Peer-reviewed version available at *Remote Sens.* **2020**, *12*, 609; doi:10.3390/rs12040609

Article

Mapping Vegetation at Species Level with High-Resolution Multispectral and Lidar Data over Large Spatial Area: A Case Study with Kudzu

Wanwan Liang[1]*, Mongi Abidi[2], Luis Carrasco[3,4], Jack McNelis[5], Liem Tran[6], Yingkui Li[6], Jerome Grant[7]

[1]Center for Geospatial Analytics, North Carolina State University, Raleigh, NC, USA
[2]Department of Electrical Engineering & Computer Science, University of Tennessee, Knoxville, TN, USA
[3]National Institute for Mathematical & Biological Synthesis, Knoxville, TN, USA
[4] Department of Ecology & Evolutionary Biology, University of Tennessee, Knoxville, TN, USA
[5]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[6]Department of Geography, University of Tennessee, Knoxville, TN, USA
[7]Department of Entomology & Plant Pathology, University of Tennessee, Knoxville, TN, USA

* Correspondence: wliang5@ncsu.edu

Abstract: Mapping vegetation species is critical to facilitate related quantitative assessment, and for invasive plants mapping their distribution is important to enhance monitoring and controlling activities. Integrating high resolution multispectral remote sensing (RS) image and lidar (light detection and ranging) point clouds can provide robust features for vegetation mapping. However, using multiple source of high-resolution RS data for vegetation mapping at large spatial scale can be both computationally and sampling intensive. Here we designed a two-step classification workflow to decrease computational cost and sampling effort, and to increase classification accuracy by integrating multispectral and lidar data to derive spectral, textural, and structural features for mapping target vegetation species. We used this workflow to classify kudzu, an aggressive invasive vine, in the entire Knox County (1,362 km^2) of Tennessee, the United States. Object-based image analysis was conducted in the workflow. The first-step classification used 320 kudzu samples and extensive coarsely labeled samples (based on national land cover) to generate an overprediction map of kudzu using random forest (RF). For the second step, 350 samples were randomly extracted from the overpredicted kudzu and labeled manually for the final prediction using RF and support vector machine (SVM). Computationally intensive features were only used for the second-step classification. SVM had constantly better accuracy than RF, and the Producer's Accuracy, User's Accuracy, and Kappa for the SVM model on kudzu was 0.94, 0.96, and 0.90, respectively. SVM predicted 1010 kudzu patches covering 1.29 km^2 in Knox County. We found the sample size of kudzu used for algorithm training impacted the accuracy and number of kudzu predicted. The proposed workflow could also improve sampling efficiency and specificity. Our workflow had much higher accuracy than the traditional method conducted in this research, and could be easily implemented to map kudzu in other regions or other vegetation species.

Keywords: detailed vegetation mapping; kudzu mapping; coarse label; two-step classification; object-based image analysis; lidar point clouds; sampling specificity

1. Introduction

Mapping distribution of target vegetation species can facilitate their quantitative assessment, such as area of spatial coverage, as well as facilitate quantitative modeling for inhabiting species, such as spread of inhabiting invasive species [1]. For invasive plants that can threat biodiversity and cause significant economic loss, mapping their distribution is important to enhance monitoring and controlling activities. Remote sensing (RS) data have been commonly used to map vegetation due to their efficiency and increasing availability [2]. RS images from satellite, such as Landsat and MODIS, have been widely used to map distribution and change dynamics of vegetation [3,4]. However, such data can only be used to classify vegetation with high-level classes due to the moderate or coarse resolution [2-4]. Hyperspectral RS data, which can provide a relatively complete spectral profile of vegetation, can be used to map vegetation at species-level [6,7]. However, hyperspectral data are not widely available thus are usually used to map vegetation for small regions [6-8]. Consequently, multispectral RS images with high spatial resolution, which are more widely available than hyperspectral data, is increasingly used for mapping vegetation at species level over large spatial area [9-12].

Conducting object-based image analysis (OBIA) with high-resolution RS data has become a common practice for vegetation mapping [11-14]. OBIA takes a patch of vegetation as a unit rather than a pixel, thus is less impacted by noise caused by within-class variation caused by high spatial resolution than pixel-based image analysis (PBIA) [11]. Additionally, different from PBIA mainly using features from the spectral domain, OBIA can use features from both spectral and spatial domains. Thus, OBIA can integrate spectral, geometric, and textural features leading to good accuracy for vegetation mapping [11-15]. With increasing availability, lidar (light detection and ranging) data have been integrated with multispectral RS data and OBIA method to provide structural features of vegetation to improve classification accuracy [14-16]. However, as a result of high-resolution and calculation of various types of features [15], integrating multispectral image and lidar data for detailed vegetation mapping at large spatial scale is computationally intensive. Thus, here we designed a workflow to decease computational cost for vegetation mapping at species level over regional scale, but to increase classification accuracy by integrating high-resolution multispectral images and lidar point clouds data.

Vegetation mapping over large spatial area often requires large sampling effort for all vegetation classes, even when the objective is to map certain target vegetation species [17-19]. For examples, Dorigo et al. [17] integrated multiple source of RS data to map an invasive plant by collecting samples for all vegetation classes, and Nguyen et al. [18] used RS data and OBIA to map multiple plant species by collecting samples for all primary vegetation classes. Using coarsely or imperfectly labeled samples extracted from open source platform [20] or land cover maps [21] is a valuable method to facilitate classification when accurately labeled samples are limited [20-23]. Langford et al. [21] used a coarse land cover map together with K-means clustering to generate training samples for vegetation mapping, and Maggiori et al. [20] first used imperfect data to train neural networks and then refined the model with small amount of accurately labeled data to detect

urban buildings. In our proposed workflow, we designed an innovative use of coarsely labeled samples for mapping target vegetation species to decrease sampling effort for non-target vegetation classes.

We used this workflow to map kudzu, *Pueraria montana*, a serious invasive plant in the southeastern United States (U.S.). The spread of kudzu has led to alteration of forest canopy structure, species biodiversity loss, and economic loss due to control fees and loss of forest production [24,25]. Additionally, kudzu also puts threat on soybean production as an alternative host to a soybean pest, kudzu bug, *Megacopta cribraria* [1]. Kudzu is also found as an invasive species in Australia, Canada, New Zealand, South Africa, and Central and South America countries [26]. Mapping the distribution of kudzu is critical for its management. Existing research on kudzu mapping used hyperspectral data for small spatial area, whereas the accuracy is insufficient [27,28].

Multiple studies evaluated the impact of training samples on the classification accuracy [29-31]. Research by Heydari and Mountrakis [30] and Millard and Richardson [31] suggested that larger sample size could result in higher classification accuracy, meanwhile Millard and Richardson [31] also found that the proportion of each class in the training samples impacted its predicted proportion over the landscape. Here, to determine the sample size of target vegetation species, we also analyzed the impact of sample size of target vegetation used for classifier training on the classification results.

Nowadays deep learning models, especially convolutional neural networks (CNNs), have been increasingly applied for image classification with RS data due to their ability of self-extracting features through convolutional layers [20,21,32]. However, CNNs is computationally redundant and require large samples to train the model, which can be a serious challenge for mapping vegetation at species level with high resolution data over large area. Therefore, in our workflow we conducted OBIA with random forest (RF) and support vector machine (SVM).

The workflow includes a two-step classification: 1) the first-step classification uses accurately labeled kudzu samples and extensive coarsely labeled vegetation samples to construct an overprediction model of kudzu, and 2) the second-step classification uses 350 randomly extracted and accurately labeled samples from the overpredicted kudzu map to refine the classification. The designed workflow decreases computational cost by only using computational-intensive features, including gray-level co-occurrence matrix (GLCM) derived textural features and lidar point clouds derived features, at the second step. The sampling effort is decreased for non-target species by using extensive coarsely labeled samples at the first step and accurately labeling a relatively small set of samples at the second step. By integrating high-resolution multispectral and lidar point clouds data and conducting OBIA, this workflow can also improve detection accuracy on target vegetation species. The proposed workflow can be easily implemented to map kudzu in other regions and other vegetation species.

2. Materials and Methods

2.1 Study Area and Target Vegetation Species

The study area includes the entire Knox County (about 1,362 km^2, Latitude: ~ 35.79° - 36.19° N, Longitude: ~ 83.65°- 84.26° W) in Tennessee (TN), U.S. According to National Land Cover Database (NLCD) 2016, 40.64%, 34.12%, and 21.5% of the study area are developed, forest, and herbaceous

areas, respectively. Kudzu is a vine plant, and its vines turn gray after the first frost. New leaves of kudzu sprout late in spring when compared to many other vegetation species. This phenology makes kudzu separable from surrounding vegetation in early spring. Therefore, the designed workflow first conducted segmentation on the image taken in early spring.

2.2 Data Details

The 4-band images (RGB and infrared) from National Agriculture Imagery Program (NAIP) were used. NAIP acquires aerial imagery for each state every two years in the continental U.S., and the timing varies among different years. NAIP images taken in June 2012, May 2014, April 2016, and October 2018 were used in this research. The infrared band of four NAIP images were used to derive textural features. Entropy, range, standard deviation (SD), and focal SD of infrared bands for vegetation objects were used as textural features as well as seven GLCM derived features (Table 1). Digital elevation model (DEM) and lidar point clouds from 3D elevation program of U.S. were used to get topographical features and canopy structural features, respectively. NLCD 2016 was used to derive coarse labels of vegetation samples. Table 1 provides a summary of data and features used in this research.

Table 1 Source and description of data and details on features used for kudzu classification

Data source and description	Feature category	Extracted features	Number of features
National agricultural imagery program, RGB and infrared spectral bands, 1-m resolution	Spectral bands	Mean value of 4-bands	16
	Vegetation index	Mean, range, and standard deviation (SD) of normalized difference vegetation index	12
	Textural features on infrared band	Entropy, SD, and focal SD	40
		Gray-level co-occurrence matrix (GLCM) mean	
		GLCM variance	
		GLCM homogeneity	
		GLCM dissimilarity	
		GLCM contrast	
		GLCM entropy	
		GLCM second moment	
3D elevation program, Elevation, 2.5-m resolution	Topographic features	Mean and SD of elevation	4
		Mean focal range and SD of elevation	

3D elevation program, lidar point clouds, 3-returns/m²	Canopy structural features	Coefficient of variation (CV), mean, variance, SD, intensity, and mode of Z-values of first return points Mode of all return points Density of all return points	8
Segments	Geometric feature	Area of each object	1
National land cover database, 30-m resolution	Coarse land cover label	Majority land cover type of all pixels in each segment	NA

2.3 Designed Workflow

The proposed workflow for target vegetation mapping was summarized in Figure 1.

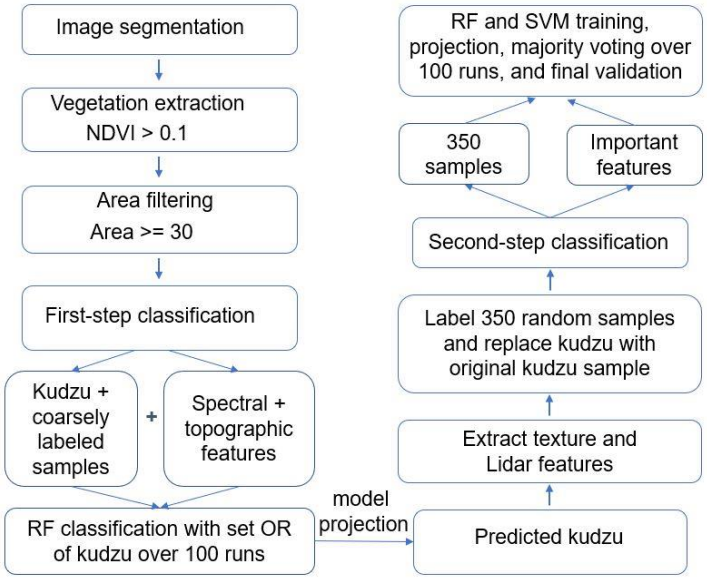

Figure 1. Workflow of proposed method on classifying kudzu with low sampling effort and computational cost

2.3.1 Image Segmentation and Vegetation Extraction

Segmentation was conducted on the 4-band NAIP imagery taken in April 2016, as kudzu was distinctive from surrounding vegetations due to its gray color at this time of year (Figure 2). Image segmentation was conducted in ArcGIS using the Mean Shift Segmentation function [33]. Mean shift is a powerful local homogenization technique that replaces pixel value with the mean of pixels in a given radius (r) of neighborhood with values varying within a given range d. Users define the r and d parameters in ArcGIS by setting the spatial and spectral detail values between 1 and 20. After comparing multiple sets of parameters, the values of spatial and spectral detail were set to 10 and 15, respectively. A scene with multiple kudzu patches from NAIP 2016 and the corresponded segmented image were shown in Figure 2.

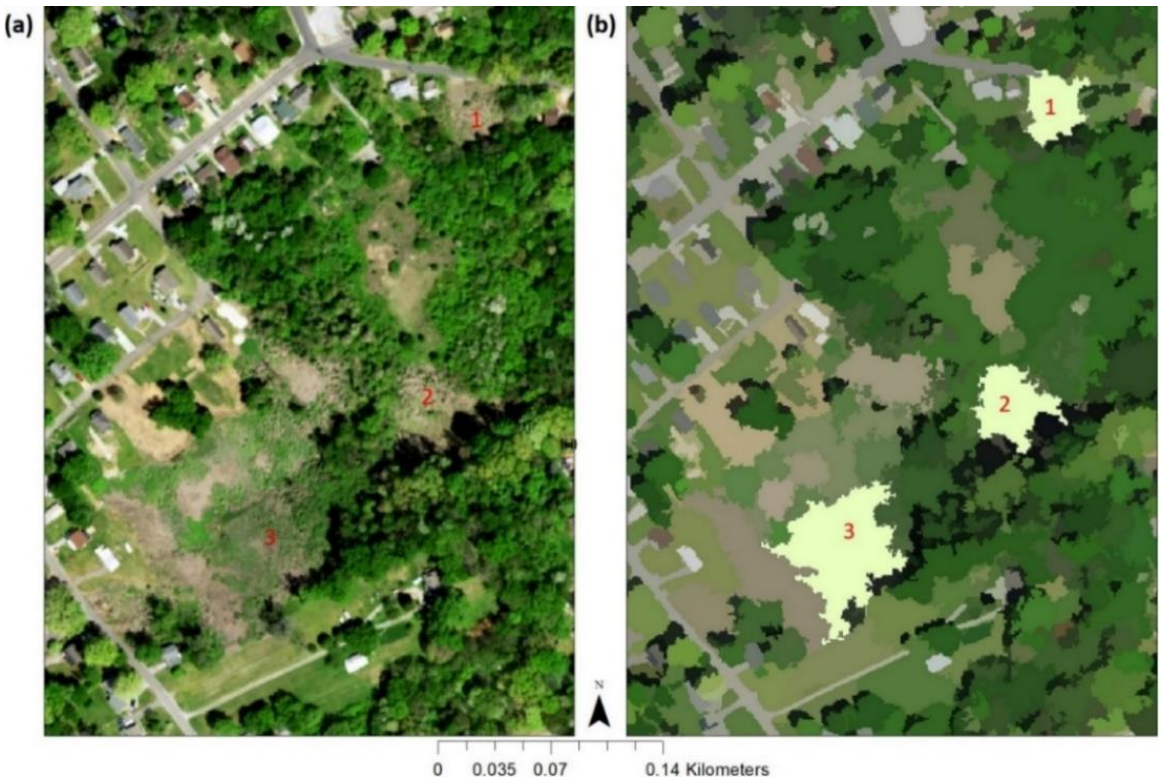

Figure 2. (a) a scene of NAIP 2016 taken in April 2016 with kudzu patches. Kudzu patch 1 and 2 representing relatively pure kudzu patches, and 3 representing kudzu patch mixing up with other vegetation. **(b)** segments of (a) with 1, 2, and 3 corresponded to the three kudzu patches in **(a)**.

NDVI derived from NAIP 2012 was used to remove non-vegetation, as the imagery was taken in summer (June 2012) when vegetation was most distinctive from non-vegetation. Objects with mean NDVI lower than 0.1 were classified as non-vegetation and removed. Among the remaining vegetation, small objects with area less than 30m^2 were further excluded to decrease potential noises. The total number of vegetation objects in the study area after each preprocessing procedure was summarized in Table 2 in Results section.

2.3.2 Two-Step Classification

We collected 400 kudzu samples from the entire study area either by field sampling or visual interpretation using Google Earth, and the corresponded vegetation segments were assigned as kudzu samples. Among the 400 samples, 20% (80) were set aside for model final validation and 80% (320) were used for model training and testing (Figure 3. (a)). GLCM derived textural features and lidar point clouds derived features were only included for the second step to decrease the computational cost.

First-Step Classification with Coarsely-labeled Samples

Accurately labeled kudzu samples together with coarsely labeled vegetation samples were used to classify kudzu for the first step (Figure 1). NLCD 2016 was used to coarsely label samples that were randomly extracted from all vegetation objects by assigning the majority land cover class within the coverage of each vegetation object. RF classifier was used for the first-step classification for its high accuracy and efficiency on large dataset. All features (Table 1), except lidar point clouds

derived and GLCM derived features, were used to construct the RF model. We randomly extracted 80% (256) kudzu from the 320 samples and a given number of coarsely labeled vegetation samples to train a RF model, and the omission rate (OR) of kudzu for each RF model was evaluated using the remaining 20% (64) samples. The RF model was run 100 times with each time a different set of kudzu training and testing samples and coarsely-labeled vegetation samples were used.

The OR of kudzu is affected by the number of coarsely labeled vegetation samples used for model construction. Thus, we first determined the number of coarsely labeled samples included in each RF model to achieve target OR of kudzu over 100 runs. We selected three testing OR, 0.01, 0.03 and 0.05, to assess which OR at the first step could produce the best final classification accuracy of kudzu. The corresponded numbers of coarsely labeled samples for OR of 0.01, 0.03, and 0.05 were 2500, 10000, and 20500, respectively. However, these numbers should vary based on the type and size of target vegetation species, so users should determine the number of coarsely-labeled samples based on their samples and expected OR. For each run, the fitted RF was used to predict vegetation class for all vegetation objects, and the majority of predicted class over 100 runs was assigned as the predicted vegetation type.

Second-Step Classification with Accurately Labeled Samples

For the second step, 350 samples were randomly extracted from the predicted kudzu by the first step and were accurately labeled. We used randomly extracted samples to make the number of samples for each vegetation class roughly representative of actual class proportions in the remaining vegetation objects [29-31,34]. Among the 350 random samples, kudzu samples were replaced with the same number of samples from the original 320 training sets. The classification models were then constructed on the new 350 training data and used to re-predict vegetation objects that were classified as kudzu by the first step (Figure 2.). RF and support vector machine (SVM) were used in the second-step classification due to their constantly reported good performance [12,35,36]. Detailed explanations of the algorithms can be found in [37,38]. Each model was run 100 times, and for each run a different set of kudzu samples randomly extracted from the 320 kudzu samples was used. The majority of prediction over 100 runs was taken as the final prediction for all remaining objects. An independent set of 80 non-kudzu objects were labeled and used as validation data for final model evaluation.

2.4 Comparison of Proposed Method with Traditional Method

To compare the proposed workflow with the traditional method, we also used the traditional method, which uses accurately labeled samples for all classes, to classify kudzu. The 320 kudzu samples were compiled with 800 randomly extracted and accurately labeled samples to construct the training set of the traditional method. The randomly extract samples included 662 forestry objects, 121 herbaceous objects (including 84 grass and 37 other herbaceous vegetation), 16 urban objects, and 1 kudzu object. As the target class was vegetation and the misclassification arose from confusion of kudzu with other vegetation classes, the small proportion of urban objects in the sample would not be a concern.

We only included lidar point clouds derived structural features and GLCM derived textural features in the second step of classification to decrease the computational cost. To make the traditional method comparable with the proposed method, the RF was also conducted 100 times

with each run using 80% and 20% kudzu sample as training and testing dataset, respectively. Similarly, the majority of prediction over 100 runs was assigned as the prediction for all objects, and only predicted kudzu objects were included in the second step. Given the apparent low accuracy of this method (details in Section 3.1 and 3.2), we also used a similar sub-sampling method for the traditional approach as follow: we randomly extracted and labeled 350 samples from the predicted kudzu by the first step and replaced kudzu with the same number of samples from the original 320 training set, and used these samples for the second-step classification.

2.5 Feature Selection, Accuracy Assessment, and Impact of Sample Size of Kudzu

Feature selection was conducted for the second-step classification. For each model, we selected the minimum set of features that generated the same classification accuracy for kudzu objects as all the features for the RF model. The same minimum set of features were then used to train both RF and SVM models. Producer's Accuracy (PA, 1-OR), User's Accuracy (UA, 1- commission rate (CR)), and Kappa coefficient [39] were used to evaluate classification accuracy. As kudzu is the target vegetation species in this research, accuracy assessment was only conducted on the kudzu class. The 80 kudzu and non-kudzu validation samples were used to calculate the PA, UA, and Kappa. To evaluated the impact of target vegetation sample size on the classification results, we varied the number of kudzu samples used for each run of model training for the proposed method with 1% OR at the first step and for the traditional method with sub-samplings.

3. Results

3.1 First-Step Classification

The first-step classification dramatically decreased the total number of objects from millions to thousands (Table 2), consequently, decreased the computational cost for extracting GLCM-textural and lidar point cloud features. For the proposed method, the number of classified kudzu objects was negatively associated with the OR of kudzu class (Table 2). The OR on validation and testing data had the same values for all procedures and methods (Table 2). The OR of the traditional method was 1%, however, the number of predicted kudzu objects was higher than the proposed method with 1% OR, suggesting higher CR of the traditional method (Table 2).

Table 2 Number of objects after each procedure and the corresponded omission rate (OR) of kudzu objects based on testing and validation dataset

Number of objects		Initial objects	Preprocessing		Number of predicted kudzu			
			Vegetation extraction	Area filtering	Proposed method			Traditional method
		9,454,240	6,911,589	3,417,188	19,548	5,306	2,815	30,268
OR	Testing data	0%	0%	0%	1%	3%	5%	1%
	Validation data	0%	0%	0%	1%	3%	5%	1%

3.2 Second-Step Classification

Table 3 listed the number of samples for each class among the 350 randomly extracted and accurately labeled samples. It is important to notice the kudzu samples in the randomly extracted samples were replaced with the same number of kudzu in the training samples instead of all kudzu training samples to avoid overprediction. The sample number of each class suggested that the major misclassification of kudzu by the first step came from the confusion of kudzu with forest and herbaceous plants (Table 3).

Table 3 Number of accurately labeled samples used for each run of model training for the second-step classification

Class	Proposed method (first-step omission rate)			Traditional method
	1%	3%	5%	Second-step sampling
Bare ground	6	20	16	2
Forest	127	73	97	96
Grass	91	43	38	155
Other herbaceous vegetation	71	98	79	64
Kudzu	33	97	106	13
Urban objects	22	19	14	20
Total number	350	350	350	350

SVM constantly produced better accuracy than RF for the proposed method (Table 4). With 1% OR at the first step, the second-step classification had underprediction of kudzu (resulting in low PA and high UA) as a result of insufficient kudzu samples included in each run (Table 4). Both RF and SVM had higher PA but lower UA when the OR was set to 3% than 5% at the first step, and both models had highest Kappa (0.88 for RF and 0.90 for SVM, Table 4) with 3% OR at the first step. Among all models, SVM with 3% OR at the first step had highest classification accuracy (PA=0.94, UA=0.96), and 1010 kudzu objects were predicted over the entire study area (Figure 3. (b), Supplementary 1).

Table 4 classification accuracy of kudzu on validation dataset by random forest (RF) and support vector machine (SVM) models

Model	Proposed method (first-step omission rate)									Traditional method					
	1% (33*) (612**, 609***)			3% (97*) (1084**, 1010***)			5% (106*) (853**, 777***)			One-step sampling (50*) (12826**, 8791***)			Two-step sampling (50*) (1726**, 1129***)		
	PA	UA	K	PA	UA	K	PA	UA	K	PA	UA	K	PA	UA	K
RF	0.66	1.00	0.66	0.94	0.94	0.88	0.86	0.95	0.81	0.93	0.63	0.38	0.83	0.94	0.80
SVM	0.70	1.00	0.70	**0.94**	**0.96**	**0.90**	0.86	0.97	0.84	0.96	0.72	0.59	0.78	0.95	0.74

PA, UA, and K are short for Producer's Accuracy, User's Accuracy, and Kappa coefficient, respectively. * indicate the number of kudzu sample included in each run. **and *** indicated number of kudzu object predicted by RF and SVM models, respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2019 doi:10.20944/preprints201912.0418.v1

Peer-reviewed version available at Remote Sens. 2020, 12, 609; doi:10.3390/rs12040609

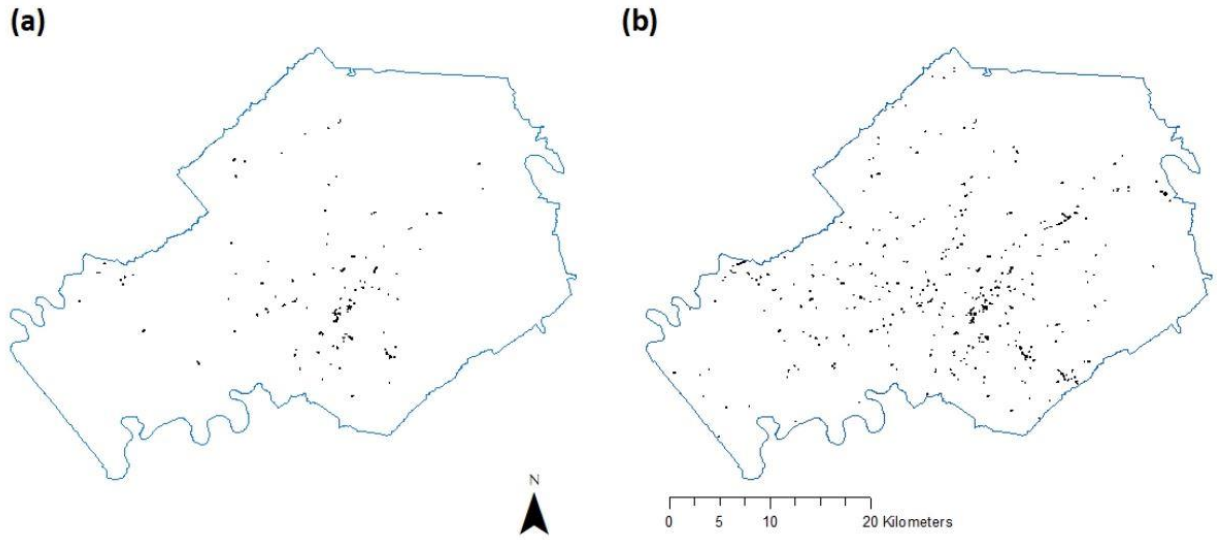

Figure 3. **(a)** 320 kudzu training samples, and **(b)** 1010 predicted kudzu objects in Knox County by support vector machine using the proposed workflow with omission rate set to 1% at the first-step classification.

Given the apparent overprediction of the traditional method and the fact that more kudzu samples included in each run lead to higher overprediction (details in section 3.3), only 50 randomly extracted kudzu samples from the 320 samples were included in each run. The traditional method with one step sampling had overprediction of kudzu, leading to high PA but low UA (Table 4). The overprediction was a result of misclassification of grass and other herbaceous plants into kudzu (Figure 4). Using a sub-sampling from the predicted kudzu by the first step dramatically improved the classification accuracy (Table 4). However, the traditional methods had constantly lower accuracy compared to the proposed methods (Table 4).

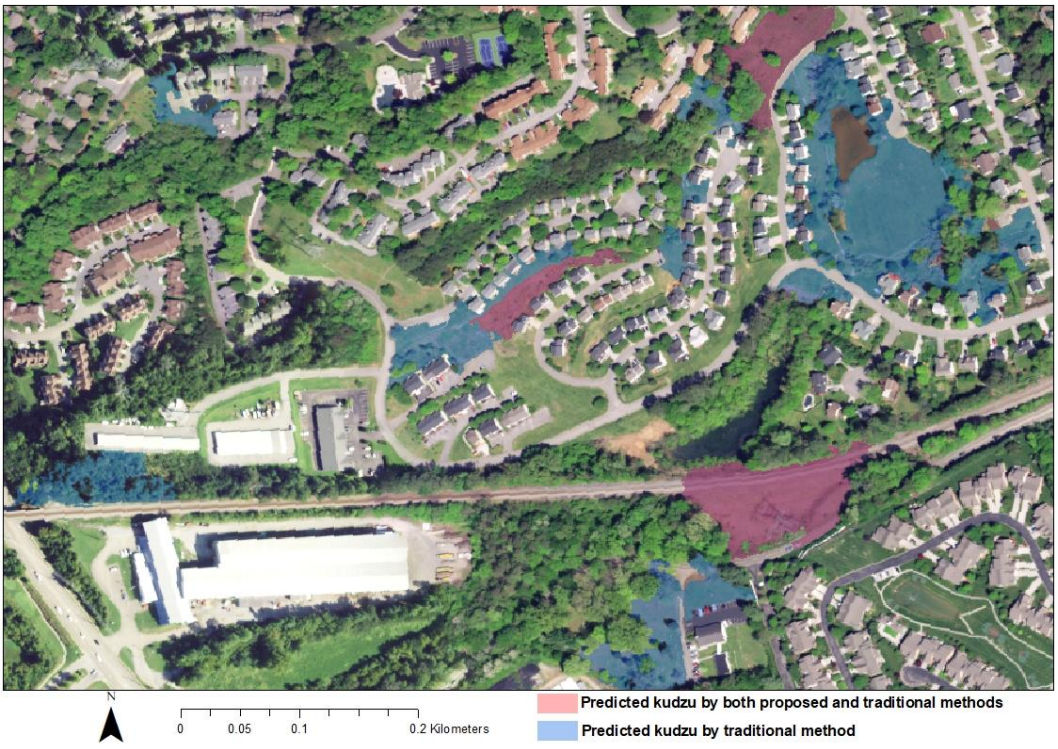

Figure 4. A scene of detected kudzu in landscape by the best support vector machine (SVM) model using proposed method and the predicted kudzu by the traditional method with one step sampling using SVM. Both methods detected all three kudzu patches in this landscape (shown in red), whereas the traditional method misclassified more herbaceous vegetation into kudzu (shown in blue).

3.3 Feature Importance

The feature importance was determined based on ranking of SVM and RF over 100 runs with the 3% OR at the first step as it led to the best final detection accuracy. The 15 most important features ranked by both SVM and RF included 5 lidar point clouds derived features, 5 infrared band derived features (mostly GLCM textural features), and 5 other features (mostly NDVI features), suggesting lidar, textural, and spectral features were all important for the detection of kudzu (Figure 5).

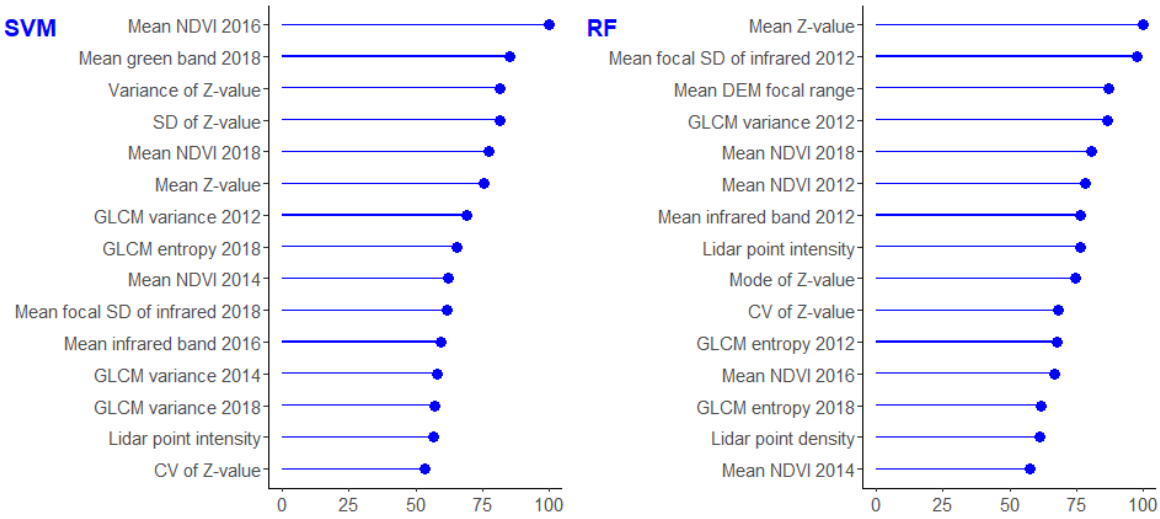

Figure 5. The 15 most important features ranked by support vector machine (SVM) and random forest (RF) over 100 runs for the second-step classification. NDVI, SD, GLCM, and CV are short for normalized difference vegetation index, standard deviation, gray-level co-occurrence matrix, and coefficient of variance, respectively.

3.4 Impact of Kudzu Sample Size on Classification Accuracy

Given the same set of training data for other vegetation classes, the number of kudzu samples included in each run was positively associated with the PA and number of predicted kudzu objects, but negatively associated with the UA (Table 5). Interestingly, we found a quadratic relation between the number of kudzu sample included in each run and the number of kudzu predicted by both SVM and RF (R^2 =1 for all models, Figure 6). Insufficient kudzu samples included in each run result in underprediction while too many samples led to overprediction (Table 5). Using 100 kudzu samples in each run generated the best accuracy (PA= 0.93, UA= 0.82) for the proposed method with 1% OR at the first step, whereas the traditional method failed to have good accuracy in all cases (Table 5).

Table 5 Mean testing accuracy on classifying kudzu objects by four machine learning models over 100 runs

Method	Model	Metrics	50	100	150	200
Proposed method with 1% OR at first-step classification	RF	PA	0.79	0.96	0.96	0.96
		UA	0.72	0.50	0.42	0.20
	SVM	PA	0.84	0.93	0.96	0.96
		UA	0.84	0.82	0.60	0.56
Traditional method with two-step sampling	RF	PA	0.83	0.94	0.96	0.96
		UA	0.50	0.40	0.40	0.24
	SVM	PA	0.78	0.91	0.96	0.96
		UA	0.65	0.34	0.40	0.36

OR, RF, SVM, PA, and UA, are short for omission rate, random forest, support vector machine, Producer's Accuracy, and User's Accuracy, respectively.

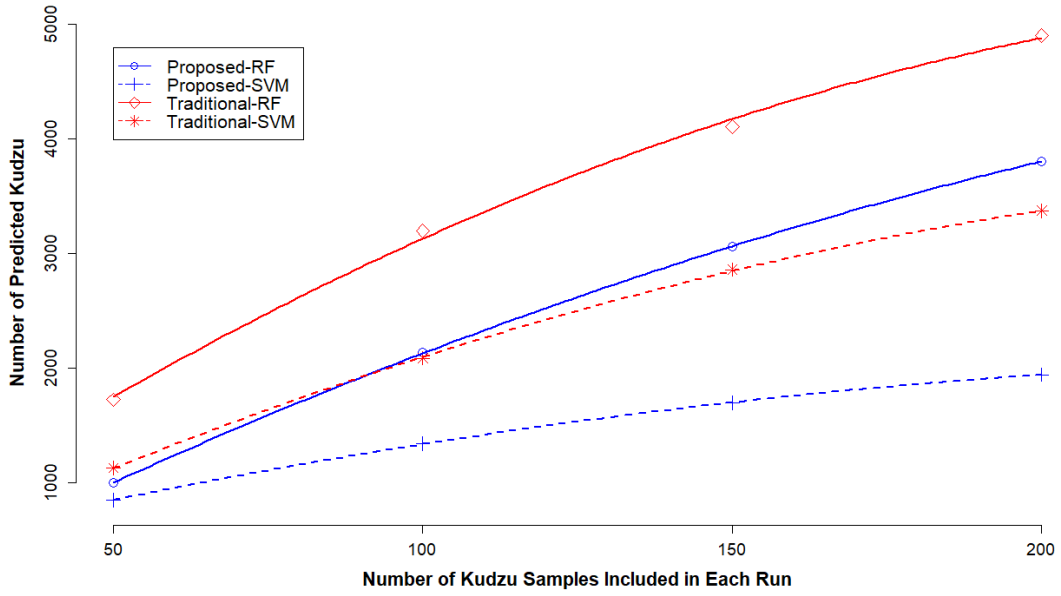

Figure 6. Quadratic relations between the number of kudzu samples included in each run and number of predicted kudzu objects by random forest (RF) and support vector machine (SVM) models for both proposed workflow and traditional method. Points patches represent the observed number of predicted kudzu and lines represent fitted quadratic models.

4. Discussion

4.1 Fusion of Multiple Remote Sensing Data for Vegetation Mapping

The need for detailed vegetation mapping with reliable accuracy is increasing, and the increasing availabilities of high-resolution RS data, including multispectral, hyperspectral, and lidar data greatly facilitate this task. Integrating RS data from multiple platform could provide a more robust set of features for better detection of target vegetation species than using a single source of RS data [14-16], and has become an increasingly used approach for vegetation mapping

at species level [4,14,40]. Using OBIA with high-resolution RS data can mitigate the potential noise caused by within-class variation associated with high-resolution [11], as well as making it applicable to use features from both spectral and spatial domains. Compared to the classification accuracy of kudzu achieved by Cheng et al. [27] using pixel-based hyperspectral RS data for small study area (PA= 0.73, UA= 0.83), our classification on kudzu over large spatial area produced higher accuracy (PA= 0.94, UA= 0.96). The improved performance of our method may suggest integrating high-resolution multispectral and lidar data to derive spectral, textural, and structural features could provide more robust predictors for vegetation mapping than a full spectral profile for some vegetation species.

4.2 Decrease of Computational Cost and Sampling Effort

Integrating multiple source of high-resolution RS data for detailed vegetation mapping at large spatial scale can be computationally intensive, especially for 3-D lidar point clouds data. For example, to map tree canopy in New York City, 307 GB lidar data needed to be processed to cover the whole study area [15], and in our case 255 GB lidar data needed to be processed. Reducing computational cost thus can greatly facilitate vegetation mapping with high resolution data, especially when the study area is regional or larger. Our designed workflow first used extensive coarsely labeled samples and accurately labeled target vegetation samples to train an overprediction map of target vegetation, and dramatically decreased vegetation objects from millions to thousands for a more detailed and accurate second-step classification. Computationally intensive features, such as GLCM derived and lidar point clouds derived features, only need to be extracted for objects that are not distinctive from the target vegetation class for a second-step classification.

Combining coarsely labeled samples with accurately labeled samples have been implemented to classify extensive animal images into detailed categories [22,23] as well as to detect urban buildings [20]. Our research suggests coarsely labeled data can also facilitate the classification of vegetation at species level, and a 30-m resolution land cover product can be a good source to coarsely label vegetation samples. Using coarsely labeled samples together with accurately labeled samples of target vegetation helps to discriminate objects with distinctive features with the target vegetation species. For vegetation objects that can not be easily distinguished from the target class, a relatively small set of accurately labeled samples, 350 in our case, extracted from the predictions by the first step can greatly refine and improve the classification accuracy.

4.3 Sub-sampling Improves Sample Specificity for Target Vegetation Class

We found even for the traditional method, a sub-sampling among the predicted objects from the first step greatly improved the final classification accuracy on the target vegetation class. The extreme overprediction (i.e., low UA) of kudzu using only one-step sampling should be a result that the samples failed to include sufficient vegetation samples that only occupied a small proportion in landscape but had similar features with kudzu. As a result, both RF and SVM classified these vegetation objects having similar features with target vegetation as kudzu. A sub-sampling among the predicted objects from the first step thus helps optimize the sampling effort by selecting samples that are similar with target vegetation class. For vegetation mapping over

large heterogenous landscape, it could be difficult to find sufficient vegetation samples that have similar features with target vegetation. The two-step sampling method used in this research, therefore, could be conducted to improve the sampling specificity for target vegetation mapping.

4.4 Impact of Sample Size of Target Vegetation

Using balanced training data with equal sample size for each class or unbalanced data with sample size of classes proportional to their area in real landscape is an unsettled concern, and both strategies have been commonly used for image classification [29-31,41]. Millard and Richardson [31] found that the predicted proportion of each class by RF was close to its proportion in real landscape when the training sample size of each class resemble its real proportion in the landscape. They also found that higher proportion of each class in the training sample led to more prediction of that class by RF [31]. Our research confirmed their findings, as increasing kudzu sample size while keeping other classes unchanged increased the number of predicted kudzu. Meanwhile we also found SVM shows the same pattern, and both models showed a quadratic relation between the number of kudzu sample included in each run of model and the number of kudzu predicted.

Researchers found that classification tree models showed best accuracy with unbalanced samples representing the proportion of each class in real landscape [29,31]. Our research confirmed this finding as the best RF occurred when the number of kudzu samples in each run was determined roughly based on their proportion in all the vegetation objects. Meanwhile, SVM also showed the same pattern. However, as suggested by the underprediction of kudzu by both RF and SVM models with 1% OR at the first step, when the target vegetation class is among the minority class, using the proportion of each class rule may result in insufficient samples leading to underprediction (i.e., low PA, Table 3). On the other hand, oversampling of the minority class can lead to overprediction (i.e., low UA, Table 4) [31,41]. Consequently, the PA for the minority class could show a positive association with the number of samples included for algorithm training, whereas UA showed a negative association. Researchers suggested to include a minimum number of training samples, like 50, for minority class with unbalanced samples [20]. Here we also suggest to determine the optimal sampling size of minority class used for algorithm training by balancing desired accuracy between PA and UA [42].

4.5 Use and Limitation of the Proposed Workflow

The proposed workflow could decrease computational cost as well as sampling effort for non-target classes for mapping target vegetation at large spatial scales. However, caution needs to be paid on the selection of OR for the first step. Higher OR at the first step lead to less computation for the second step but also lead to high OR of the final classification model. In our case, 3% OR at the first step generated the best classification accuracy in terms of UA and Kappa. Although a 1% OR for the first step led to a highly overprediction of kudzu, using an appropriate kudzu sample size for training the classification algorithms at the second step still produced a good final accuracy. This suggests that selecting an appropriate sample size of target vegetation at the second step can overcome the overprediction by the first step. We therefore suggest a low OR, such as 3% and 1%, at the first step for higher final classification accuracy. Additionally, we expect our workflow work

effectively for vegetation species that are separable from surroundings at a certain time of year to be segmented for OBIA.

5. Conclusions

Integrating high-resolution multispectral image and lidar point clouds data and applying OBIA can provide robust set of features for vegetation mapping at species level. The designed workflow using two step classification can dramatically reduce computational cost associated with deriving computationally intensive features from high-resolution RS data over large spatial area. Using accurately labeled samples of target vegetation species and extensive coarsely labeled vegetation samples to first train an overprediction model can be an efficient method for mapping vegetation over large spatial area to decrease sampling effort for non-target vegetation. A land cover map with 30 m resolution can be a good source to coarsely label vegetation samples. We suggest a low OR of target vegetation species at the first step to achieve good final accuracy. The workflow can also improve sampling efficiency and specificity, as a sub-sampling among the predicted objects by the first step optimizes the sampling effort by selecting samples that are not distinctive from target vegetation species. Caution need to be paid on the sample size of target vegetation used to train the classification algorithm. Insufficient number of samples lead to underprediction, whereas too many samples lead to overprediction of the target vegetation classes. We expect the proposed workflow work effectively for vegetation classes that are separable from surroundings to be segmented for OBIA.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Shapefile S1: 1010 predicted kudzu patch by SVM

Acknowledgments: We thank Ming Shen, Dr. Monica Papeş, Dr. Scott Stewart, and Dr. Gregory Wiggins at the University of Tennessee for valuable discussions on this research.

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

References

1. Liang, W.; Tran, L.; Wiggins, G.J.; Grant, J.F.; Stewart, S.D.; Washington-Allen, R. Determining spread rate of kudzu bug (Hemiptera: Plataspidae) and its associations with environmental factors in a heterogeneous landscape. *Environ. Entomol.* **2019**, *48*, 309-317.
2. Xie, Y.; Sha, Z.; Yu, M. Remote sensing imagery in vegetation mapping: a review. *J. Plant Ecol.* **2008**, *1*, 9–23.
3. Wardlow, B.; Egbert, S.; Kastens, J. Analysis of time-series MODIS 250 m vegetation index data for crop classification in the U.S. Central Great Plains. *Remote Sens. Environ.* **2007**, *108*, 290–310.
4. Gilmore, M.S.; Wilson, E.H.; Barrett, N.; Civco, D.L.; Prisloe, S.; Hurd, J.D.; Chadwick, C. Integrating multi-temporal spectral and structural information to map wetland vegetation in a lower Connecticut River tidal marsh. *Remote Sens. Environ.* **2008**, *112*, 4048–4060.
5. Yang, J.; Weisberg, P.J.; Bristow, N.A. Landsat remote sensing approaches for monitoring long-term tree cover dynamics in semi-arid woodlands: Comparison of vegetation indices and spectral mixture analysis. *Remote Sens. Environ.* **2012**, *119*, 62–71.
6. Zhong, Y.; Wang, X.; Xu, Y.; Wang, S.; Jia, T.; Hu, X.; Zhao, J.; Wei, L.; Zhang, L. Mini-UAV-borne hyperspectral remote sensing: from observation and processing to applications. IEEE Geosci. *Remote Sens. Mag.* **2018**, *6*, 46–62.

7. Zhang, L.; Zhang, L.; Tao, D.; Huang, X. On combining multiple features for hyperspectral remote sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 879–893.
8. Burai, P.; Deák, B.; Valkó, O.; Tomor, T. Classification of herbaceous vegetation using airborne hyperspectral imagery. *Remote Sens.* **2015**, *7*, 2046–2066.
9. Alvarez-Taboada, F.; Paredes, C.; Julián-Pelaz, J. Mapping of the invasive species hakea sericea using unmanned aerial cehicle (UAV) and WorldView-2 imagery and an object-oriented approach. *Remote Sens.* **2017**, *9*, 913.
10. Ahmed, O.S.; Shemrock, A.; Chabot, D.; Dillon, C.; Williams, G.; Wasson, R.; Franklin, S.E. Hierarchical land cover and vegetation classification using multispectral data acquired from an unmanned aerial vehicle. *Int. J. Remote Sens.* **2017**, *38*, 2037–2052.
11. Yu, Q.; Gong, P.; Clinton, N.; Biging, G.; Kelly, M.; Schirokauer, D. Object-based detailed vegetation classification with airborne high spatial resolution remote sensing imagery. *Photogramm. Eng. Remote Sens.* **2006**, *72*, 799–811.
12. Fu, B.; Wang, Y.; Campbell, A.; Li, Y.; Zhang, B.; Yin, S.; Xing, Z.; Jin, X. Comparison of object-based and pixel-based Random Forest algorithm for wetland vegetation mapping using high spatial resolution GF-1 and SAR data. *Ecol. Indic.* **2017**, *73*, 105–117.
13. Erker, T.; Wang, L.; Lorentz, L.; Stoltman, A.; Townsend, P.A. A statewide urban tree canopy mapping method. *Remote Sens. Environ.* **2019**, *229*, 148–158.
14. Ke, Y.; Quackenbush, L.J.; Im, J. Synergistic use of QuickBird multispectral imagery and LIDAR data for object-based forest species classification. *Remote Sens. Environ.* **2010**, *114*, 1141–1154.
15. MacFaden, S.W.; O'Neil-Dunne, J.P.M.; Royar, A.R.; Lu, J.W.T.; Rundle, A.G. High-resolution tree canopy mapping for New York City using LIDAR and object-based image analysis. *J. Appl. Remote Sens.* **2012**, *6*, 063567.
16. Suárez, J.C.; Ontiveros, C.; Smith, S.; Snape, S. Use of airborne LiDAR and aerial photography in the estimation of individual tree heights in forestry. *Comput. Geosci.* **2005**, *31*, 253–262.
17. Dorigo, W.; Lucieer, A.; Podobnikar, T.; Čarni, A. Mapping invasive *Fallopia japonica* by combined spectral, spatial, and temporal analysis of digital orthophotos. *Int. J. Appl. Earth Obs. Geoinformation* **2012**, *19*, 185–195.
18. Nguyen, U.; Glenn, E.P.; Dang, T.D.; Pham, L.T.H. Mapping vegetation types in semi-arid riparian regions using random forest and object-based image approach: A case study of the Colorado River Ecosystem, Grand Canyon, Arizona. *Ecol. Inform.* **2019**, *50*, 43–50.
19. Narumalani, S.; Mishra, D.R.; Wilson, R.; Reece, P.; Kohler, A. Detecting and mapping four invasive species along the floodplain of North Platte River, Nebraska. *Weed Technol.* **2009**, *23*, 99-107.
20. Maggiori, E.; Tarabalka, Y.; Charpiat, G; Alliez, P. Convolutional neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645-657.
21. Langford, Z.; Kumar, J.; Hoffman, F.; Breen, A.; Iversen, C. Arctic vegetation mapping using unsupervised training datasets and convolutional neural networks. *Remote Sens.* **2019**, *11*, 69.
22. Ristin, M.; Gall, J.; Guillaumin, M.; Van Gool, L. From categories to subcategories: large-scale image classification with partial class label refinement. In *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.* **2015**, 231-239.

23. Lei, J.; Guo, Z.; Wang, Y. Weakly supervised image classification with coarse and fine labels. In *Proc. 14th Conf. Comput. Robot Vis.* **2017**, 240–247.
24. Sun, J.-H.; Liu, Z.-D.; Britton, K.O.; Cai, P.; Orr, D.; Hough-Goldstein, J. Survey of phytophagous insects and foliar pathogens in China for a biocontrol perspective on kudzu, *Pueraria montana* var. lobata (Willd.) Maesen and S. Almeida (Fabaceae). *Biol. Control* **2006**, *36*, 22–31.
25. Forseth, I.N.; Innis, A.F. Kudzu (Pueraria montana): History, Physiology, and Ecology Combine to Make a Major Ecosystem Threat. *Crit. Rev. Plant Sci.* **2004**, *23*, 401–413.
26. Follak, S. Potential distribution and environmental threat of *Pueraria lobata*. *Open Life Sci.* **2011**, *6*, 457–469.
27. Cheng, Y.-B.; Tom, E.; Ustin, S.L. Mapping an invasive species, kudzu (*Pueraria montana*), using hyperspectral imagery in western Georgia. *J. Appl. Remote Sens.* **2007**, *1*, 013514.
28. Li, J.; Bruce, L.M.; Byrd, J.; Barnett, J. Automated detection of *Pueraria montana* (kudzu) through Haar analysis of hyperspectral reflectance data. *Proc. IEEE IGARSS* **2001**, *5*, 2247–2249.
29. Colditz, R.R. An evaluation of different training sample allocation schemes for discrete and continuous land cover classification using decision tree-based algorithms. *Remote Sens.* **2015**, *7*, 9655–9681.
30. Heydari, S.S.; Mountrakis, G. Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites. *Remote Sens. Environ.* **2018**, *204*, 648–658.
31. Millard, K.; Richardson, M. On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. *Remote Sens.* **2015**, *7*, 8489–8515.
32. Ball, J.E.; Anderson, D.T.; Chan, C.S. Comprehensive survey of deep learning in remote sensing: theories, tools, and challenges for the community. *J. Appl. Remote Sens.* **2017**, *11*, 042609.
33. ESRI. ArcGIS Desktop and Spatial Analyst Extension: Release 10.1. *Environmental Systems Research Institute,* **2012**, Redlands, CA.
34. Ramezan, A.C.; Warner, A.T.; Maxwell, E.A. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sens.* **2019**, *11*, 185.
35. Qi, Z.; Yeh, A.G.-O.; Li, X.; Lin, Z. A novel algorithm for land use and land cover classification using RADARSAT-2 polarimetric SAR data. *Remote Sens. Environ.* **2012**, *118*, 21–39.
36. Feng, Q.; Liu, J.; Gong, J. UAV Remote sensing for urban vegetation mapping using random forest and texture analysis. *Remote Sens.* **2015**, *7*, 1074–1094.
37. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
38. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
39. Cohen, J. A Coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46.
40. Juel, A.; Groom, G.B.; Svenning, J.-C.; Ejrnæs, R. Spatial application of Random Forest models for fine-scale coastal vegetation classification using object based analysis of aerial orthophoto and DEM data. *Int. J. Appl. Earth Obs. Geoinformation* **2015**, *42*, 106–114.

41. Colditz, R.R.; Schmidt, M.; Conrad, C.; Hansen, M.C.; Dech, S. Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions. *Remote Sens. Environ.* **2011**, *115*, 3264–3275.
42. Jin, H.; Stehman, S.V.; Mountrakis, G. Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado. *Int. J. Remote Sens.* **2014**, *35*, 2067–2081.