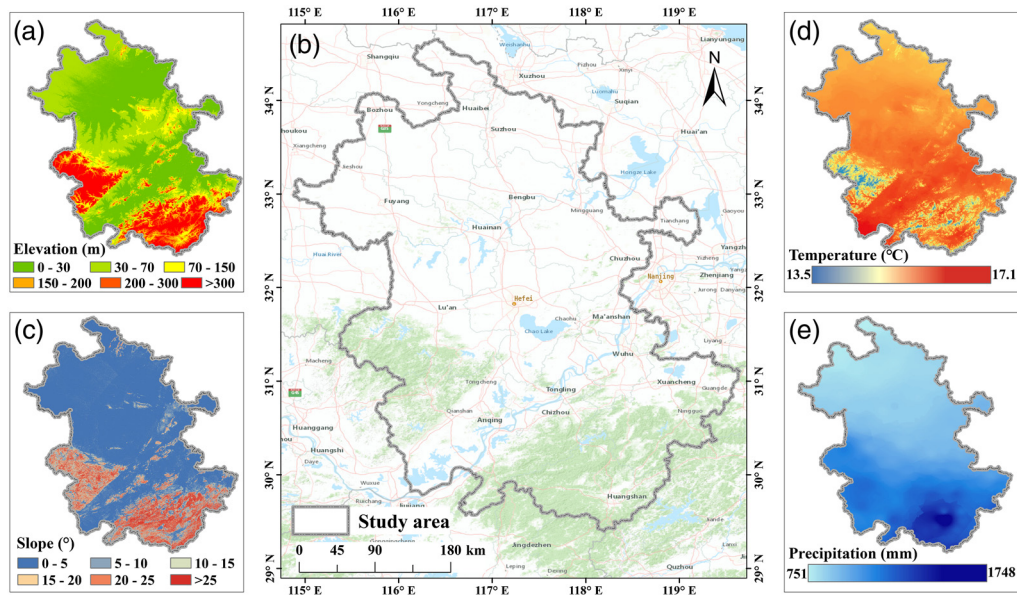
Fig. 1 Location of the study area. Maps showing the (a) location, (b) elevation, (c) slope, (d) annual average temperature, and (e) annual average precipitation in the study area.

In general, the natural conditions of northern and southern Anhui differ greatly.[29] The terrain is generally high in the west and south and low in the east and north. The north is dominated by plains, whereas the south is dominated by hills and mountains. Anhui is in the mid-latitude zone, which belongs to the transitional region between the warm temperate and subtropical zones. The annual average temperature in Anhui is ~14°C to 17°C, and the annual average precipitation is 750 to 1750 mm.[26] Precipitation is higher in the south and the mountainous areas and is lesser in the north, the plains, and the hills. Rainfall is abundant in summer, accounting for 40% to 60% of the annual precipitation.[27] The unique climatic and topographical conditions have made Anhui Province one of the major tea-producing provinces in China.[6] In 2019, the tea plantation area reached 187,058 hectares, and the total output of tea was 121,980 tons.[4]

2.2 Basic Data and Data Preprocessing

The data used in this study include four types of raster data, specifically, Sentinel-2 MSI data, land use and land cover data, topographic data, and Google Earth CNES/Airbus imageries, and two types of vector data, specifically, administrative boundary data and field survey data. All data were open-access datasets, except for the field survey data. See Table 1 for detailed descriptions of the data used in this study.

The Sentinel-2 data were obtained by a satellite cluster composed of two identical satellites, Sentinel-2 A and Sentinel-2 B.[30] The revisit periods of the two satellites were 5 days and were simultaneous.[31] In this study, Sentinel-2 satellite data were selected as the main remote sensing image data for accurately identifying the distribution of tea plantations. Sentinel-2 satellite data products include three levels: level-0, level-1C, and level-2A. Among them, level-2A images are bottom-of-atmosphere reflectance in cartographic geometry after atmospheric correction and, thus, are ready to use.[32] In this study, we used level-2A products; only the identification of clouds and cloud shadows was required prior to use. We used the Fmask 4.0 algorithm by Qiu et al. for cloud detection and achieved good results.[33]

Topographic features (elevation, slope, and aspect) are important factors that affect tea tree growth; therefore, they can effectively improve the classification accuracy for tea plantations.[11] To describe the topographic features, we used a digital elevation model (DEM) with a 30-m resolution generated by the Shuttle Radar Topography Mission (SRTM).[34] This DEM is a post-processed elevation dataset that is widely used because of its high accuracy and extensive coverage.[35]

Table 1 Data included in this study and their usage.

Data type	Data name	Spatial resolution	Usage
Raster	Sentinel-2 MSI[a]	10 m	Identify tea gardens and other land use types
	Google Earth CNES/Airbus imageries[a]	0.3 m	Generate sample points through visual interpretation
	GlobeLand30	30 m	Cross-validate land use and land cover types except for tea plantations through visual interpretation
	SRTM V3 DEM[a]	30 m	Assist remote sensing image classification
Vector	Administrative boundary	1:10,000	Determine the extent of the study area
	Field survey data	—	Generate training and verification sample points

Note: SRTM: Shuttle Radar Topography Mission.
[a]It can be called on Google Earth Engine directly.

High-precision reference data and an appropriate classification system are prerequisites for classifications. In this study, we combined imagery observations and field investigations to determine the reference points. We collected field survey data acquired with a handheld GPS/GNSS receiver in point-positioning mode in Huangshan and Wuhu City in 2020. The imagery observation data for other cities were incrementally added via Google Earth CNES/Airbus during the progressive random forest classification process. According to Giuseppe Pulighe et al., an error of ~1 m occurred in visualizing added points.[36] Considering that our classification results have an accuracy of 10 m, these points can be used for classification and accuracy verification.

GlobeLand30 was used for qualitative cross-validation with Sentinel-2 MSI classification results through visual interpretation. Then, according to the distribution of the ground objects in the study area, they were reclassified into six LULC (land use/land cover) types: cropland, forest, grassland, water bodies, built-up land, and unused.[37]

After the data were collected, the inputs were resampled to a spatial resolution of 10 m, which corresponds to the fine resolution of the Sentinel-2 MSI. Finally, all data were projected into the WGS84 coordinate system (EPSG: 4326) for tea plantation classification.

3 Methods

Figure 2 shows the overall workflow for the classification of tea plantations in the study area in 2020. First, to enhance the separability of the tea plantations from the other LULC types, we used original band features, remote sensing index features, texture features, tasseled cap transformation (TCT) features, and terrain features to generate overall classification features. Then, based on maintaining the separability between the tea plantations and other LULC types, the J-M distance model was applied to obtain the optimal classification features. Finally, we used a progressive random forest classification method for high-precision classification.

3.1 Selection and Optimization of Classification Features

3.1.1 Spectral features of the original bands

The spectral features of the original bands of the remote sensing images record the reflection intensity of different ground objects.[38] By analyzing the different spectral characteristics of these objects, it is possible to distinguish vegetation, water bodies, and other types of ground objects.[39] The 60-m bands in the Sentinel-2 MSI data are mainly used for atmospheric correction; therefore, they were excluded from this study. The other 10 original bands in the Sentinel-2 MSI data were selected for classification, specifically, blue, green, red, red edge1-4, NIR, and SWIR1-2. To analyze the phenological characteristics of the vegetation, we calculated the maximum,
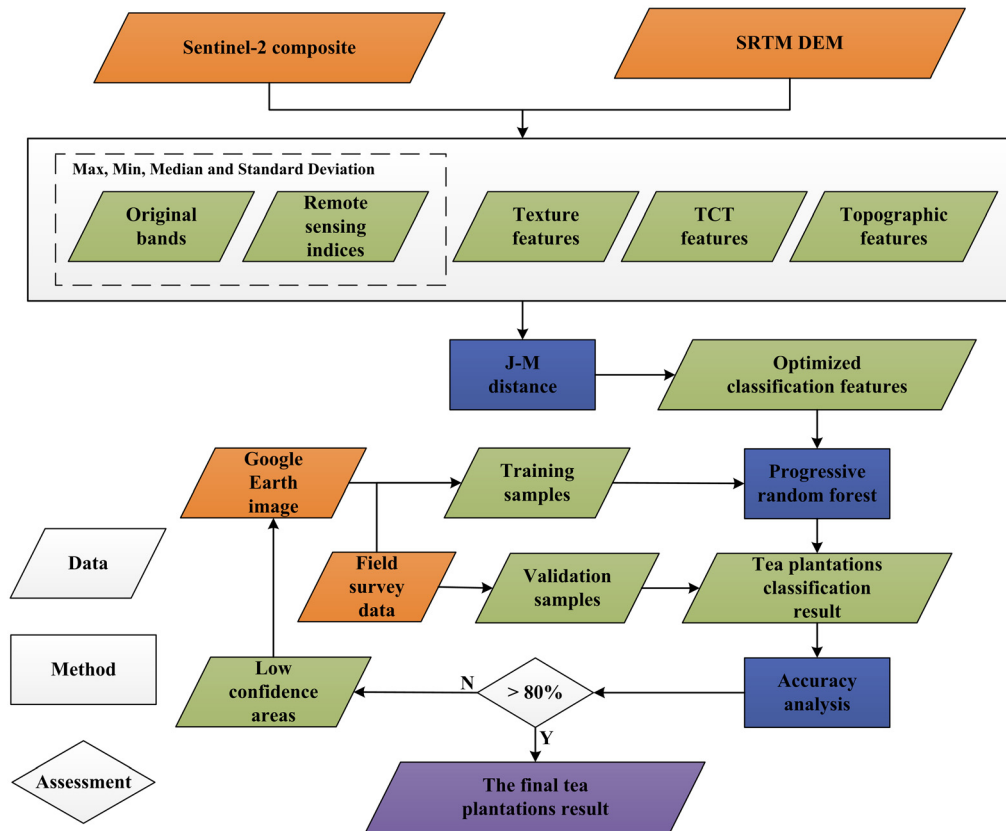
Fig. 2 Overall workflow chart for the classification of tea plantations.

minimum, median, and standard deviation of the digital number of all original bands in the study area in 2020. Tea trees and natural forests show strong spectral similarity; therefore, in addition to the original spectral features, other classification features, including remote sensing indices, red-edge vegetation indices, TCT features, and topographic features, were utilized in this study.

3.1.2 *Features of remote sensing indices*

Due to the surface environment, the same objects may have different spectral features, and different objects may have the same spectral features, which directly impacts the classification accuracy.[40] To overcome this issue, this study selected five commonly used remote sensing indices to improve classification accuracy. These indices were mainly used to enhance the distinction among other LULC types, rather than tea plantations. Nevertheless, because of the strong spectral similarity between tea trees and forests, eight red-edge vegetation indices were used to further distinguish tea plantations from forests.

The five common remote sensing indices include the normalized difference vegetation index (NDVI), soil adjusted vegetation index (SAVI), normalized difference water index (NDWI), modified normalized difference water index (MNDWI), and normalized difference built-up index (NDBI).[41–45] The eight red-edge vegetation indices include the normalized difference vegetation index Red-edge1-3 ($NDVI_{re1}$, $NDVI_{re2}$, and $NDVI_{re3}$), normalized difference Red-edge1,2 (ND_{re1} and ND_{re2}), inverted red-edge chlorophyll index (IRECI), MERIS terrestrial chlorophyll index (MTCI), and red-edge chlorophyll index (CI_{re}).[42,46] The equations to calculate these indices are listed in Table 2.

3.1.3 *Texture features*

Different ground objects have different texture features; therefore, adding texture features helps enable the full capabilities of remote sensing image information.[47,48] To facilitate tea harvests and

Table 2 Remote sensing indices used in this study and their equations for calculation.

Name	Calculation equation	Number
NDVI	$NDVI = (B8 - B4)/(B8 + B4)$	(1)
SAVI	$SAVI = (B8 - B4) \times (1+0.5)/(B8 + B4 + 0.5)$	(2)
NDWI	$NDWI = (B8 - B11)/(B8 + B11)$	(3)
MNDWI	$MNDWI = (B3 - B11)/(B3 + B11)$	(4)
NDBI	$NDBI = (B11 - B8)/(B11 + B8)$	(5)
$NDVI_{re1}$	$NDVI_{re1} = (B8A - B5)/(B8A + B5)$	(6)
$NDVI_{re2}$	$NDVI_{re2} = (B8A - B6)/(B8A + B6)$	(7)
$NDVI_{re3}$	$NDVI_{re3} = (B8A - B7)/(B8A + B7)$	(8)
ND_{re1}	$ND_{re1} = (B6 - B5)/(B6 + B5)$	(9)
ND_{re2}	$ND_{re2} = (B7 - B5)/(B7 + B5)$	(10)
IRECI	$IRECI = (B7 - B4)/(B5/B6)$	(11)
MTCI	$MTCI = (B6 - B5)/(B5 - B4)$	(12)
CI_{re}	$CI_{re} = B7/B5 - 1$	(13)

Note: NDVI, SAVI, NDWI, MNDWI, and NDBI represent the normalized difference vegetation index, soil adjusted vegetation index, normalized difference water index, modified normalized difference water index, and normalized difference built-up index, respectively. $NDVI_{re1-3}$, $ND_{re1,2}$, IRECI, MTCI, and CI_{re} represent the normalized difference vegetation index Red-edge1-3, normalized difference Red-edge1,2, inverted red-edge chlorophyll index, MERIS terrestrial chlorophyll index, and red-edge chlorophyll index, respectively. B3, B4, B5, B6, B7, B8, B8A, B11, and B12 are blue, green, red, red-edge 1, red-edge 2, red-edge 3, NIR, red-edge 4, SWIR1, and SWIR2, respectively.

obtain adequate nutrition (e.g., CO_2 and soil nutrients), tea trees are planted at certain spatial intervals.[1] Due to this unique cultivation method, tea plantations have obvious textural features that differ significantly from natural vegetation such as forests and grass.[9] Currently, the most common method of texture analysis is the gray-level co-occurrence matrix (GLCM).[49] Because the GLCM has demonstrated good adaptability and strong robustness, it was selected to extract texture information.[50]

The GLCM method requires an appropriate band when calculating texture features.[49] Because the texture features calculated based on Sentinel-2 MSI images are the most effective for distinguishing tea plantations from forests, the selected band should be highly capable of detecting vegetation changes. Considering that the resolution for our classification results was 10 m, we chose the 10-m resolution NDVI to calculate the GLCM matrix. In addition, to manage the interannual variation in NDVI, the median value of the annual NDVI was utilized in calculating the GLCM matrix. This is calculated in GEE by applying the "glcmTexture()" function; however, three parameters must first be set: size, kernel, and average.[51] Because the individual tea plantations in the study area were generally small and scattered, the size was set to 3, which means that each neighborhood was a 3×3 area. The remaining parameters used default values. We used 12 texture features with lower correlations, including angular second moment, contrast, correlation, inverse difference moment, entropy, sum average, sum entropy, difference entropy, information measure of correlation 1, information measure of correlation 2, cluster shade, and cluster prominence.

3.1.4 *Tasseled cap transformation features*

TCT features refer to those obtained by TCT, which is a result of special principal component analysis.[52] Unlike general principal component analysis, the conversion coefficient of TCT is a fixed transformation matrix.[53] TCT transforms the original multispectral image into the actual

Table 3 TCT matrix coefficients for Sentinel-2 images.

TCT features	Blue	Green	Red	NIR	SWIR1	SWIR2
Brightness	0.0822	0.1360	0.2611	0.3895	0.3882	0.1366
Greenness	−0.1128	−0.1680	−0.3480	0.3165	−0.4578	−0.4064
Wetness	0.1363	0.2802	0.3072	−0.0807	−0.4064	−0.5602

physical meaning of brightness (comprehensive surface reflectance), greenness (integration of the degree of surface vegetation coverage), and wetness (comprehensive surface water conditions) three-dimensional feature space, which can fully reflect surface reflection, vegetation coverage, and moisture information.[32] The TCT process can reduce the feature dimensions and enhance remote sensing image information. The specific transformation equation is as follows:

$$Y = CX + a, \tag{1}$$

where Y represents the image after TCT, C represents the coefficient matrix corresponding to the TCT, X represents the original multispectral image before transformation, and a is a constant added to avoid negative values. The coefficient matrix C is a fixed transformation matrix that depends on the specific sensor. The TCT matrix coefficients of Sentinel-2 images are shown in Table 3. A total of six bands are required for TCT, including blue, green, red, NIR, SWIR1, and SWIR2.

3.1.5 Topographic features

Topographic features are widely used to distinguish ground objects.[54] Tea trees have extremely strict environmental requirements for growth; thus, the terrain in a tea tree plantation should be neither too flat nor too steep.[9] In flat areas, rainfall will accumulate at the bottom of the tea trees and cause the roots to rot. In steep areas, it is difficult to maintain essential soil nutrients and moisture for tea tree growth, and tea farmers have difficulty picking tea. Therefore, tea trees can only grow in less steep mountainous or hilly areas. Therefore, terrain features can effectively improve the recognition accuracy for tea plantations and other objects, owing to the different terrain requirements for each object.[2] For example, the objects in plains are typically cropland, built-up land, and water bodies, whereas those in mountainous and hilly areas are mainly forests and grasslands with only a small percentage of cropland and built-up land.

Three topographic features were considered in this study: elevation, slope, and aspect. Elevation directly affects the temperature in mountainous areas, thereby indirectly affecting the distribution of vegetation.[11] The slope expresses the degree of steepness of a surface, which affects not only the surface energy exchange rate but also human activities and ultimately the spatial distribution of ground objects.[55] The aspect, which is the facing direction of the slope, indirectly affects the growth of vegetation and ultimately affects its spatial distribution.[5] In GEE, these topographic features can be directly calculated through the "ee.Algorithms.Terrain()" function.

3.1.6 Optimization of classification features

Considering a phenological perspective, the maximum, minimum, median, and standard deviation were calculated for the original spectral features, remote sensing indices, and red-edge vegetation indices, resulting in a total of 110 classification features. Notably, strong correlations can exist among these features, resulting in low classification accuracy and efficiency.[56] Therefore, this does not mean that more features are better, and appropriate classification features must be selected.

The J-M distance method is relatively simple and intuitive for directly measuring the distinguishability of different classes; therefore, we utilize it for feature optimization.[57] Essentially, the J-M distance uses the distance between samples of different classes to measure their

separability, thereby determining distinguishable features based on the training samples.[58] The larger the J-M distance is, the higher the separability between classes is, and the easier classification it is to them. The calculation method for J-M distance is shown as

$$\text{J-M}(c_i, c_j) = \int (\sqrt{p(x|c_i)} - \sqrt{p(x|c_j)})^2 dx, \tag{2}$$

where x is a time series, c_i and c_j are the two classes to be compared, and $p(x|c_i)$ and $p(x|c_j)$ are the conditional probability density functions for the time series x. When only considering the separability between two classes, Eq. (27) can be simplified to

$$\text{J-M} = \sqrt{2(1 - e^{-B_{ij}})}, \tag{3}$$

$$B_{ij} = \frac{1}{8}(M_i - M_j)^T \left(\frac{V_i + V_j}{2}\right)^{-1}(M_i - M_j) + \frac{1}{2} \ln\left[\left|\frac{V_i + V_j}{2}\right| / \sqrt{|V_i||V_j|}\right], \tag{4}$$

where B_{ij} is the Bhattacharyya distance of classes i and j; V_i and V_j are the sample covariance matrices of i and j, respectively; and M_i and M_j are the mean vectors of the corresponding samples. The values of the J-M distance range from 0 to $\sqrt{2}$. The larger the J-M distance is, the stronger the separability between the two classes is, with $\sqrt{2}$ corresponding to the largest separation. If $JM_{ij} \geq 1$, there is no overlap between classes i and j; if $JM_{ij} < 1$, overlap exists between classes i and j.

According to Bruzzone et al.,[59] when performing multiclass cases, the most important classification features for classification accuracy are determined by the classes with the lowest discrimination. In our classification process, tea plantations and natural forests were frequently misclassified, whereas combinations of other LULC types were easy to classify. The distinguishability between tea plantations and natural forests directly impacted the overall accuracy (OA) of classification results; therefore, the J-M distance between tea plantations and natural forests was calculated to optimize the classification features.

3.2 Progressive Random Forest Classification Approach

The progressive random forest classification method used in this study combines prior knowledge and human intervention regarding the use of training samples for machine learning. Progressive random forest classification is similar to classical supervised classification; the goal of both is to obtain the highest possible accuracy through human–computer interaction. The difference is that supervised classification improves classification accuracy by marking a large number of samples before classification, whereas progressive random forest classification uses a strategy based on classification accuracy. After a classification accuracy assessment, which sorts the probability of correct classification for each pixel from low to high, some of the pixels with probability values below a specific value (80% was used in this study) are manually added to the training sample. From the perspective of practical application, progressive random forest classification assumes that the lower accuracy samples are always located in areas that are difficult to discriminate, which can quickly and continuously promote the performance of the classifier. The overall performance of the progressive random forest classification model is controllable and does not fall into a local optimum.

The progressive random forest classification model reflects the human learning process, progressing from easy to difficult. The model is trained using a small number of training samples based on classification accuracy, and complex samples are gradually added to improve the classification accuracy. In our study, a typical tea plantation distribution area was first selected, and a small number of training samples were used for classification. Subsequently, the validation samples were used to calculate the OA and the confidence of each pixel. If the OA was high, the classification area was expanded. If the OA was low, low-confidence pixels were added to the training sample and reclassified. After reclassification, if the OA improved, the newly added training samples were retained; if the OA decreased, the newly added training samples were

removed and reselected. The above steps were repeated, expanding the classification area by gradually increasing the number of training samples; thus, the classification accuracy was gradually improved until the entire study area achieved a satisfactory classification accuracy.

As this study aimed to classify tea plantations over a large area, it was preferable to add low-precision tea plantation sample points before other types when applying the progressive random forest method. This ensured that the classification result for each pixel point within a tea plantation has high confidence. When applying the random forest model in GEE, two parameters must be set: numberOfTrees and minLeafPopulation. After repeated verification, we set numberOfTrees to 100 and minLeafPopulation to 10.

3.3 *Postprocessing and Accuracy Assessment*

There were a small number of blank values in the classification result, which directly impacted the classification accuracy. There are two main reasons that blank values are generated. One reason is that some pixels have classification features outside the normal range, resulting in empty classification results. For these pixels, we rechecked the values for all classification features, eliminated outliers, and then reclassified them. The other reason is that some pixels are recognized as clouds by the cloud recognition method. For these, a multivalue within the 3×3 neighborhood of these pixels was assigned.

Tea trees prefer shade; hence, many tea plantations are in the shadows of hills. Thus, some of these plantations were easily misclassified as natural forests; therefore, we also used the 3×3 neighboring pixels to eliminate the influence of the hill shadows.

After postprocessing, we used a 10-fold cross-validation to assess the classification accuracy. We calculated the classification error matrix based on the classification results, which were quantified using the OA. The OA is the sum of the diagonal elements divided by the sum of all of the elements of the confusion matrix. Then, we calculated the producer's accuracy (PA), user's accuracy (UA), and F_1-score. The F_1-score is particularly useful for class-level accuracy assessment because it assigns equal importance to the PA and UA. Finally, the above process was repeated 10 times, and the average OA, PA, UA, and F_1-score were calculated. The PA, UA, and F_1-score are calculated for each class as follows:

$$\text{PA}_i = \sum_{j=1}^{r} \frac{n_{ii}}{n_{ij}}, \tag{5}$$

$$\text{UA}_i = \sum_{j=1}^{r} \frac{n_{ii}}{n_{ji}}, \tag{6}$$

$$(F_1)_i = \frac{2 \times \text{PA}_i \times \text{UA}_i}{\text{PA}_i + \text{UA}_i}, \tag{7}$$

where r is the number of classes and n_{ij} is the element of the confusion matrix in the i'th column and j'th row, that is, the count of elements of class j classified as class i. PA_i, UA_i, and $(F_1)_i$ stand for the PA, UA, and F_1-score for class i, respectively.

4 Results

4.1 *Optimized Classification Features*

In this study, 110 classification features were optimized to ensure classification efficiency and achieve a higher classification accuracy. Both the topographic and TCT features had only three features and weak correlations; hence, they were excluded from the feature optimization process. Therefore, the rest of the classification features, including the original bands, remote sensing indices, and texture features, were optimized using the J-M distance. The feature optimization process was conducted in the following two steps. First, 1000 tea plantation samples and 1000 natural forest samples were randomly selected from the training samples. Then, the J-M distance

Table 4 Optimization results for the classification features.

	Before optimization		After optimization	
	Number	J-M	Number	J-M
Original bands	40	1.030	27	0.870
Remote sensing indices	20	0.983	15	0.895
Red-edge indices	32	1.399	7	1.320
Texture features	12	1.198	9	1.183
TCT features	3	0.528	3	0.528
Topographic features	3	0.370	3	0.370
Total	116	1.414	64	1.412

for each feature was calculated (see Table 8 for details), and the features with J-M distance values lower than a certain standard were eliminated.

The J-M distance can determine the discrimination between different classification bands but not the number of optimal features for classification. Here, we selected the optimal feature bands based on Lorenzo Bruzzone et al.[59] As shown in Table 8, the maximum, minimum, and median values of the J-M distance of the original bands were 0.483, 0.085, and 0.229, respectively. Thus, the tea plantations and natural forests were difficult to distinguish. The original bands with J-M distance values greater than 0.2 were selected. Similarly, the characteristics of tea plantations and natural forests were also very similar in the remote sensing indices. Those with J-M distances greater than 0.2 were also selected. However, the distinguishability between tea plantations and natural forests was very strong for the red-edge vegetation indices, which are more sensitive concerning vegetation. Therefore, the red-edge vegetation indices with J-M distances greater than 0.4 were selected. A brief description of the optimized classification features, including the number and J-M distance values, is shown in Table 4 (for details, see Table 9). The total number of classification features was reduced from 116 to 64, a 45% decrease, which dramatically improved the classification efficiency. The overall J-M distance values varied from 1.414 for the original classification features to 1.412 for the optimized classification features, which is a difference of 0.14%; therefore, a comparable result was obtained after the feature optimization process. From the perspective of balancing the classification efficiency and accuracy, the feature optimization process performed excellently.

4.2 Classification Results for the Tea Plantations and Other Land Cover Types

Figures 3 and 4 show the results of the tea plantation classification using the progressive random forest classification model from a regional and local perspective, respectively. As shown in Fig. 3, tea plantations were scattered throughout in Anhui Province. Tea plantations were mainly distributed in the southern and western study area, which predominantly have a mountainous terrain. In the middle of the study area, there were a small number of tea plantations in hilly terrain. There were few tea plantations in the northern study area or along the Yangtze River; these places are plains that are not suitable for the growth of tea trees. The city with the most tea plantations was Huangshan City, which was consistent with the statistical data.

Figure 4 shows the classification results for three typical regions. The tea plantations identified using the progressive classification model were generally consistent with reality. In hilly areas, tea plantations were well identified, and the cropland and forest classifications were accurate [Fig. 4(b)]. In the transitional regions between hills and mountains, the overall tea plantation classifications were also relatively accurate; however, a small amount of forest was misclassified as tea plantations [Fig. 4(d)]. The classification results for forests and water bodies were more accurate; however, some cropland was misclassified as built-up land. In mountainous areas, the
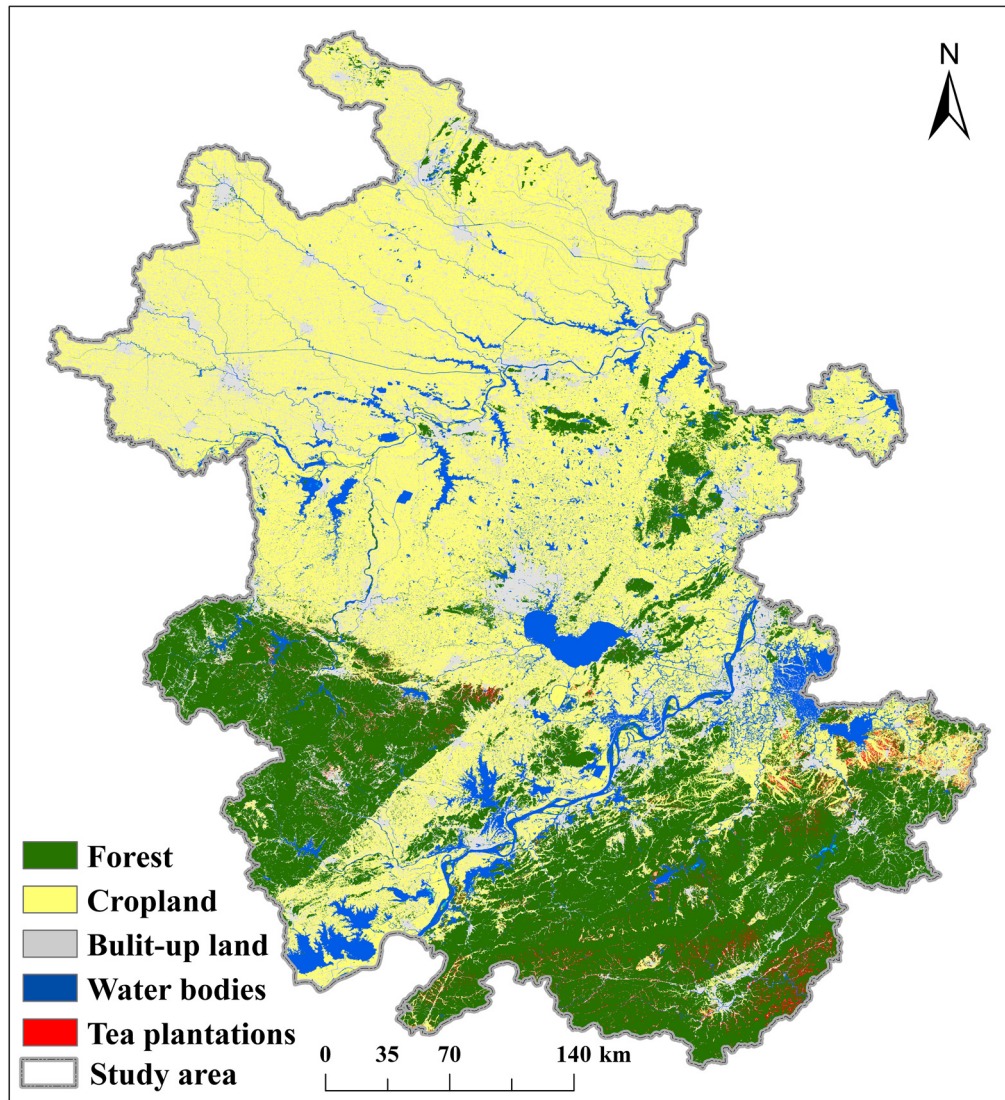
Fig. 3 Map of the tea plantations in the study area in 2020.

classification results for tea plantations and forests were accurate; however, some cropland was again misclassified as built-up land [Fig. 4(f)]. In general, tea plantations were accurately classified in the main distribution areas.

4.3 Classification Accuracy Analysis

To test the classification accuracy for tea plantations, the OA, UA, PA, and F_1-score were calculated quantitatively using the confusion matrix. In the quantitative analysis, a 10-fold cross-validation was used to determine the stability of the classification model. The calculation results are presented in Table 5.

The average OA reached 89%. The classification accuracy for cropland, forest, and water bodies was greater than 85%, and that for tea plantations reached 82%. With the spatial resolution and the large area of Anhui Province, this result is generally satisfactory. However, the classification accuracies for grassland and unused land were poor, which was mainly caused by their small distribution areas within the study area and the small number of sample points used during classification. Nevertheless, these low accuracies had little influence on the identification of tea plantations because of their small areas and low probability of misclassification as tea plantations.
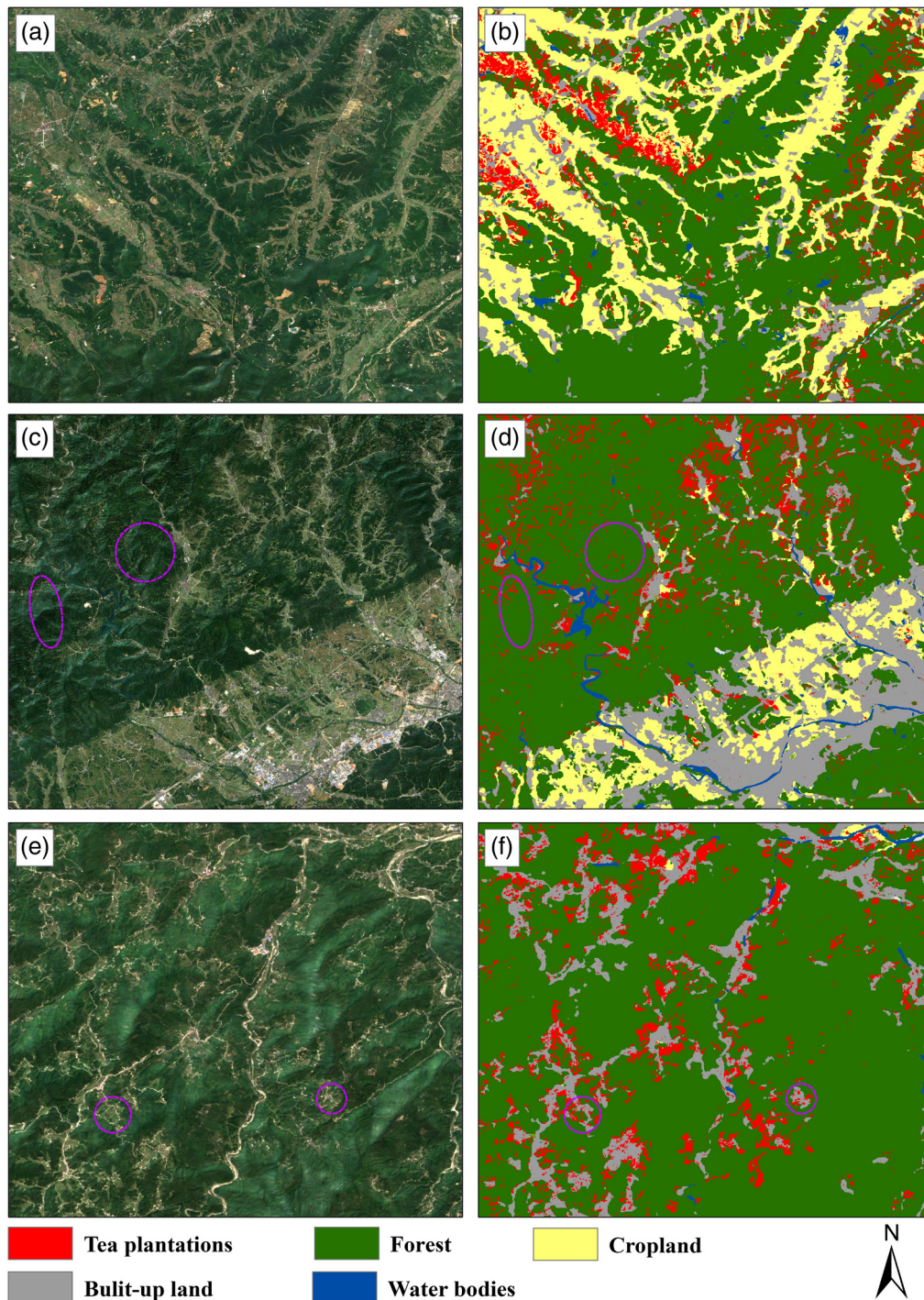
Fig. 4 Comparison of the classification results in typical areas. (a), (c), (e) Sentinel-2 MSI true-color images of a hilly region, hill-to-mountain transition region, and mountain region, respectively, and (b), (d), and (f) the corresponding classification results.

5 Discussion

5.1 *Impacts of Different Feature Selection Schemes*

To compare the impacts of different classification features on the classification accuracy for tea plantations, this study designed 10 different classification feature combination schemes. Because the red-edge bands had the largest J-M distance, comparisons with and without these bands were

Table 5 Classification accuracy for results in the study area.

	UA (%)	PA (%)	F_1-score
Tea plantations	82.04 ± 0.03	86.33 ± 0.02	0.8413
Cropland	88.67 ± 0.02	87.67 ± 0.01	0.8817
Forest	95.50 ± 0.02	94.74 ± 0.01	0.9512
Grassland	76.08 ± 0.04	64.95 ± 0.10	0.7007
Water bodies	90.15 ± 0.01	91.97 ± 0.01	0.9105
Built-up land	85.60 ± 0.02	87.58 ± 0.01	0.8658
Unused land	58.49 ± 0.06	46.56 ± 0.11	0.5185
OA = 89.27 ± 0.03%, Kappa = 0.8831			

Note: UA, PA, and OA represent the user's accuracy, producer's accuracy, and overall accuracy, respectively. The values of the UA, PA, and OA are the mean ± standard deviation.

carried out to analyze whether they have prominent effects when designing the feature combination method. Due to the large study area and the number of comparative experiments, only Huangshan City, which has the most tea plantations, was selected for comparative experiments. The experimental results are presented in Table 6.

Table 6 shows that the combination of the original bands and the remote sensing indices obtained a high OA value of 83.99% because the constructed classification feature space could extract the phenological features of each class. The red-edge indices had the most significant impacts on improving the classification accuracies for tea plantations due to their high sensibility for vegetation. The topographical features had the second largest positive effect on classification accuracy. Tea plantations have strict environmental requirements; therefore, elevation, slope, and aspect all have direct effects on the growth of tea trees. The TCT features also showed positive effects because they comprehensively reflect the brightness, greenness, and wetness of ground objects. For example, the water in the roots of tea plantations and natural forests differs, resulting in a difference in TCT features. Tea trees are planted at certain spatial intervals to facilitate both growth and harvesting; thus, their greenness is generally lower than that of natural forests. The

Table 6 Classification accuracies of different classification feature combinations in Huangshan city.

	UA (%)	PA (%)	OA (%)	F_1-score
Scene1 (original bands+ remote sensing indices)	86.10	96.03	83.99	0.9079
Scene2 (original bands+ remote sensing indices+ red-edge indices)	87.45	96.1	85.39	0.9157
Scene1 + texture features	87.9	95.32	85.1	0.9146
Scene2 +texture features	88.61	95.17	85.9	0.9177
Scene1 + TCT features	87.47	95.08	84.58	0.9112
Scene2 + TCT features	88.49	95.02	85.65	0.9164
Scene1 + topographic features	86.89	95.81	84.63	0.9113
Scene2 + topographic features	87.8	96.1	85.82	0.9176
All	89.12	95.25	86.50	0.9208
Optimized features	88.67	95.33	86.41	0.9188

smallest improvement in the accuracy was from the texture features. This was mainly because the spatial resolution of the Sentinel-2 data was 10 m, which does not fully reflect the textural features of the tea plantations, resulting in a minor improvement.

When using all features, the UA for tea plantations was 89.12%, the PA was 95.25%, and the F_1-score was 0.9208. In addition, these accuracy indices obtained using the optimized features were 88.67%, 95.33%, and 0.9188, respectively. Although the UA is slightly decreased when using the optimized features, the PA improved, and the F_1-score, which measures the comprehensive classification accuracy, dropped by 0.002. The OA using all features was 86.50%, whereas that using the optimized features was 86.41%, which is a decrease of 0.09%. Although the use of optimized features resulted in a slight loss of accuracy, the number of classified features was dramatically reduced from 116 to 64. A reduction in the number of features can significantly improve the classification efficiency. When using the progressive random forest model, it is necessary to continuously add training samples and then reclassify them. Therefore, compared with the slight loss of accuracy, the improvement in the classification efficiency is more important.

5.2 Improvement in Classification Accuracy by Progressive Random Forest

In this study, we designed a progressive random forest model that combines random forest and progressive models. The progressive random forest model improves classification accuracy by gradually increasing the number of training samples on GEE. These newly added training samples were based on the probable classification result for each pixel. A total of 10 iterative classifications were performed, and the results are shown in Fig. 5.

Figure 5 shows that, in the process of iterative classification, the OA, UA, and PA of the tea plantations constantly improved. This shows that the proposed progressive random forest model was effective. For the first iteration, the samples were derived from the results of our field surveys in Huangshan and Wuhu. Huangshan and Wuhu are typical mountainous and hilly tea-producing cities in the study area, respectively. For the first iteration, the samples were concentrated in mountainous and hilly areas, resulting in low OA, UA, and PA for the tea plantations. Then, the visually selected training samples for other cities were gradually added according to the accuracies of the classification results. With additional training samples, the OA, UA, and PA greatly improved. On the 10th iteration, the OA and PA exceeded 85%, which met our objectives; hence, we stopped the classification process.

The first visual samples were from the areas surrounding Huangshan and Wuhu; subsequently, the area was gradually expanded. The results of the visually selected additions and land survey samples are shown in Fig. 6. The figure shows that there were more samples in the
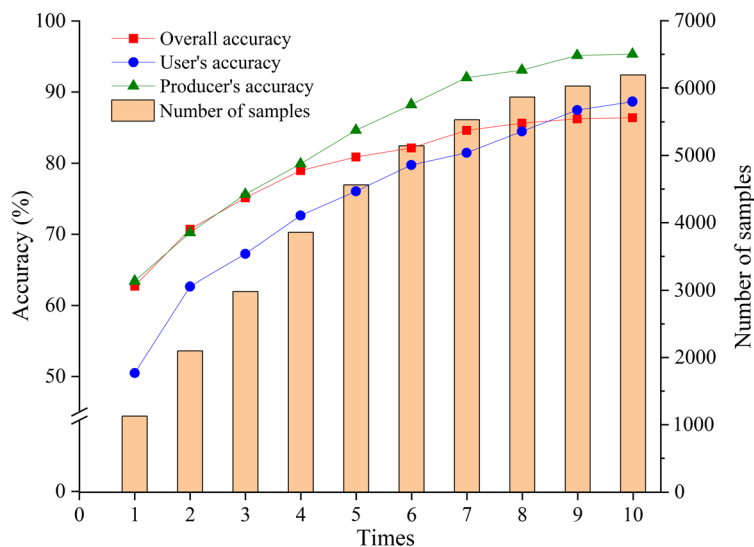

Fig. 5 The accuracy of the progressive random forest classification with the number of samples used during the classification process.
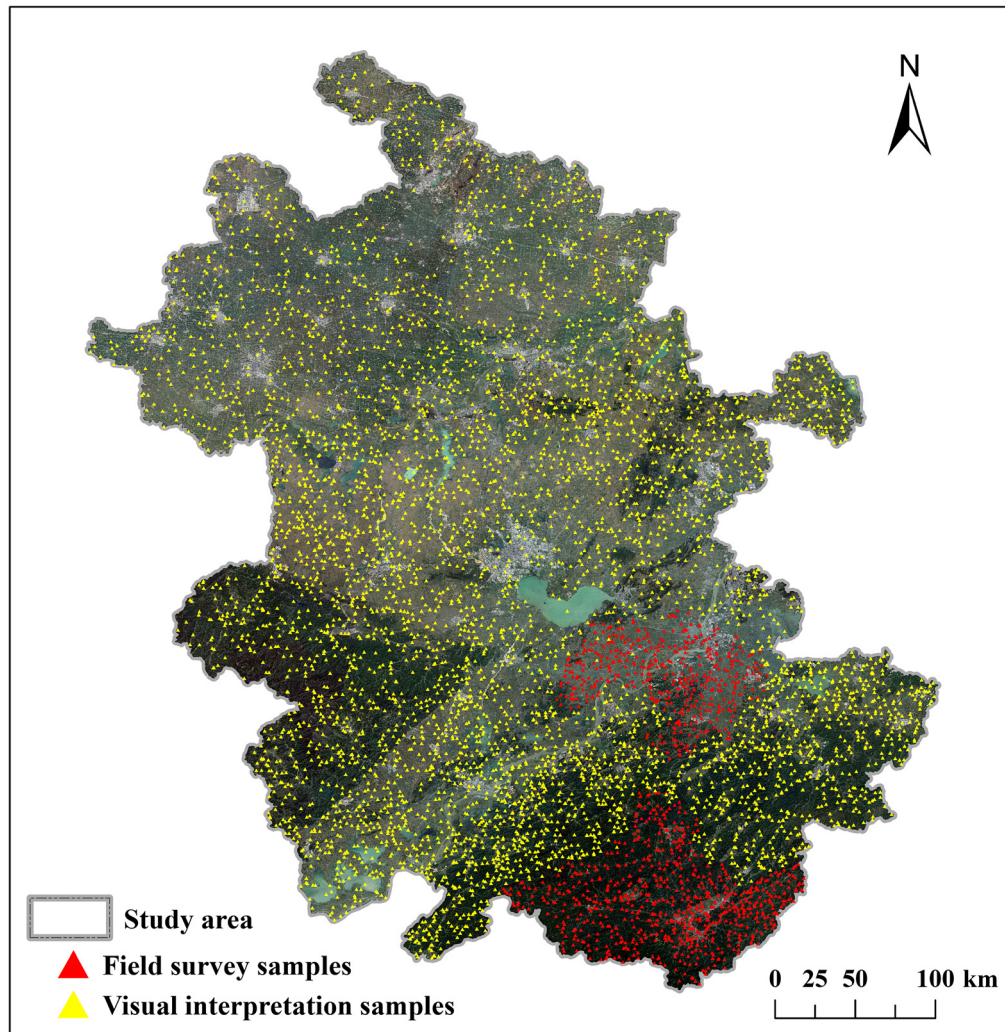
Fig. 6 The results of the field survey and the addition of visually selected training samples.

southern study area but few samples in the northern region. This is because tea plantations and natural forests were initially misclassified; thus, we added samples of tea plantations and natural forests. With the increase in natural forest samples, tea plantations were mainly misclassified as croplands. Therefore, we added samples of cropland in the middle of the study area. When the UA and PA of the tea plantations exceeded 80%, we added samples from other classes to improve the OA.

5.3 *Advantages and Disadvantages*

Because the spectral bands of tea plantations and natural forests are highly similar, obtaining high-precision classification of tea plantations over a large area remains challenging. With ready-made remote-sensing cloud platforms (e.g., GEE) and the use of nonparametric classifiers, the main challenge becomes selecting suitable classification bands and training samples.

The high classification accuracy and F_1-score for tea plantations and other main land use classes demonstrate the practicality of the progressive random forest method. The effectiveness of this method depends on accessing sufficient computing power. In this study, this was not a problem because the GEE provided sufficient computing power.

In this study, a new method for the classification of sentinel-2 MSI data, which was combined with a random forest and progressive model running on GEE, was proposed. This is advantageous for the following reasons: (1) it retains the advantages of the random forest and the progressive model, which can effectively and robustly run high-dimensional data for a large

Table 7 Overall accuracies of different comparative experiment models in Huangshan city.

Number of samples	CART (%)	SVM (%)	DNN (%)
1128	44.05	65.37	10.19
2098	50.19	70.80	22.36
2980	55.05	76.56	30.50
3856	59.39	79.60	39.23
4563	62.24	83.08	46.77
5143	67.52	83.89	54.38
5532	72.76	83.08	62.56
5864	77.28	83.18	69.77
6027	79.52	83.08	73.83
6193	76.90	83.60	75.96

area. (2) It can classify sparse objects (such as the tea plantations in this study) with a high classification accuracy. (3) The entire classification process is carried out on GEE to ensure highly efficient overall workflow.

We used classification and regression trees (CART), support vector machine (SVM), and deep neural networks (DNN) to verify the generality and reliability of our conclusions. In the comparative experiments, the samples used for the CART, SVM, and DNN models were the same as for the RF models. Default parameters were set for CART and SVM in GEE. The Keras library was utilized for the DNN hyperparameter optimization in Google Colab, and the key parameters of DNN were set as follows: 64–128–256 for the hidden layers, 10 for the epochs, ReLU (the rectified linear units) for the activation function, Adam for the optimizer, and 20% for the dropout ratio. The OA values of the comparative experiment classifications are shown in Table 7. In general, when a small number of samples was used, a low OA value was obtained. In addition, the OA values of the compared classification models increased with the increase in the number of samples, indicating that increasing the number of samples could significantly improve classification accuracy, which was similar to the RF model. However, the contributions of increasing samples to the OA were found to be different among the compared classification models. The OA values of the CART model were moderately improved with the increase of samples, but the highest OA of the CART model (79.52%) was significantly lower than that of the RF model (89.27%). The SVM model only needs a small number of samples to obtain high classification accuracy, but it easily overfits and its highest OA (83.89%) was also lower than that of the RF model. Increasing the number of samples can dramatically improve the classification accuracy of the DNN model, which needs a large number of samples to obtain high classification accuracy. Although the highest OA of DNN (75.96%) was significantly lower than that of the RF model, it is worth noting that the classification accuracy value of the DNN model continuously improved as the samples increased, which means that it may be close to or exceed the highest classification accuracy of the RF model if the number of samples is more than 6193. For balancing the classification accuracy and the limit number of samples, the combination of the progressive model and the RF model is the optimal choice.

Although our progressive random forest model has many advantages, it also has some disadvantages. First, because the GEE visualization interactive platform only provides current remote sensing images, the training data added by manual visual inspection only utilizes the current data, and the generation of sample points using historical data requires other platforms. Second, progressive random forest is semiautomated. Manual assignment is required each time that new samples are added to the training sample dataset. In future studies, we will consider introducing models (such as LandTrender) to generate long-term sample data. We are also considering adding other existing land use/land cover products to reduce manual assignment and improve the automation of the classification workflow.

6 Conclusions

A methodology was proposed for the classification of tea plantations in Anhui Province in 2020. The J-M model distance was applied to the original bands, remote sensing indices, texture features, TCT features, and terrain features to select the optimal combination of classification features. A progressive random forest model was used to classify 3554 scenes from Sentinel-2 MSI images. Using these methods, we obtained a map of tea plantations in Anhui Province in 2020 and analyzed its classification accuracy.

The progressive random forest classification method used in this study combines the advantages of the random forest and progressive models. First, remote sensing classification for a large area was carried out quickly by utilizing random forest. Moreover, owing to the advantages of the progressive model, the OA of classification remained high even with a small number of training samples. Furthermore, the progressive random forest classification model accurately identified the spatial distribution of tea plantations.

By comparing the J-M distance, the number of classification features was reduced dramatically, which improves the classification efficiency. Moreover, high classification accuracy was obtained using the reduced classification features.

In addition, when using only the original bands and remote sensing indices, good classification results were still obtained. This is because the calculations of statistical values, such as the maximum, minimum, and standard deviation, reflect the phenological characteristics of the vegetation. Furthermore, the red-edge indices are the most effective for the identification of tea plantations, followed by topographical features. The TCT and texture features have minimal positive effects on improving the classification accuracy for tea plantations.

In future studies, we will explore the effects of applying other machine learning models to the classification process along with the progressive model to obtain a more precise map of tea plantations. Meanwhile, we will consider extending our research workflow using Landsat time-series data to obtain long-term spatial distributions and changes in tea plantations.

7 Appendix A

7.1 *J-M Distance Results of Classification Features*

Table 8 shows the details of the calculated J-M distance for each feature. The low value of the J-M distance represents low separability and the features with J-M distance values lower than a certain standard (i.e., 0.2) were further eliminated. Table 9 shows the details of the optimized classification features, including the number and the J-M distance values. The total number of classification features was reduced from 116 to 64, a 45% decrease, which dramatically improved the classification efficiency.

Table 8 J-M distance results of classification features.

Name	J-M	Name	J-M	Name	J-M	Name	J-M
B2-max	0.085	B2-min	0.351	B2-median	0.361	B2-std	0.124
B3-Max	0.093	B3-Min	0.341	B2-median	0.400	B3-Std	0.125
B4-Max	0.095	B4-Min	0.347	B3-median	0.400	B4-Std	0.116
B5-Max	0.116	B5-Min	0.311	B4-median	0.450	B5-Std	0.114
B6-Max	0.171	B6-Min	0.316	B5-median	0.440	B6-Std	0.150
B7-Max	0.207	B7-Min	0.321	B6-median	0.453	B7-Std	0.148
B8-Max	0.213	B8-Min	0.306	B7-median	0.446	B8-Std	0.104
B8A-Max	0.229	B8A-Min	0.309	B8-median	0.455	B8A-Std	0.120

Table 8 (*Continued*).

Name	J-M	Name	J-M	Name	J-M	Name	J-M
B11-Max	0.451	B11-Min	0.264	B8A-median	0.483	B11-Std	0.387
B12-Max	0.350	B12-Min	0.263	B11-median	0.473	B12-Std	0.347
NDVI-Max	0.279	NDVI-Min	0.171	B12-median	0.291	NDVI-Std	0.297
SAVI-Max	0.279	SAVI-Min	0.170	NDVI-median	0.292	SAVI-Std	0.296
NDWI-Max	0.206	NDWI-Min	0.307	SAVI-median	0.374	NDWI-Std	0.281
MNDWI-Max	0.141	MNDWI-Min	0.253	MNDWI-median	0.452	MNDWI-Std	0.238
NDBI-Max	0.216	NDBI-Min	0.049	NDBI-median	0.238	NDBI-Std	0.163
$NDVI_{re1}$-Max	0.164	$NDVI_{re1}$-Min	0.160	$NDVI_{re1}$-median	0.195	$NDVI_{re1}$-Std	0.265
$NDVI_{re2}$-Max	0.283	$NDVI_{re2}$-Min	0.234	$NDVI_{re2}$-median	0.296	$NDVI_{re2}$-Std	0.294
$NDVI_{re3}$-Max	0.636	$NDVI_{re3}$-Min	0.438	$NDVI_{re3}$-median	0.226	$NDVI_{re3}$-Std	0.568
ND_{re1}-Max	0.221	ND_{re1}-Min	0.158	ND_{re1}-median	0.243	ND_{re1}-Std	0.293
ND_{re2}-Max	0.139	ND_{re2}-Min	0.174	ND_{re2}-median	0.174	ND_{re2}-Std	0.272
IRECI-Max	0.256	IRECI-Min	0.243	IRECI-median	0.296	IRECI-Std	0.133
MTCI-Max	0.920	MTCI-Min	0.286	MTCI-median	0.153	MTCI-Std	0.583
CI_{re}-Max	1.09	CI_{re}-Min	0.163	CI_{re}-median	0.152	CI_{re}-Std	0.947
ASM	0.411	CONTRAST	0.428	CORR	0.171	IDM	0.421
ENT	0.370	SAVG	0.325	SENT	0.468	DENT	0.466
IMCORR1	0.432	IMCORR2	0.584	SHADE	0.476	PROM	0.589

Note: Max, Min, median, and Std represent the maximum, minimum, median, and standard deviation, respectively.

7.2 *Optimization Results of Classification Features*

Table 9 Optimization results of classification features.

Type	Before optimization			After optimization		
	Original features	Number	J-M	Optimized features	Number	J-M
Original bands	The maximum, minimum, median, and standard deviation of the original bands 2-8, 8A, 11, 12	40	1.03	B7-Max, B8-Max, B8A-Max, B11-Max, B12-Max, B2-Min, B3-Min, B4-Min, B5-Min,B6-Min, B7-Min, B8-Min, B8A-Min, B11-Min, B12-Min, B2-median, B3-median, B4-median, B5-median, B6-median, B7-median, B8-median, B8A-median, B11-median, B12-median, B11-Std, B12-Std	27	0.870
Remote sensing indices	The maximum, minimum, median, and standard deviation of NDVI, SAVI, NDWI, MNDWI, and NDBI	20	0.983	NDVI-Max, SAVI-Max, NDWI-Max, NDBI-Max, NDWI-Min, MNDWI-Min, NDVI-median, SAVI-median, NDWI-median, NDBI-median, MNDWI-median, NDVI-Std, SAVI-Std, NDWI-Std, MNDWI-Std	15	0.895

Table 9 (*Continued*).

Type	Before optimization			After optimization		
	Original features	Number	J-M	Optimized features	Number	J-M
Red-edge indices	The maximum, minimum, median, and standard deviation of $NDVI_{re1}$, $NDVI_{re2}$, $NDVI_{re3}$, ND_{re1}, ND_{re2}, IRECI, MTCI, and CI_{re}	32	1.399	CI_{re}-Max, CI_{re}-Std, MTCI-Max, $NDVI_{re3}$-Max, MTCI-Std, $NDVI_{re3}$-Std, $NDVI_{re3}$-Min	7	1.320
Texture features	ASM,CONTRAST, CORR,IDM,ENT,SAVG, SENT,DENT,IMCORR1, IMCORR2,SHADE, PROM	12	1.198	ASM, CONTRAST, IDM, SENT, DENT, IMCORR1, IMCORR2, SHADE, PROM	9	1.183
TCT features	Brightness, greenness, wetness	3	0.528	Brightness, greenness, wetness	3	0.528
Topographic features	Elevation, slope, aspect	3	0.370	Elevation, slope, aspect	3	0.370
ALL	—	116	1.414	—	64	1.412

Acknowledgments

References

1. W. Xu et al., "The pruning phenological phase-based method for extracting tea plantations by field hyperspectral data and Landsat time series imagery," *Geocarto Int.*, 1–21 (2020).
2. Y. Chen and S. Tian, "Feature-level fusion between Gaofen-5 and Sentinel-1A data for tea plantation mapping," *Forests* **11**(12), 1357 (2020).
3. W. Xu et al., "Quantifying spatial-temporal changes of tea plantations in complex landscapes through integrative analyses of optical and microwave imagery," *Int. J. Appl. Earth Obs. Geoinf.* **73**, 697–711 (2018).
4. Bureau of Statistics of Anhui, "Anhui Statistical Yearbook 2020," http://www.ahtjj.gov.cn/tjj/web/tjnj_view.jsp (accessed 10 July 2021).
5. B. Wang et al., "Mapping tea plantations from multi-seasonal landsat-8 OLI imageries using a random forest classifier," *J. Indian Soc. Remote Sens.* **47**(8), 1315–1329 (2019).
6. Y.-C. Wang et al., "Diverse Colletotrichum species cause anthracnose of tea plants (*Camellia sinensis* (L.) O. Kuntze) in China," *Sci. Rep.* **6**(1), 35287 (2016).
7. S. Su, "Economic benefit and ecological cost of enlarging tea cultivation in subtropical China: characterizing the trade-off for policy implications," *Land Use Policy* **66**, 183–195 (2017).
8. P. Phan et al., "Using multi-temporal MODIS NDVI data to monitor tea status and forecast yield: a case study at Tanuyen, Laichau, Vietnam," *Remote Sens.* **12**(11), 1814 (2020).
9. Z. Tang, M. Li, and X. Wang, "Mapping tea plantations from VHR images using OBIA and convolutional neural networks," *Remote Sens.* **12**(18), 2935 (2020).
10. R. R. Nidamanuri, "Hyperspectral discrimination of tea plant varieties using machine learning, and spectral matching methods," *Remote Sens. Appl. Soc. Environ.* **19**, 100350 (2020).

11. N. Li et al., "Mapping the spatial distribution of tea plantations using high-spatiotemporal-resolution imagery in northern Zhejiang, China," *Forests* **10**(10), 856 (2019).

12. P. Mishra et al., "Fusing spectral and textural information in near-infrared hyperspectral imaging to improve green tea classification modelling," *J. Food Eng.* **249**, 40–47 (2019).

13. J. Zhu et al., "An improved multi-temporal and multi-feature tea plantation identification method using Sentinel-2 imagery," *Sensors* **19**(9), 2087 (2019).

14. L. Valderrama-Landeros et al., "Extrapolating canopy phenology information using Sentinel-2 data and the Google Earth Engine platform to identify the optimal dates for remotely sensed image acquisition of semiarid mangroves," *J. Environ. Manage.* **279**, 111617 (2021).

15. R. Sonobe, T. Sano, and H. Horie, "Using spectral reflectance to estimate leaf chlorophyll content of tea with shading treatments," *Biosyst. Eng.* **175**, 168–182 (2018).

16. H.-J. Chu et al., "Integration of full-waveform LiDAR and hyperspectral data to enhance tea and areca classification," *GISci. Remote Sens.* **53**(4), 542–559 (2016).

17. X. Liu et al., "Large-scale crop mapping from multisource remote sensing images in Google Earth Engine," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **13**, 414–427 (2020).

18. J. M. A. Duncan et al., "Observing climate impacts on tea yield in Assam, India," *Appl. Geogr.* **77**, 64–71 (2016).

19. Y.-C. Chuang and Y.-S. Shiu, "A comparative analysis of machine learning with worldview-2 pan-sharpened imagery for tea crop mapping," *Sensors* **16**(5), 594 (2016).

20. M. Calderón-Loor, M. Hadjikakou, and B. A. Bryan, "High-resolution wall-to-wall land-cover mapping and land change assessment for Australia from 1985 to 2015," *Remote Sens. Environ.* **252**, 112148 (2021).

21. L. Liu et al., "Mapping cropping intensity in China using time series Landsat and Sentinel-2 images and Google Earth Engine," *Remote Sens. Environ.* **239**, 111624 (2020).

22. N. Gorelick et al., "Google Earth Engine: planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.* **202**, 18–27 (2017).

23. M. N. Jahromi et al., "Google Earth Engine and its application in forest sciences," in *Spatial Modeling in Forest Resources Management*, pp. 629–649, Springer, Cham (2021).

24. L. Kumar and O. Mutanga, "Google Earth Engine applications," *Remote Sens.* **11**(5), 591 (2019).

25. L. Chu et al., "Assessing the link between human modification and changes in land surface temperature in Hainan, China using image archives from Google Earth Engine," *Remote Sens.* **12**(5), 888 (2020).

26. M.-S. Zhao et al., "Spatial-temporal change of soil organic carbon in Anhui Province of East China," *Geoderma Regional* **26**, e00415 (2021).

27. Z.-G. Cao et al., "Spatio-temporal pattern of schistosomiasis in Anhui Province, East China: potential effect of the Yangtze River-Huaihe river water transfer project," *Parasitol. Int.* **67**(5), 538–546 (2018).

28. Y. Chen et al., "Spatiotemporal characteristics of winter wheat waterlogging in the middle and lower reaches of the Yangtze River, China," *Adv. Meteorol.* **9**, 1–11 (2018).

29. Q.-Q. Zhu et al., "Epidemiological characteristics of pulmonary tuberculosis in Anhui Province, Eastern China from 2013 to 2018," *PLoS One* **15**(8), e0237311 (2020).

30. P. Griffiths et al., "Towards national-scale characterization of grassland use intensity from integrated Sentinel-2 and Landsat time series," *Remote Sens. Environ.* **238**, 111124 (2020).

31. C. Granero-Belinchon et al., "Phenological dynamics characterization of alignment trees with Sentinel-2 imagery: a vegetation indices time series reconstruction methodology adapted to urban areas," *Remote Sens.* **12**(4), 639 (2020).

32. P. Macintyre, A. van Niekerk, and L. Mucina, "Efficacy of multi-season Sentinel-2 imagery for compositional vegetation classification," *Int. J. Appl. Earth Obs. Geoinf.* **85**, 101980 (2020).

33. S. Qiu, Z. Zhu, and B. He, "Fmask 4.0: improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery," *Remote Sens. Environ.* **231**, 111205 (2019).

34. P. Li et al., "A new quality validation of global digital elevation models freely available in China," *Surv. Rev.* **48**(351), 409–420 (2016).

35. A. Mouratidis and D. Ampatzidis, "European digital elevation model validation against extensive global navigation satellite systems data and comparison with SRTM DEM and ASTER GDEM in Central Macedonia (Greece)," *ISPRS Int. J. Geo-Inf.* **8**(3), 108 (2019).

36. G. Pulighe, V. Baiocchi, and F. Lupia, "Horizontal accuracy assessment of very high resolution Google Earth images in the city of Rome, Italy," *Int. J. Digital Earth* **9**(4), 342–362 (2016).

37. J. J. Arsanjani, A. Tayyebi, and E. Vaz, "GlobeLand30 as an alternative fine-scale global land cover map: challenges, possibilities, and implications for developing countries," *Habitat Int.* **55**, 25–31 (2016).

38. V. Upadhyay and A. Kumar, "Hyperspectral remote sensing of forests: technological advancements, opportunities and challenges," *Earth Sci. Inf.* **11**(4), 487–524 (2018).

39. J. M. Peña-Barragán et al., "Object-based crop identification using multiple vegetation indices, textural features and crop phenology," *Remote Sens. Environ.* **115**(6), 1301–1316 (2011).

40. P. Hurskainen et al., "Auxiliary datasets improve accuracy of object-based land use/land cover classification in heterogeneous savanna landscapes," *Remote Sens. Environ.* **233**, 111354 (2019).

41. P. D'Odorico et al., "Experimental evaluation of Sentinel-2 spectral response functions for NDVI time-series continuity," *IEEE Trans. Geosci. Remote Sens.* **51**(3), 1336–1348 (2013).

42. W. J. Frampton et al., "Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation," *ISPRS J. Photogramm. Remote Sens.* **82**, 83–92 (2013).

43. X. Yang et al., "Mapping of urban surface water bodies from Sentinel-2 MSI imagery at 10 m resolution via NDWI-based image sharpening," *Remote Sens.* **9**(6), 596 (2017).

44. Y. Du et al., "Water bodies' mapping from Sentinel-2 imagery with modified normalized difference water index at 10-m spatial resolution produced by sharpening the SWIR band," *Remote Sens.* **8**(4), 354 (2016).

45. M. Piragnolo et al, "Comparison of vegetation indices from RPAS and sentinel-2 imagery for detecting permanent pastures," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **XLII-3**(3), 1381–1387 (2018).

46. C. Li et al., "Estimating apple tree canopy chlorophyll content based on Sentinel-2A remote sensing imaging," *Sci. Rep.* **8**(1), 1–10 (2018).

47. S. Filin, "Surface clustering from airborne laser scanning data," *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **34**(3/A), 119–124 (2002).

48. Q. Feng, J. Liu, and J. Gong, "UAV remote sensing for urban vegetation mapping using random forest and texture analysis," *Remote Sens.* **7**(1), 1074–1094 (2015).

49. P. Mohanaiah, P. Sathyanarayana, and L. GuruKumar, "Image texture feature extraction using GLCM approach," *Int. J. Sci. Res. Publ.* **3**(5), 1–5 (2013).

50. D. Gadkari, *Image Quality Analysis Using GLCM*, University of Central Florida Orlando, Florida (2004).

51. A. Tassi and M. Vizzari, "Object-oriented LULC classification in Google Earth Engine combining SNIC, GLCM, and machine learning algorithms," *Remote Sens.* **12**(22), 3776 (2020).

52. T. Shi and H. Xu, "Derivation of tasseled cap transformation coefficients for Sentinel-2 MSI at-sensor reflectance data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **12**(10), 4038–4048 (2019).

53. J. Lastovicka et al., "Sentinel-2 data in an evaluation of the impact of the disturbances on forest vegetation," *Remote Sens.* **12**(12), 1914 (2020).

54. Y. Chen et al., "Hierarchical object oriented classification using very high resolution imagery and LIDAR data over urban areas," *Adv. Space Res.* **43**(7), 1101–1110 (2009).

55. J. Zhang et al., "Machine learning-based spectral library for crop classification and status monitoring," *Agronomy* **9**(9), 496 (2019).

56. F. Löw et al., "Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using support vector machines," *ISPRS J. Photogramm. Remote Sens.* **85**, 102–119 (2013).

57. Y. Wang, Q. Qi, and Y. Liu, "Unsupervised segmentation evaluation using area-weighted variance and Jeffries-Matusita distance for remote sensing images," *Remote Sens.* **10**(8), 1193 (2018).
58. R. Sen, S. Goswami, and B. Chakraborty, "Jeffries-Matusita distance as a tool for feature selection," in *Int. Conf. Data Sci. and Eng.*, pp. 15–20, IEEE (2019).
59. L. Bruzzone, F. Roli, and S. B. Serpico, "An extension of the Jeffreys-Matusita distance to multiclass cases for feature selection," *IEEE Trans. Geosci. Remote Sens.* **33**(6), 1318–1321 (1995).

Le'an Qu received his ME degree in geographic information science from Chang'an University, Xi'an, China, in 2009. He is currently working toward his PhD in geography from Nanjing University, Nanjing, China. His research interests include machine learning and remote sensing applications.

Manchun Li received his PhD in cartography from Nanjing University, Nanjing, China, in 1992. He is currently a professor at Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University. His research interests include GIS and remote sensing applications.

Zhenjie Chen received his PhD in cartography and geographic information science from Nanjing University, Nanjing, China, in 2006. He is currently an associate professor at Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University. His research interests include GIS design and remote sensing applications.

Wangbing Liu received his MS degree in agricultural remote sensing and information technology from Zhejiang University, Hangzhou, China, in 2013. He is currently an engineer with Anhui Province Institute of Land Surveying and Planning, Key Laboratory of JiangHuai Arable Land Resources Protection and Eco-restoration, Ministry of Natural Resources. His research interests include remote sensing applications.

Junjun Zhi received his PhD in agricultural remote sensing and information technology from Zhejiang University, Hangzhou, China, in 2014. He is currently an associate professor at the School of Geography and Tourism, Anhui Normal University. His research interests include machine learning and agricultural remote sensing applications.

Lechun Zhang is currently working toward his BSc degree in GIS from Anhui Normal University, Wuhu, China. His research interests include GIS design and applications.