International Journal of Applied Earth Observation and Geoinformation 49 (2016) 87–98

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

A systematic comparison of different object-based classification techniques using high spatial resolution imagery in agricultural environments

Manchun Li[a], Lei Ma[a,b,*], Thomas Blaschke[b], Liang Cheng[a], Dirk Tiede[b]

[a] Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University, 210023 Nanjing, China
[b] Department of Geoinformatics—Z_GIS, University of Salzburg, Hellbrunner Str. 34, A-5020 Salzburg, Austria

ARTICLE INFO

Article history:
Received 27 November 2015
Received in revised form 31 January 2016
Accepted 31 January 2016

Keywords:
GEOBIA
OBIA
Random Forest
Segmentation scale
Training set size
Feature selection
Mixed object
Classification
High spatial resolution

ABSTRACT

Geographic Object-Based Image Analysis (GEOBIA) is becoming more prevalent in remote sensing classification, especially for high-resolution imagery. Many supervised classification approaches are applied to objects rather than pixels, and several studies have been conducted to evaluate the performance of such supervised classification techniques in GEOBIA. However, these studies did not systematically investigate all relevant factors affecting the classification (segmentation scale, training set size, feature selection and mixed objects). In this study, statistical methods and visual inspection were used to compare these factors systematically in two agricultural case studies in China. The results indicate that Random Forest (RF) and Support Vector Machines (SVM) are highly suitable for GEOBIA classifications in agricultural areas and confirm the expected general tendency, namely that the overall accuracies decline with increasing segmentation scale. All other investigated methods except for RF and SVM are more prone to obtain a lower accuracy due to the broken objects at fine scales. In contrast to some previous studies, the RF classifiers yielded the best results and the k-nearest neighbor classifier were the worst results, in most cases. Likewise, the RF and Decision Tree classifiers are the most robust with or without feature selection. The results of training sample analyses indicated that the RF and adaboost. M1 possess a superior generalization capability, except when dealing with small training sample sizes. Furthermore, the classification accuracies were directly related to the homogeneity/heterogeneity of the segmented objects for all classifiers. Finally, it was suggested that RF should be considered in most cases for agricultural mapping.

1. Introduction

Recent advances in the airborne and spaceborne remote sensing technology and image segmentation techniques offer new opportunities for remote sensing agricultural mapping (Wulder and Coops, 2014; Ma et al., 2014; Zhang et al., 2015), whilst the OBIA/GEOBIA ((Geographic) Object-based Image Analysis) paradigm in the field of remote sensing classification is already widely accepted (Liu et al., 2006; Blaschke et al., 2014). Plenty of classification approaches are documented within the GEOBIA framework, especially the implementation of expert rule-sets, which make use of the extremely extended feature space spanned by the use of the available object-specific features at several segmentation scales (context features/neighborhood relation, scaled-hierarchy relations, form features etc.) (Benz et al., 2004; Blaschke, 2010; Tiede et al., 2010; Strasser and Lang, 2015). Nevertheless, supervised classification algorithms based on objects rather than pixels as classification units are still very important. According to previous studies, the comparison of classification approaches within a GEOBIA framework can be divided into two general topics: 1) a comparison of GEOBIA and traditional per-pixel image analysis; and 2) a comparison of different classification techniques within GEOBIA only. Although there is general agreement regarding the former (Yan et al., 2006; Duro et al., 2012), the selection of a suitable classifier is still a problem for any per-pixel and GEOBIA method due to the diversity of data sources, the training set size and feature and, especially, the selection of segmentation parameters (e.g. scale/size of objects) and spectrally mixed objects bringing some uncertainty into the comparison of methods (Yu et al., 2008). In the following sub-section we

* Corresponding author at: Jiangsu Provincial Key Laboratory of Geographic Information Science and Technology, Nanjing University, 210023 Nanjing, China.
E-mail address: maleinju@gmail.com (L. Ma).

http://dx.doi.org/10.1016/j.jag.2016.01.011
0303-2434/

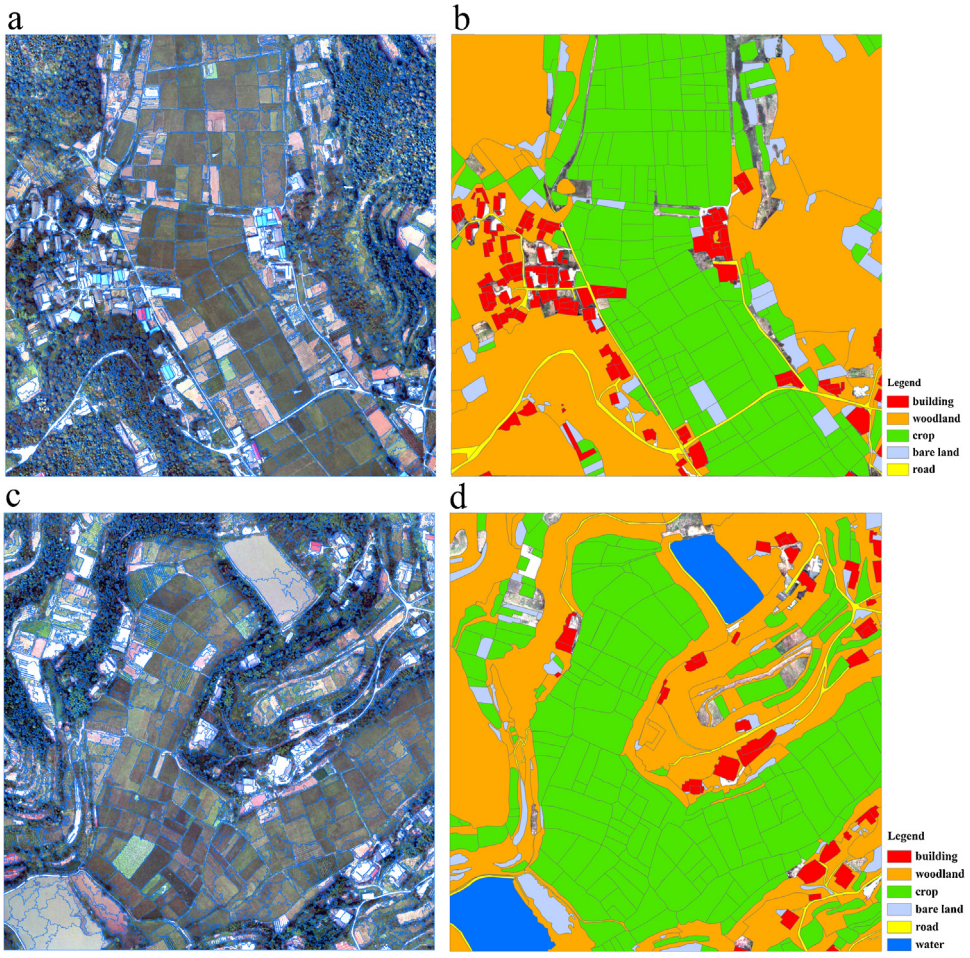

Fig. 1. Study sites. The segmented layers and the corresponding reference layers. (a) The segmented layer for area 1 at a scale of 100; (b) the manual interpretation layer for area 1; (c) the segmented layer for area 2 at a scale of 100; and (d) the manual interpretation layer for area 2.

conduct a brief literature review and succinctly identify the benefits and the main difficulties when comparing classification techniques.

Despite the expert rule-set classifications in GEOBIA, researchers already try using statistical and machine-learning classification techniques, including Linear Discriminant Analysis (LDA) (Pu and Landry, 2012), Random Forest (RF) (Stumpf and Kerle, 2011), Decision Tree (DT) (Mallinis et al., 2008), K-Nearest Neighbor (KNN) (Luque et al., 2013), naiveBayes (Dronova et al., 2012), Support Vector Machines (SVM) (Heumann, 2011). Since around 2010, the Adaboost technique, as another ensemble classifier, has received more attention in remote sensing classification due to the high accuracy (Chan and Paelinckx, 2008). So far, only a few GEOBIA applications have used ensemble classifications beyond RF. Thus, Adaboost as another example of ensemble classifier was used with GEOBIA other than RF.

In per-pixel analysis, the classification accuracy is usually accredited to the classification technique (Rogan et al., 2008). Chan and Paelinckx (2008) also evaluated Random Forest and Adaboost tree-based ensemble classifications using airborne hyperspectral imagery and yielded almost the same accuracy results as established per-pixel classifiers. Brenning (2009) compared eleven classification algorithms in automatic rock glacier detection using terrain analysis and multispectral remote sensing, and found that mapping results of PLDA (Penalized Linear Discriminant Analysis) are significantly better than all other classifiers, including both SVM and RF. For the purpose of land-cover classification, Shao and Lunetta (2012) compared the support vector machine, neural network, and CART (classification and Regression Trees) algorithms using Moderate Resolution Imaging Spectroradiometer time-series data, and found that SVM was the superior algorithm. Xu et al. (2014) compared seven classification techniques for marine oil spill identification using RADARSAT-1 imagery and showed that the classification was able to benefit from ensemble techniques (bundling and bagging). These few examples demonstrate the importance of selecting optimal classifiers for remote sensing classification or prediction.

We hypothesize that for GEOBIA it is not sufficient to analyze the choice of the classifier only, because the resulting accuracies also depend on the segmentation scale, on the selection of features, and on the existence of spectrally mixed objects (Ma et al., 2015). It therefore seems to be impossible to generically advise on the selection of a specific classification technique for a specific application case. For instance, Laliberte et al. (2006) and Mallinis et al. (2008) found that the overall classification accuracies of the classification tree was better than that of the K-NN algorithm. In contrast, Tehrany et al. (2014) suggested that K-NN generally performed better for land-cover mapping, while they compared with DT and SVM using SPOT 5 imagery. In addition, previous researches (Duro et al., 2012; Ghosh and Joshi, 2014) demonstrated a superior capability of producing higher classification accuracies by SVM or RF, but Dronova et al. (2012) concluded that RF always performed worse, while they examined six families of statistical machine-learning classifiers. We assumed that these inconsistent results are related to the unsystematic studies of comparison, while, as mentioned before, the vast majority of comparisons analyzed only a single factor (i.e., scale) or relatively few classification algorithms.

The objectives of this study are to implement a comprehensive comparison of classification performances in GEOBIA using a variety of statistical and machine-learning classification methods, and to generate robust results using advanced statistical evaluation methods, and then to determine the most suitable classifier for agricultural mapping in different cases using UAV (Unmanned Aerial Vehicle) high spatial resolution imagery. According to the earlier studies, many factors, including scales, features, samples and mixed objects, notably influence the classification performance (Ma et al., 2015). In the first part of this article, we conducted a visual assessment and multiple comparisons of the repeated classification measures using a variety of scale groups. Then, we assessed the implications of additional feature selection techniques for each of the selected classification algorithms. Moreover, since we hypothesize that the sizes of the training samples strongly affect the remote sensing classification results, we additionally conducted experiments for each classifier by changing the training samples sizes while employing the same statistical assessment methods. Finally, an assessment of "mixed objects" – spectrally mixed objects created through under-segmentation – was performed to contribute to uncertainty research in GEOBIA. Recommendations of the most suitable algorithms were made for different cases using high spatial resolution earth observation imagery.

2. Methods

2.1. Data set and preprocessing

The data used for training and assessing the classifiers stems from a project on high resolution imagery collection in Deyang city (Ma et al., 2013) situated in the hilly northeast area of the Chengdu plain in Sichuan Province, China. The images, including three visible bands (blue, green and red) with a 0.2 m spatial resolution, were acquired with an unmanned aerial vehicle at a height of 750 m in August 2011. Subsequently, the digital orthophoto map (DOM), as a 500 × 500 m standard map sheet, was produced by digital photogrammetry software using the collected control points (Ma et al., 2013). Both standard map sheets were then used to examine the effects of segmentation scale, training set size, feature selection, and mixed objects on each classification technique.

Due to the very high spatial resolution, the classified images were also used as a reference image to delineate sample units (polygons) for several species/groups. The reference vector layers (Fig. 1) were produced by manual interpretation of both areas, and was prepared for sampling and validation. Study area 1 comprised a variety of land cover types, but consisted mainly of cropland (38%) with paddy fields and dry land. The remainder of study area 1 was characterized by the presence of woodland (43%), buildings (6%), bare land (5%), and roads (2%). Study area 2 comprised cropland (45%), woodland (37%), buildings (4%), bare land (4%), roads (1%), and water (5%).

2.2. Segmentation and sampling

Based on the generated DOM data, a series of segmented patches with 19 different segmentation scales were generated using the multi-resolution segmentation algorithm (Baatz and Schaepe, 2000) as implemented in the eCognition software package (Trimble Geospatial), starting with a scale of 20 and ending with 200 at an interval of 10. A bottom-up region merging technique was used in eCognition from one-pixel objects, and then smaller image objects were merged into larger ones. The weights of color/shape and smoothness/compactness were the same at all scales. The color/shape parameters were set to 0.9/0.1, respectively, because we wanted the spectral information to have the dominant role during segmentation. The smoothness/compactness ratio was set to 0.5/0.5 because we did not want to favor compact or non-compact segments. The weights of the different image layers were equal for all three bands to avoid any bias. Following the segmentation, we exported the segmented objects of each scale together with the values for 30 features (including spectral, texture and shape) in order to apply the various classification methods in different software packages.

In order to implement stratified random sampling (Congalton and Green, 2009), we first labelled all segmented objects using a GIS overlay ratio rule between the segmented objects and reference polygons. When studying the effect of the scale, feature selection and training set size, the segmented object was defined as the class covering more than 50% of the reference polygon. Subsequently, all segmented objects were divide into groups (5 groups for area 1 and 6 groups for area 2 according to the classes of manual interpretation map in Fig. 1b and d) and then each stratum was randomly sampled with the same training set ratio (proportion of the segmented objects used for training in a certain segmented layer). In this study, a training set ratio of 30% sampling, which proved to be useful for multi-scale (scale < 200) comparisons (Ma et al., 2015), was used to evaluate the influence of scale and feature selection on each classifier. Then, the reference polygons were used as reference set to validate all re-labelled objects by each classifier. For the assessment of the training set size, we utilized the number of training objects instead of the ratio, for which the scale was fixed at 80, approximating an optimal segmentation scale according our previous works (Ma et al., 2015). Finally, with a training set ratio of 30% sampling, the effect of mixed objects to classifiers was investigated by changing the overlay ratio, i.e., using overlay ratios of 0.7 and 0.9 in addition to the standard ratio of 0.5. Although the training sets used in later analyses varied, the same testing set was used in the evaluation of all classifications. This testing set comprised all re-labelled objects of each class.

2.3. Correlation-based feature selection

Previous studies found that most of the large number of features in GEOBIA were strongly correlated (Ma et al., 2015). To obtain the optimal feature subset, correlation-based feature selection (CFS) was used to measure the quality of a subset of features. This was achieved by aggregating the best-first search strategy based on the results of the feature importance evaluation (Hall and Holmes, 2003). As mentioned above, we performed CFS on all of the calculated features before each classification, including many spectral measures: mean blue, mean green, mean red, max difference, standard deviation blue, standard deviation green, standard deviation red, and brightness; the texture measures: GLCM (Gray-Level Co-occurrence Matrix) homogeneity, GLCM contrast, GLCM dissimilarity, GLCM entropy, GLCM std. dev., GLCM correlation, GLCM ang. 2nd moment, GLCM mean, GLDV (Gray-Level Difference Vector) ang. 2nd moment, GLDV entropy, GLDV mean, and GLDV contrast; the shape measures: area, compactness, density, roundness, main direction, rectangular fit, elliptic fit, asymmetry, border index, and shape index. Hierarchical features and class-related features (features that are only applicable after initial classifications) were not taken into consideration in accordance with the test design implemented in this study.

2.4. Classifiers

2.4.1. Support vector machines

A SVM is a non-parametric supervised learning classifier which has become increasingly popular in remote sensing applications (Otukei and Blaschke, 2010). To facilitate the SVM algorithm, our study employed the R package 'e1071', which was implemented in

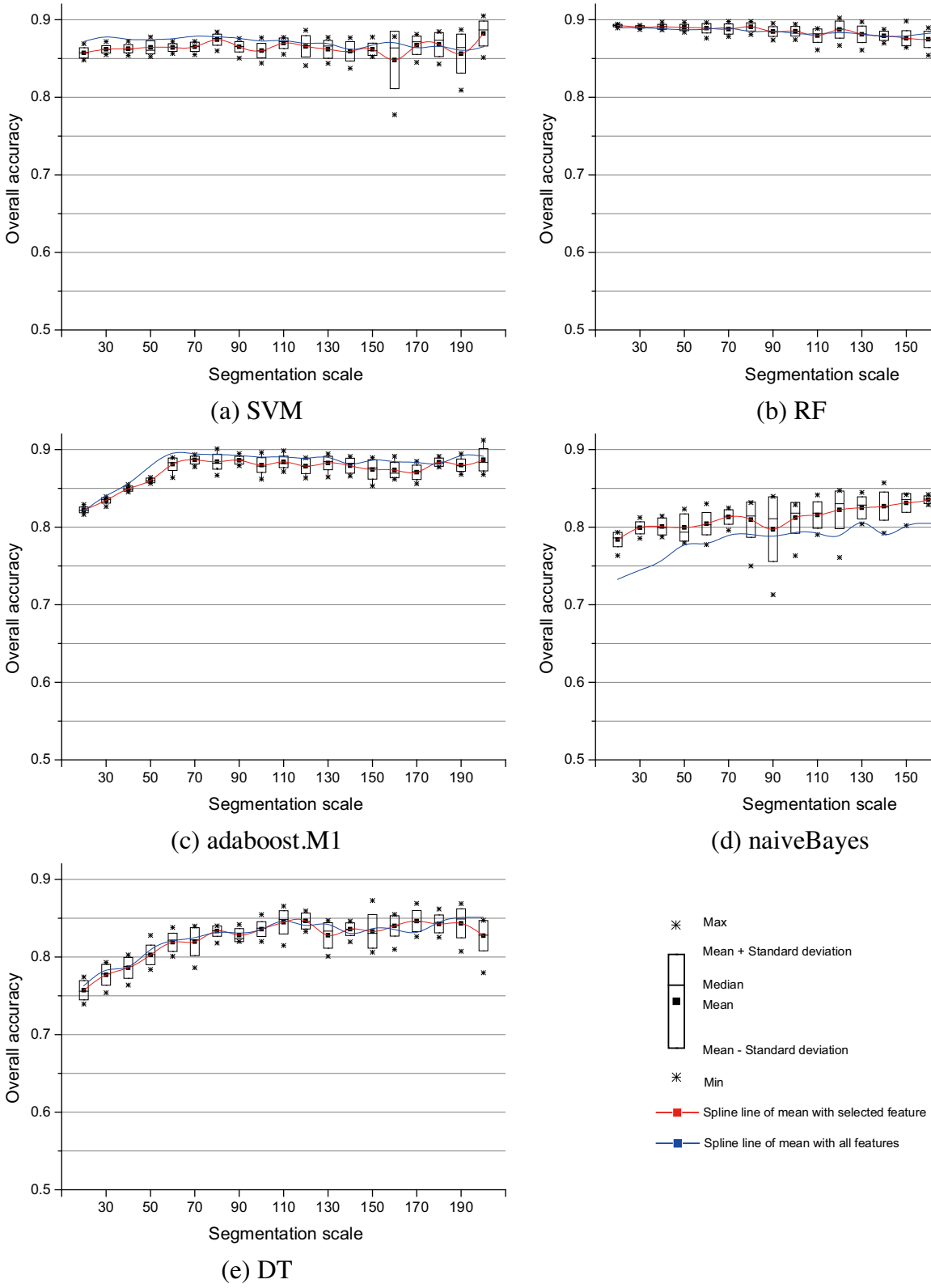

Fig. 2. Overall accuracy with different scales for Area 1 shown for the best five classifiers. For each scale, ten independent stratified random samples were implemented, and the statistics (mean, median, standard deviation, minimum, and maximum) of ten OAs were shown. Red lines represent the fitted spline line of means classified using selected features while blue lines indicate the fitted spline line of means classified using all features.(For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the LIBSVM library by Chang and Lin (2011) and provides four different kernels. Following the recommendation of Hsu et al. (2010), we adopted the radial basis function (RBF) kernel in this study. For the RBF kernel, two parameters need to be acquired beforehand, including the penalty parameter C and the kernel parameter γ. In order to find the best values for these parameters, the grid-

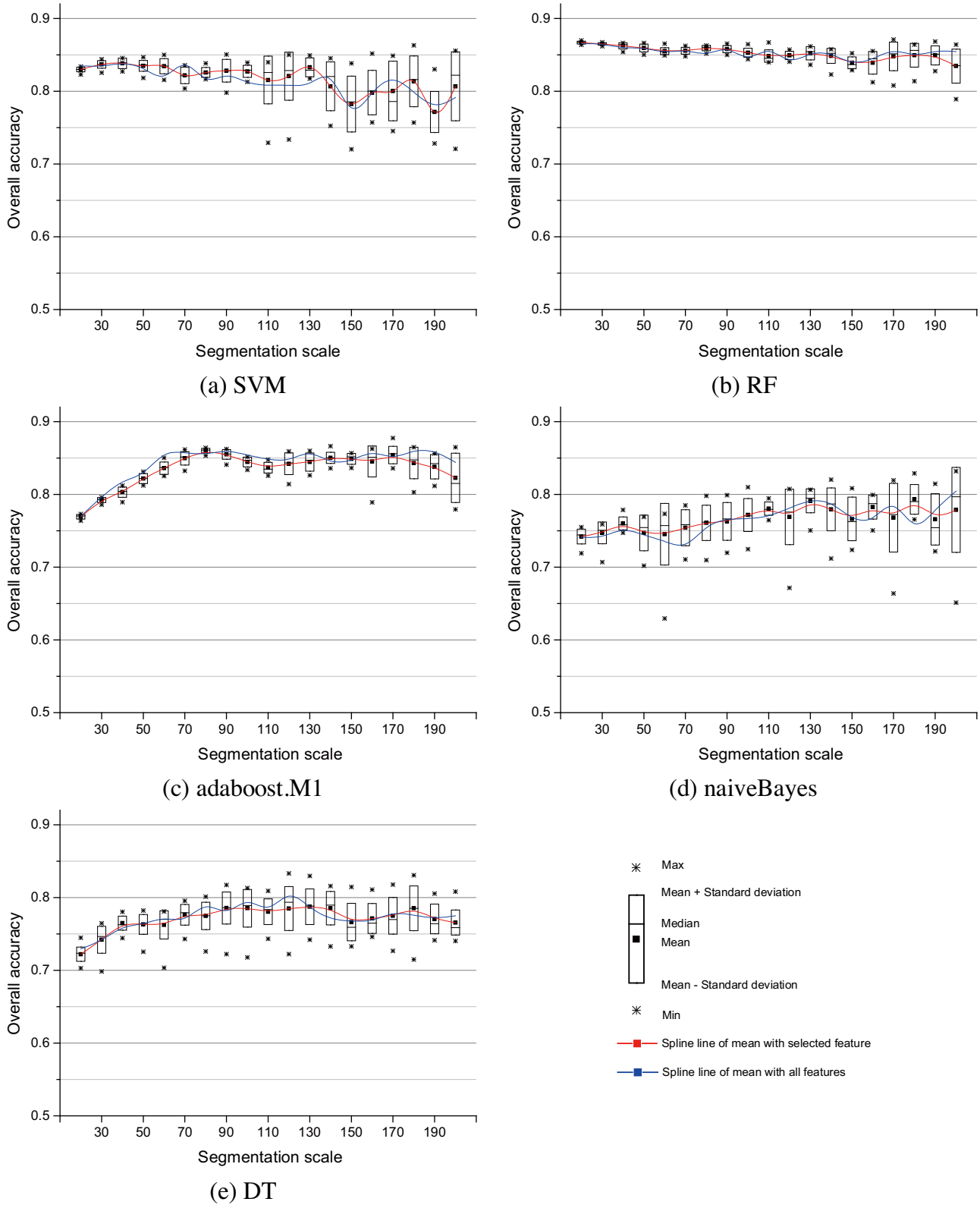

Fig. 3. Overall accuracy for Area 2 using the best five classifiers with different scales. For explanation see Fig. 2.

search method was used to identify the pair (C, γ) that achieved the best cross-validation accuracy for the training and validation sets. A coarse grid, which is a two dimensional parameter space along 2^d, where $d = -4, -3.5, -3, \ldots, 1$ for γ and $d = -4, -1.5, -1, \ldots, 4$ for C, was used to avoid a complete grid-search.

2.4.2. Random Forest

Since the RF classifier was proposed (Breiman, 2001), it has been improved continuously in the field of remote sensing image information extraction, where it has been shown to be a robust classifier (Chan and Paelinckx, 2008). The RF employs a random method to

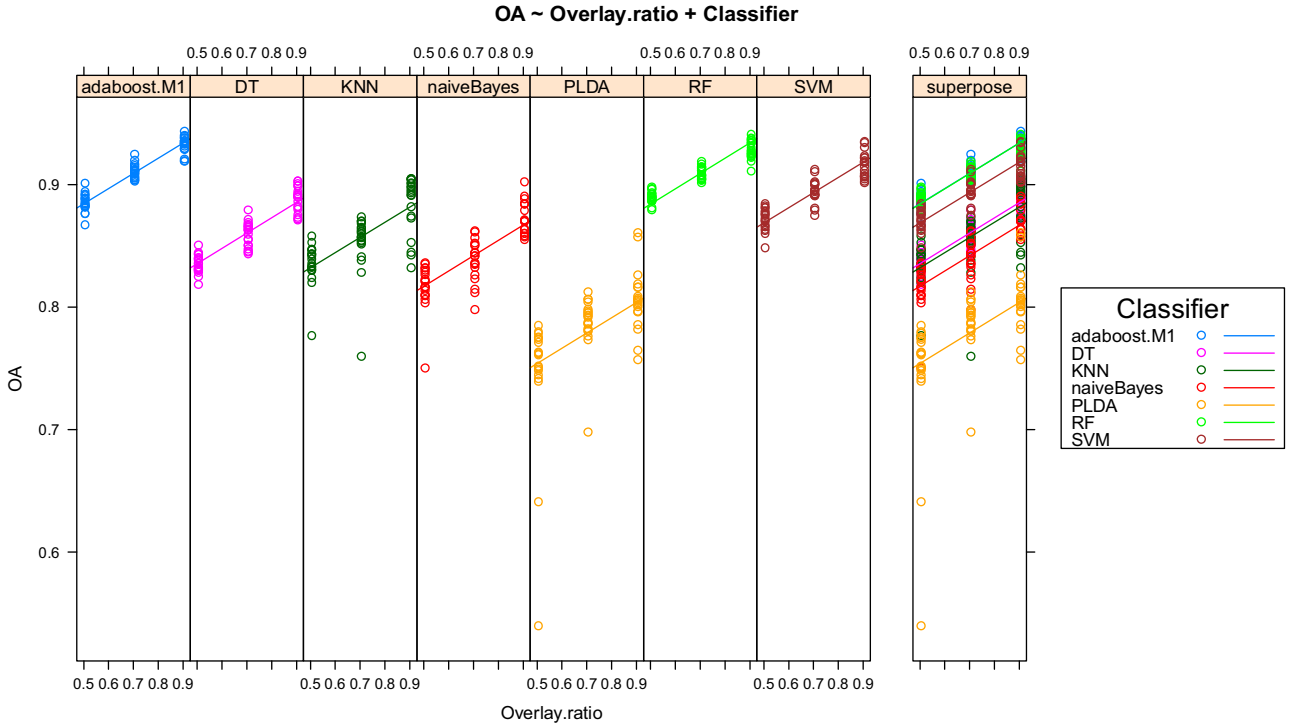

Fig. 4. Scatterplot of overall accuracy against overlay ratio for each classifier in Area 1. Adjusted lines from linear regression analyses are shown for each classifier.

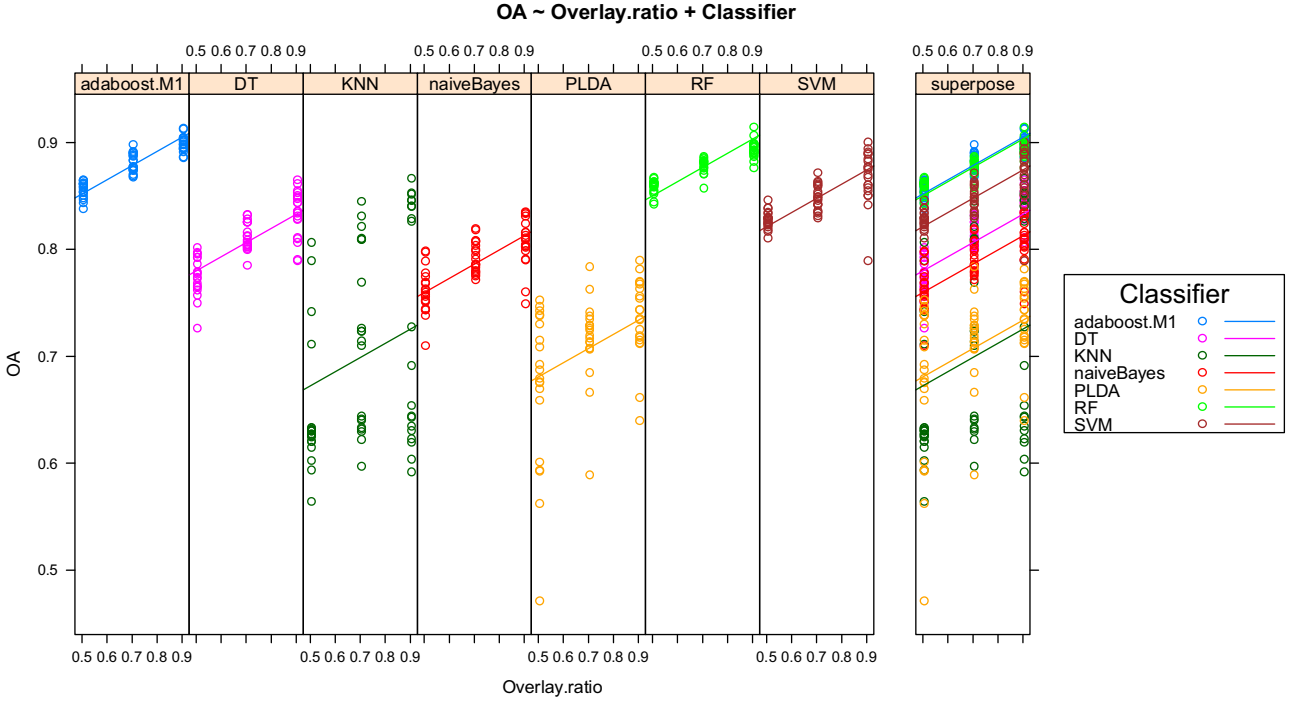

Fig. 5. Scatterplot of overall accuracy against overlay ratio for each classifier in area 2. Adjusted lines from linear regression analyses are shown for each classifier.

establish a forest comprising many mutually independent decision trees. After obtaining the forest using the training set, we let each decision tree in the forest make a judgment about the unlabeled sample before the unlabeled sample was predicted as the category that was voted for most frequently. The RF classifier only requires the definition of two parameters to generate a prediction model: the number of classification trees desired (k) and the number of prediction variables (n) used in each node to make the tree grow. When considering the parameter set to be used for a remote sensing classification with the RF classifier, Rodriguez-Galiano et al. (2012) showed that it was preferable to use a large number of trees (k) and a small number of split variables (n) to reduce the generalization error and the correlation between trees. Based on their research results, our RF model used large RFs with 479 trees to avoid the over-fitting problem and one single randomly split variable (feature) for UAV optical imagery classification. The over-

all classification procedure was performed automatically using the package 'RandomForest' in R. Bagging techniques are not discussed here due to the fact that the RF classifier is implemented in this study.

2.4.3. K-Nearest Neighbor

In general, the KNN has been used in many GEOBIA workflows (Luque et al., 2013) due to its simplicity and flexibility. Since the KNN is the most frequently used classification method in the widely used OBIA software framework eCognition (Trimble Geospatial), it is necessary to include it in this comparison. In contrast to model-based learning, the KNN assigns the object to the class based on nearest neighbors in the feature space rather than learning from a model. To predict a new object, the closest K neighbors are found from the training set and then used to vote for the final prediction. K is a tunable parameter and a typically small (i.e., 1, 2, . . .) positive integer. In this study, we used an optimal K parameter based on cross-validation, and bootstrap samples are employed to search the best K value in the R package 'e1071', where K can range from 1 to 10 in steps of 1. The best K was then applied to the KNN algorithm, which was implemented in the R package 'class'.

2.4.4. Decision tree

The use of decision trees for remote sensing image classification has increased in recent years (Peña-Barragán et al., 2011). For GEOBIA, the most important phase is the construction of an image interpretation model (knowledge) for the segmented objects. However, it may be difficult to execute in combination with other classifiers, considered to be a "black box", while DTs are like a "white box": users are easily able to interpret the links between the response variables of classes and the explanatory variables of remote sensing data. Consequently, researchers are particularly interested in classification trees within GEOBIA workflows (Laliberte and Rango, 2009). In this study, a tree is grown by binary recursive partitioning using the response in the specified formula and choosing splits from the terms of the right-hand-side. The split which maximizes the reduction in impurity is chosen, the data set is then split and the process repeated. Splitting continues until the terminal nodes are too small or too few to be split. Here, the tree growth is limited to a depth of 31, and the package 'tree' in R is employed to classify our data.

2.4.5. Adaboost.M1

In addition to bagging, boosting is another popular method for ensembles of trees (Alfaro et al., 2013). Adaboost is the popular approach used for boosting, and has recently been in implemented the study of remote sensing image classification (Chan and Paelinckx, 2008). Adaboost can process data with weights, and the misclassification rate for each trial is used to update the distribution over the training samples. The weights of misclassified samples are increased, while the weights of the rightly classified samples are decreased by controlling the misclassification rate and forcing the classifier to focus on the hardest samples in the next iteration (Alfaro et al., 2013). Finally, the voting for the labels is weighted by the accuracy of each classifier. In this work, the R package 'adabag' was used to implement Freund and Schapire's Adaboost.M1 (Freund and Schapire, 1996) algorithm using classification trees as individual classifiers. Here, the constant factor alpha as a learning rate was calculated using a function of the error and weights recommended by Freund and Schapire (1996), and then a bootstrap sample of the training set was drawn using the weights for each trial on that iteration. Both the number of iterations and the number of trees were set to 100, which is the default in both cases.

2.4.6. naiveBayes

Bayes Network is a powerful probabilistic representation and reasoning tool when dealing with conditions of uncertainty (Ouyang et al., 2006). It has also been widely used as a strategy or single classifier for remote sensing classifications owing to its highly scalable and incremental learning (Yang and Wang, 2012). Recent comparative research has also focused on this classification algorithm in GEOBIA (Dronova et al., 2012). From a probabilistic viewpoint, if X indicates the feature of an object and Y represents the type of classes, the predictive problem can be viewed as a conditional probability estimation, trying to find Y where $P(Y \mid X)$ is maximized ($P(Y \mid X) = P(X \mid Y) \times P(Y)/P(X)$). This is equivalent to finding Y where $P(X \mid Y) \times P(Y)$ is maximized. The standard naive Bayes classifier assumes independence of the predictor variables, at least within this study. The problem is to find Y to maximize $\prod_{i=1}^{n} P\left(\frac{X_i}{Y}\right) \times P(Y)$. Here, we implement the classifier using the function with default parameters in the R package 'e1071'.

2.4.7. Penalized Linear Discriminant Analysis

Since Penalized Linear Discriminant Analysis (PLDA) has been proven to be capable of processing data with a large number of highly correlated variables, it has been widely used in hyperspectral remote sensing images classification (Yu et al., 1999), but also for other purposes such as, for example, forest biomass estimation (Fassnacht et al., 2014). Furthermore, various features calculated for each object are similar to the hyperspectral data, especially for the high spatial resolution images with more spectra (i.e., Worldview 3), and hence we also examined the PLDA to label the objects with correlated co-variables. PLDA is a general approach for penalizing the discriminant vectors in Fisher's discriminant problem in such a way as to increase interpretability (Witten and Tibshirani, 2011). In this study, PLDA was implemented in the R package 'penalizedLDA'. The lasso penalty tuning parameter lambda = 0.14 was used by default, and the number of discriminant vectors was equal to the number of classes minus 1, since it must be no greater than the number (classes −1) and we do not want to lose important information.

2.5. Accuracy assessment and statistical comparison

There are many ways to determine accuracy. In general, the overall accuracy has widely been recommended as the primary accuracy measure (Congalton and Green, 2009), which summarizes the percentage of correct classification by confusion matrix. We therefore employed the overall accuracy as the main measure of comparison. For this study, we divided the area sum of the major diagonal in confusion matrix by the total area of all objects involved in classification to calculate the overall accuracy. It is a single summary measure directly related to commission and omission accuracies of every class derived from the confusion matrix. Additionally, due to the uncertainty of segmented objects, the area-based (polygon-based) accuracy assessment method for the overall accuracy is selected to calculate the confusion matrix (Whiteside et al., 2014).

In order to select several suitable tests and measures, we first reviewed the current practice for statistical comparison of remote sensing classification accuracy. In practice, the McNemar test has widely been used to assess whether significant differences between classification accuracies exist (Foody et al., 2006), but it must follow a chi-square distribution or one following approximately a normal distribution. Since the area-based accuracy assessment method used in this study is time-consuming in calculating each individual accuracy, we could not repeat the classification too many times, especially for fine scales. Consequently, the samples were not large enough to perform a statistical analysis with parametric

assumptions. In such cases, it has been recommended to use non-parametric test methods without assumptions to assess whether statistically significant differences between classifications exists (Brenning, 2009).

In this study, we first employed the Kruskal-Wallis test, which made no assumptions, to ascertain whether there were any differences between different scale groups for each classifier. Furthermore, in order to compare the performance of classifiers in respect to scale and training set size, and taking into account that the Friedman test indicates significance globally, the *post-hoc* test according to Nemenyi was employed to conduct multiple comparisons on all pairs of classifiers (Demšar, 2006). In addition, Friedman tests were conducted to compare the differences with and without feature selection. Finally, we increased the times of the repeated classification at scale 80 to collect more accuracy samples. Subsequently, ANCOVA was employed to estimate the interaction effect of the overlay ratio and classifier.

3. Results

3.1. Responses of classifiers to segmentation scale

3.1.1. Visual assessment

Figs. 2 and 3 reveal different patterns of accuracy changes from finer to coarser segmentation scales for individual classifiers. In contrast, Dronova et al. (2012) reported that the magnitudes of the overall accuracy values and change were relatively similar among different classifiers for GEOBIA. In order to analyze the response of classifiers to the segmentation scale, the statistics, including means, medians, standard deviations, minimums, and maximums, of the 10 overall accuracies from the ten training datasets of stratified random sampling (using stable training set ratio of 30% sampling) for each scale were displayed. For these two areas, SVM and RF show consistent trends along with increasing segmentation scale, that overall accuracies decrease and standard deviations increase. In order to compare the differences between scales for each classifier, we also implemented the Kruskal-Wallis test for different scale groups (i.e., (20, 30), (20, 30, 40), . . ., (20, 30, . . ., 200)). Consequently, for the RF and SVM classifiers, the Kruskal-Wallis test on the null hypothesis, stating that the performances between scales cannot be distinguished in Area 1, was accepted at the 5% significance level (p-value: <0.001) until the scale 110 was involved in this test. In area 2, the same effect occurred at scale 80 for the SVM, and at scale 60 for the RF. In other words, the classification accuracy stabilized at finer scales with SVM and RF classifiers, but it did not stabilize at coarse scales. These differences from statistical tests may be used further to detect optimal segmentation scales.

The performance of naiveBayes and DT were also associated with increasing segementation scales. In contrast to SVM and RF, the higher the scale parameter for naiveBayes and DT, the better the achieved accuracy. The direction that maximizes the segmentation scales should be favoured for the naiveBayes and DT classifier, though they were influenced slightly by the mixed objects at coarser scales. Adaboost.M1 outperformed the DT classifier, which may be a typical advantage of an ensemble classifier. Notably, for both study areas adaboost.M1 performed worse in fine scales when it was not expected because of the similar ensemble learning technique with RF, but showed decreasing standard deviations for finer scales. KNN and PLDA were not shown here because of their worst performances.

3.1.2. Multiple comparisons between classifiers

We conducted multiple statistical comparisons to evaluate the performances of the best five classifiers studied. To facilitate the comparisons at different scale levels, we divided the scales into four groups, namely group 1 (20 ~ 50), group 2 (60 ~ 100), group 3 (110 ~ 150), and group 4 (160 ~ 200). As the Friedman test indicates significance globally ($p < 0.001$) for each group, we continued to conduct multiple comparisons in order to identify differences between the classifiers. According to the Nemenyi *post-hoc* test for multiple joint samples of overall accuracy, test statistics (q) for each comparison were calculated as shown in Table 1 (KNN and PLDA not shown here). Here, the critical test statistics were identified to be 4.17 ($\alpha = 0.05$, ∞ degrees of freedom). For medium and coarse scale groups, the null hypothesis, stating that there is no significant performance difference between RF and Adaboost.M1, was accepted ($q < 4.17$). DT and naiveBayes also showed no significant performance difference for all scale groups in both areas. Consequently, we can conclude the single cliques of two algorithms (RF and Adaboost.M1, DT and naiveBayes) and a group with one algorithm only (SVM). More specifically, based on the comparative rank scores of the classifiers (Table 1), the results also indicate that RF is significantly better than all other classifiers except for Adaboost.M1. When ranking the classifiers, RF is first; SVM performs significantly better than DT and naiveBayes.

Table 1
Test statistics (q) for the Nemenyi *post-hoc* tests between classifiers for four scale groups.

		RF	SVM	Adaboost.M1	naiveBayes
Area1	SVM	3.29*			
		5.30			
		3.93*			
		3.60*			
	Adaboost.M1	5.49	2.20*		
		1.57*	3.73*		
		0.00*	3.93*		
		0.07*	3.54*		
	naiveBayes	10.03	6.73	4.54	
		13.88	8.58	12.31	
		11.59	7.66	11.59	
		8.58	4.98	8.51	
	DT	12.52	9.22	7.03	2.49*
		11.65	6.35	10.08	2.23*
		9.49	5.56	9.49	2.09*
		9.30	5.70	9.23	0.72*
Area2	SVM	4.14*			
		7.22			
		8.33			
		9.26			
	Adaboost.M1	8.64	4.50		
		2.22*	5.00		
		0.79*	7.55		
		1.76*	7.50		
	naiveBayes	15.37	11.23	6.73	
		17.13	9.91	14.91	
		14.58	6.25	13.79	
		12.45	3.19*	10.69	
	DT	15.47	11.33	6.83	0.10*
		14.54	7.31	12.31	2.59*
		13.79	5.46	13.01	0.79*
		13.47	4.21	11.71	1.02*

The symbol "*" indicates that the difference is not statistically significant, because the values are below the critical value of 4.17 ($\alpha = 0.05$, ∞ degree of freedom).

3.2. Comparisons between selected and all features

The Friedman test was also conducted to test whether there was a significant difference in the classification accuracy of repeated measurements between models with and without certain features selected for each scale group (scale 20, 30, 40 and 50 in one group, and every five scales in another group, to provide more samples for each Friedman test). For most of the scale groups, there was a significant difference in the accuracy indices of adaboost.M1 between having selected features and all features ($p < 0.01$), while these dif-

Table 2
p-values from the Friedman test for each scale group using CFS features and using all features.

Group		1	2	3	4
Scale		20 ~ 50	60 ~ 100	110 ~ 150	160 ~ 200
Area1	SVM	0**	0**	0.0237	0.7773
	RF	0.0578	0.0477	1	0.1573
	naiveBayes	0**	0**	0**	0**
	DT	0.0578	0.5716	0.7773	0.3961
	adaboost.M1	0**	0**	0.0007**	0.0047*
Area2	SVM	1	0.5716	0.2579	0.7773
	RF	0.0114	0.0237	0.3961	0.0237
	naiveBayes	0.0008**	0.0278	0.6892	0.5485
	DT	0.8231	0.1615	0.3173	0.8415
	adaboost.M1	0.0005**	0.0019*	0.0109	0.0019*

ferences were not statistically significant for DT and RF classifiers at $p = 0.01$ for both areas (Table 2). For the SVM, the results also revealed that the accuracy using all features was significantly better than when using selected features at fine scales for Area 1 ($p < 0.01$), while there was no significant difference at $p = 0.01$ for area 2. In addition, naiveBayes with selected features always significantly ($p < 0.01$) improved the accuracy for Area 1, while this also occurred at scale group 1 for area 2. Therefore, it was suggested that RF and DT were most robust classifiers with or without feature selection, while the other classifiers were affected more or less due to feature selection or the redundant data with all features.

3.3. The effect of training sample size on classification accuracy

To assess the impact of the training set size on each classifier, we fixed the scale parameter at 80, which was assumed to be the optimal scale according to the previous analysis. Firstly, the classifier sensitivity to different training sample sizes was evaluated after various numbers of training samples were determined (20, 40, 60, 80, 100, 200, 300, 400, 500, 600, 700 and 800) instead of ratio of training samples. For each set of training sample sizes, objects for every sample size were randomly selected as the same ratio from each class samples, while at least one object was forced to be included from each land cover class. The results for both areas demonstrated that there is a positive relationship between the size of the training set and the classification accuracy for all tested classifiers (here the results not shown), that this is not surprising and in accordance with earlier findings (Ma et al., 2015).

To further investigate the performance of the classifiers on different sample sizes, the sample sizes were considered as three groups: a small sample size group (20, 40, 60 and 80), a medium sample size group (100, 200, 300 and 400) and a large sample size group (500, 600, 700 and 800). For each sample size group, we implemented multiple comparisons using the Nemenyi *post-hoc* test. There are no significant performance differences between the algorithms RF, adaboost.M1, SVM and DT in the small sample size group of both area 1 and 2 (Table 3). There is always statistically significant evidence at $\alpha = 0.05$ to indicate that there is no statistically significant difference between RF and Adaboost.M1 for all sample sizes in both areas. In addition, the comparative rank scores of classifiers for the medium sample size group and the large sample size group also show that RF achieved significantly better performance than the other classifiers except for the Adaboost.M1.

3.4. The effect of homogeneous and heterogeneous objects

A last experiment was performed to test the effect of the covariate (overlay ratio), which represented the proportion of segmented objects containing a specific reference object, and the class of the sampling object was determined using the class of the dominant

Table 3
The test statistics (q) for each comprison from the *post-hoc* test according to Nemenyi between classifiers for three training sample size groups.

		RF	Adaboost.M1	SVM	DT
Area1	Adaboost.M1	0.54*			
		1.20*			
		1.62*			
	SVM	0.68*	1.21*		
		9.27	8.06		
		8.68	7.06		
	DT	2.5*	3.04*	1.82*	
		13.85	12.65	4.59	
		18.98	17.36	10.30	
	naiveBayes	5.34	5.88	4.66	2.84*
		15.10	13.90	5.84	1.25*
		23.42	21.80	14.74	4.44
Area2	Adaboost.M1	0.80*			
		0.94*			
		1.12*			
	SVM	2.72*	3.52*		
		7.22	6.28		
		8.83	7.72		
	DT	1.35*	2.15*	1.37*	
		16.53	15.59	9.31	
		19.79	18.67	10.96	
	naiveBayes	2.53*	3.33*	0.19*	1.18*
		14.21	13.27	6.99	2.32*
		19.13	18.01	10.30	0.66*

The symbol "*" indicates that the difference is not statistically significant, while the values are below the critical value of 4.17 ($\alpha = 0.05$, ∞ degree of freedom).

overlay ratio. We repeated twenty classifications at scale 80 using overlay ratios of 0.5, 0.7 and 0.9 for both test sites, and the segmented object was defined as the class covering more than 50%, 70% and 90% of the reference polygon respectively for three overlay ratios. The ANCOVA was employed to quantify the relationships between the overlay ratios (the covariate) and OAs. Regarding both sites, Figs. 4 and 5 show scatterplots and adjusted regression lines displaying the relationship between the overlay ratio and OA for each of the seven classifiers (different colors and symbols). The results indicate that the relationships of overlay ratios and OAs among classifiers have no significantly difference ($p = 0.08708$ and 0.1717 for both areas, respectively). In the other words, the slopes of the lines for each classifier are very similar. Figs. 4 and 5 clearly indicate that there was a positive relationship between the overlay ratio and OA for all of the classifiers, because the decreased mixed objects with the increased overlay ratio are able to improve the accuracy. In addition, RF and Adaboost.M1 achieved almost identical regression lines, followed by SVM, DT and naiveBayes. It should be noted that these results were obtained at scale 80 and do not provide a complete picture of the overall performance of all classifiers used, but the results clearly show the effects of the mixed objects are serious for OBIA, since more mixed objects with low overlay ratio lead to smaller accuracy.

4. Discussion

In GEOBIA, the selection of specific parameters influencing the classification of objects (segmentation, segmentation scale/object size, training set size, sample scheme, the extended feature space etc.), supervised classification and the selection of adequate classifiers for GEOBIA still constitute a challenge. Our literature review also showed a large diversity of methods used to label objects from segmentation layers without a clear agreement on what methods perform the best. This study aimed to comprehensively investigate the performance differences of advanced classification techniques in agricultural environments. We therefore tested the sensitivity to scale (repeated 10 times for each scale) and training set size (repeated 20 times for each size at scale 80) for two case stud-

ies, and analyzed the effect of features and mixed objects on the overall classification accuracy of seven classifiers. We also stress that our results are based on high resolution UAV images covered most of cropland. Further studies could also determine whether our findings are altered using other resolution images, such as medium or low resolution remote sensing images. In addition, the well-designed study framework could have well transferability of exploring the uncertainty of object-based classification on other study sites or data sets by ensuring enough times of repeated classification. PLDA and KNN were not discussed further due to the lack of competitive performance.

4.1. Comparison of classifiers

In general, our studies indicated that the ensemble classifiers yielded significantly better results than the other classification techniques considered. This is in line with earlier findings for per-pixel studies (Chan and Paelinckx, 2008). We attribute this to the schemes of combination (bagging or boosting) for the single classifiers. We also agree that the performance of Adaboost.M1 is slightly better than the RF classifier (Alfaro et al., 2013), while no statistically significant differences ($q < 4.17$) were observed for scale groups 2 to 4 (see Table 1) between these two classifiers. Nevertheless, Adaboost.M1 performed worse than RF at a fine scale, and this can be attributed to the stability of RF with a large number of trees compared to Adaboost.M1 (Chan and Paelinckx, 2008). On the one hand, the increased objects are able to contribute more classification information for RF with a subsampling algorithm; on the other hand, the increasing objects, which generate increasingly similar features due to the similar spectra and small area, were more difficult to distinguish, since Adaboost.M1 always increased weights of misclassified samples, which would likely reduce the classified samples with distinct features. Furthermore, this is also why the accuracy of classification with feature selection was better than one without feature selection for Adaboost.M1, especially at a fine scale.

Although RF significantly outperformed SVM at all scales, there were apparently consistent trends of the overall accuracies decreasing with the segmentation scale for both classifiers; the metric's variances were increasingly larger and the metric's mean values were ever smaller with a decrease in scale. This is also in agreement with the assumption that mixed objects lead to a natural error of classification due to under-segmentation (Ma et al., 2015). SVM and RF were strongly capable of overcoming the effect of the broken objects at a fine scale compared to the other classifiers considered, where more segments with similar features were generated. In addition, our results indicate that the performance of SVM was consistently better compared to those obtained from the other single classifier, for example the DT and naiveBayes, and this was also in accordance with earlier findings (Duro et al., 2012). It can therefore be expected that SVM and RF should frequently be employed for fine scales due to their stable performance in regard to the scale.

The trend of the overall accuracy of DT changing along with the scale parameter in this study is in agreement with the findings of Laliberte and Rango (2009), further proving that DT is a very stable classification method for different segmentation scales in GEOBIA. Our results also indicate that DT cannot overcome the effect of the broken objects at a fine scale as indicated by the fact that the classification tree at a fine scale yielded worse results than those produced at a medium or coarse scale. Nonetheless, DT was more frequently used for lots of research (Mallinis et al., 2008; Peña-Barragán et al., 2011), and acceptable performance was reported. We assume that this is related to the estimation of the segmented scale parameter using expert based trial and error or automated (statistical) methods (i.e., ESP (Drăguţ et al., 2014)) to find the best suited scales for the classification. Due to the observed trend of DT performing significantly worse than the SVM at all scales coupled with the continuous improvement of the accuracy gap in terms of the scores between DT and SVM with increasing scale can be interpreted as evidence of DT's stability at an optimal (e.g. medium) scale. We therefore conclude that the DT classification performance is predisposed to be approximately close to that of the SVM at the '*optimal*'segmentation scale. Additionally, we also encourage the use of DT due to its good interpretability, although the results must be regarded with some caution.

Ideally, a good classifier for GEOBIA should: 1) follow the trend of the overall accuracy gradually declining with increasing scale; and 2) perform significantly better than other classifiers for all scales used. The hypothesis behind this is that with increased scale the accuracy metrics essentially decrease irrespective of which classifier is used and so called "mixed objects" appear. These mixed objects are basically results of under-segmentation and they are more likely to occur at coarser scales. The results (Fig. 2a and b, Fig. 3a and b) demonstrated the potential of the SVM and RF algorithms to overcome the effect of the broken objects at finer scales and confirm the hypothesis that the overall accuracies decline with increasing scale along with increasing instances of under-segmentation. Although SVM and RF performed better than other classifiers at fine scales, we need to point to a difficult discussion in GEOBIA concerning the fact that accuracy metrics would ultimately decline at the finest scales close to pixel level (Ma et al., 2015). Therefore, when referring to a "fine scale" this should still be distinctly coarser than the pixel level – otherwise GEOBIA methods do not make much sense (see discussion in Blaschke et al. (2014)).

4.2. Differences between selected and all features

Most studies in literature claim that classifications were able to benefit from feature selection (Van Coillie et al., 2007). In contrast to previous studies, our results highlight that a robust classification technique has to be capable of overcoming the difficulty of a significant difference in the accuracy measures occurring in the classifier between selected features and all features. Nevertheless, the results revealed that some classifiers showed significant accuracy differences for the general cases with or without feature selection, namely naiveBayes and adaboost.M1. In most cases, naiveBayes significantly benefited from feature selection, while adaboost.M1 was negatively affected: in these cases the classification error increased along with a decreasing numbers of features used. We assume that adaboost.M1 generally benefits from larger numbers of features due to the weighted samples. Consequently, no feature selection is suggested for adaboost.M1, while it seems to be necessary for naiveBayes. Finally, it should be noted that in general DT and RF were the most robust classification techniques with or without feature selection, since no significant differences between selected and all features were observed for these two classifiers, followed by SVM, which slightly benefits from feature selection.

4.3. Sensitivity analysis to training set size

Since training samples are in most cases costly and therefore a limitation factor, we analyzed the sensitivity to the respective training set size using absolute numbers of sampled objects instead of the sample ratio (Ma et al., 2015). Concerning the Nemenyi *post-hoc* test results, there is no significant difference between the classifiers for limited training samples. For limited training samples, RF's accuracy metrics fluctuated with higher variance and changed rapidly along with increasing sample sizes for both areas. We assumed that this is related to lack of training data to lead to the over-fitting problem (Verikas et al., 2011). It can be concluded that RF is sensitive to limited training samples and it greatly benefits from a larger sam-

ple size (Fassnacht et al., 2014). However, we may conclude that in limited training samples each of the best five classifiers can be used. For large training sample sizes, we recommend RF and adaboost.M1 which significantly outperform other classifiers.

4.4. Homogeneous and heterogeneous objects in GEOBIA

To date, no reports have analyzed the effect of the mixed objects (heterogeneous objects) for supervised classification techniques in GEOBIA, rather than simply labeling the objects. Radoux and Bogaert (2014) also paid more attention to interpreting how important the sampling unit of the polygon was, but did not specifically analyze how mixed objects – which will ultimately occur at too coarse segmentation scales – affect the accuracy. Concerning the ANCOVA results, accuracy performances appear to be good when homogeneous objects are used as training data or test data. In other words, the performance of classification accuracy was strongly affected by the number of mixed objects. Although mixed objects do not necessarily have dominant sub-object cover proportions, they still need to be labeled as only one cover class to provide a traditional hard label approach (Shao and Lunetta, 2012). Consequently, such types of classification errors occur when using area-based accuracy assessment techniques. For both study areas, further analysis also indicated that the change of accuracy measures for each classifier was consistent with the change in the number of mixed objects. This can be attributed to the decrease of mixed objects or the increase of dominant cover for each object as the overlay ratio increases. Furthermore, the similar slopes of the lines of each classifier (Figs. 4 and 5) suggest that there is an equal response to the mixed objects phenomenon between all classification techniques used. Although only scale 80 with an approximately optimal segmentation scale was tested, it can be predicted that the use of coarser scales will substantially contribute to statistical accuracy errors due to the increased probability of mixed objects.

5. Conclusions

In this study, we presented a systematic comparison of seven classification techniques used within GEOBIA workflows. Our primary interest was to analyze the behavior of each classifier at different segmentation scales for agricultural mapping, with and without feature selection, with different training set sizes and the existence of "mixed objects" (in reality consisting of different objects due to under-segmentation). The results indicated that the overall accuracy only changed linearly along with the segmentation scale change and the expected change of the resulting accuracies when applying the SVM and RF classification techniques. For both study areas, the RF classifier provided the highest overall accuracy at all of the scales used, especially when using larger training samples. In addition, DT and RF were the most stable classification techniques with and without feature selection, while all other classifiers were influenced, to a greater or lesser extent, by feature selection. We do not recommend using adaboost.M1 with feature selection, while feature selection poses no problem for naiveBayes. RF and adaboost.M1 show almost the same accuracy performance over all training set sizes and yield significantly better results than the other classifiers, while generally benefitting from larger sample sizes. Further analysis also led to the conclusion that the change of accuracy measures for each classifier was consistent with changes in the number of mixed objects. Overall, when generalizing and providing advice to GEOBIA users, we may conclude that the results of this comprehensive study suggest that the Random Forest classifier performed the best overall when using high spatial resolution imagery in GEOBIA. These findings can support many researchers in their decisions regarding which GEOBIA classification strategy to use for agricultural mapping. However, the suitability of these approaches should be investigated further for other medium or low resolution remote sensing image in agricultural environments. Additional studies with other land cover communities are also expected to determine if there are appeared these similar trends, for example, urban area.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant Nos. 41371017 and 41001238), the National Key Technology R&D Program of China (Grant Nos. 2012BAH28B02), the National Study Abroad Funding of China, the program A/B for Outstanding PhD candidate of Nanjing University and the National Study Abroad Funding of China. Sincere thanks are given for the comments and contributions of anonymous reviewers and members of the editorial team.

References

Alfaro, E., Gamez, M., Garcıa, N., 2013. adabag: An R package for classification with boosting and bagging. J. Stat. Softw. 54 (2), 1–35.
Baatz, M., Schaepe, A., 2000. Multiresolution Segmentation: an optimization approach for high quality multi-scale image segmentation. Angewandte Geographische Informationsverarbeitung XII. Beiträge zum AGIT-Symposium Salzburg 2000, pp. 12–23, Karlsruhe, Herbert Wichmann Verlag.
Benz, U.C., Hofmann, P., Willhauck, G., Lingenfelder, I., Heynen, M., 2004. Multi-resolution, object-oriented fuzzy analysis of remote sensing data for GIS-ready information. ISPRS. J. Photogram. Remote Sens. 58 (3/4), 239–258.
Blaschke, T., 2010. Object based image analysis for remote sensing. ISPRS J. Photogram. Remote Sens. 65 (1), 2–16.
Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Queiroz Feitosa, R., van der Meer, F., van der Werff, H., van Coillie, F., Tiede, D., 2014. Geographic Object-Based Image Analysis—towards a new paradigm. ISPRS J. Photogram. Remote Sens. 87, 180–191.
Breiman, L., 2001. Random Forests. Mach. Learn. 45 (1), 5–32.
Brenning, A., 2009. Benchmarking classifiers to optimally integrate terrain analysis and multispectral remote sensing in automatic rock glacier detection. Remote Sens. Environ. 113 (1), 239–247.
Chan, J.C.W., Paelinckx, D., 2008. Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. Remote Sens. Environ. 112 (6), 2999–3011.
Chang, C.C., Lin, C.J., 2011. LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol., 2011, Software available at. http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
Congalton, R.G., Green, K., 2009. Assessing the accuracy of remotely sensed data: Principles and practices, 2nd. CRC Press, Boca Raton, pp. 2009.
Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.
Drăguţ, L., Csillik, O., Eisank, C., Tiede, D., 2014. Automated parameterisation for multi-scale image segmentation on multiple layers. ISPRS J. Photogram. Remote Sens. 88, 119–127.
Dronova, I., Gong, P., Clinton, N.E., Wang, L., Fu, W., Qi, S., Liu, Y., 2012. Landscape analysis of wetland plant functional types: the effects of image segmentation scale, vegetation classes and classification methods. Remote Sens. Environ. 127, 357–369.
Duro, D.C., Franklin, S.E., Dubé, M.G., 2012. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5HRG imagery. Remote Sens. Environ. 118, 259–272.
Fassnacht, F.E., Hartig, F., Latifi, H., Berger, C., Hernández, J., Corvalán, P., Koch, B., 2014. Importance of sample size: data type and prediction method for remote sensing-based estimations of aboveground forest biomass. Remote Sens. Environ. 154, 102–114.
Foody, G.M., Mathur, A., Sanchez-Hernandez, C., Boyd, D.S., 2006. Training set size requirements for the classification of a specific class. Remote Sens. Environ. 104 (1), 1–14.
Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: Proceedings of the Thirteenth International Conference on Machine Learning. Morgan Kaufmann, pp. 148–156.
Ghosh, A., Joshi, P.K., 2014. A comparison of selected classification algorithms for mapping bamboo patches in lower Gangetic plains using very high resolution WorldView 2 imagery. Int. J. Appl. Earth Obs. Geoinf. 26, 298–311.
Hall, M.A., Holmes, B., 2003. Benchmarking attribute selection techniques for discrete class data mining. IEEE Trans. Knowl. Data Eng. 15 (6), 1–16.
Heumann, B.W., 2011. An object-based classification of mangroves using a hybrid decision tree—support vector machine approach. Remote Sens. 3 (11), 2440–2460.

Hsu, C.W., Chang, C.C., Lin, C.J., 2010. A practical guide to support vector classification. Department of Computer Science, National Taiwan University, Taipei 106, Taiwan, last updated: April 15 2010 http://www.csie.ntu.edu.tw/~cjlin/.

Laliberte, A.S., Koppa, J.S., Fredrickson, E.L., Rango, A., 2006. Comparison of nearest neighbor and rule-based decision tree classification in an object-oriented environment. In: IEEE international Geoscience and Remote Sensing Symposium Proceedings, Denver, Colorado.

Laliberte, A.S., Rango, A., 2009. Texture and scale in object-based analysis of subdecimeter resolution unmanned aerial vehicle (UAV) imagery. IEEE Trans. Geosci. Remote Sens. 47 (3), 761–770.

Liu, Y.X., Li, M.C., Mao, L., Xu, F., Huang, S., 2006. Review of remotely sensed imagery classification patterns based on object-oriented image analysis. Chin. Geog. Sci. 16 (3), 282–288.

Luque, I.F., Aguilar, F.J., Álvarez, M.F., Aguilar, M.Á., 2013. Non-parametric object-based approaches to carry out ISA classification from archival aerial orthoimages. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 6 (4), 2058–2071.

Ma, L., Li, M.C., Tong, L.H., Wang, Y.F., Cheng, L., 2013. Using unmanned aerial vehicle for remote sensing application. In: 2013 21st International Conference on Geoinformatics, Kaifeng, China, pp. pp. 1–5.

Ma, L., Cheng, L., Han, W.Q., Zhong, L.S., Li, M.C., 2014. Cultivated land information extraction from high-resolution unmanned aerial vehicle imagery data. J. Appl. Remote Sens. 8 (1), 836731–8367325.

Ma, L., Cheng, L., Li, M.C., Liu, Y.X., Ma, X.X., 2015. Training set size scale, and features in Geographic Object-Based Image Analysis of very high resolution unmanned aerial vehicle imagery. ISPRS J. Photogram. Remote Sens. 102, 14–27.

Mallinis, G., Koutsias, N., Tsakiri-Strati, M., Karteris, M., 2008. Object-based classification using quickbird imagery for delineating forest vegetation polygons in a Mediterranean test site. ISPRS J. Photogram. Remote Sens. 63 (2), 237–250.

Otukei, J.R., Blaschke, T., 2010. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. Int. J. Appl. Earth Obs. Geoinf. 12 (1), S27–S31.

Ouyang, Y., Ma, J., Dai, Q., 2006. Bayesian multi-net classifier for classification of remote sensing data. Int. J. Remote Sens. 27 (21), 4943–4961.

Peña-Barragán, J.M., Ngugi, M.K., Plant, R.E., Six, J., 2011. Object-based crop identification using multiple vegetation indices, textural features and crop phenology. Remote Sens. Environ. 115 (6), 1301–1316.

Pu, R., Landry, S., 2012. A comparative analysis of high spatial resolution IKONOS and WorldView-2 imagery for mapping urban tree species. Remote Sens. Environ. 124, 516–533.

Radoux, J., Bogaert, P., 2014. Accounting for the area of polygon sampling units for the prediction of primary accuracy assessment indices. Remote Sens. Environ. 142, 9–19.

Rogan, J., Franklin, J., Stow, D., Miller, J., Woodcock, C., Roberts, D., 2008. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. Remote Sens. Environ. 112 (5), 2272–2283.

Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogram. Remote Sens. 67, 93–104.

Shao, Y., Lunetta, R.S., 2012. Comparison of support vector machine neural network, and CART algorithms for the land-cover classification using limited training data points. ISPRS J. Photogram. Remote Sens. 70, 78–87.

Strasser, T., Lang, S., 2015. Object-based class modelling for multi-scale riparian forest habitat mapping. Int. J. Appl. Earth Obs. Geoinf. 37, 29–37.

Stumpf, A., Kerle, N., 2011. Object-oriented mapping of landslides using Random Forests. Remote Sens. Environ. 115 (10), 2564–2577.

Tehrany, M.S., Pradhan, B., Jebuv, M.N., 2014. A comparative assessment between object and pixel-based classification approaches for land use/land cover mapping using SPOT 5 imagery. Geocarto Int. 29, 351–369.

Tiede, D., Lang, S., Albrecht, F., Hölbling, D., 2010. Object-based class modeling for cadastre constrained delineation of geo-objects. Photogram. Eng. Remote Sens. 76 (2), 193–202.

Van Coillie, F., Verbeke, L., De Wulf, R.R., 2007. Feature selection by genetic algorithms in object-based classification of IKONOS imagery for forest mapping in Flanders. Belgium Remote Sens. Environ. 110 (4), 476–487.

Whiteside, T.G., Maier, S.W., Boggs, G.S., 2014. Area-based and location-based validation of classified image objects. Int. J. Appl. Earth Obs. Geoinf. 28, 117–130.

Witten, D.M., Tibshirani, R., 2011. Penalized classification using Fisher's linear discriminant. J. R. Stat. Soc. B 73 (5), 753–772.

Wulder, M.A., Coops, N.C., 2014. Make Earth observations open access. Nat 513, 30–31.

Xu, L.L., Li, J., Brenning, A., 2014. A comparative study of different classification techniques for marine oil spill identification using RADARSAT-1 imagery. Remote Sens. Environ. 141, 14–23.

Yan, G., Mas, J.F., Maathuis, B.H.P., Xiangmin, Z., Van Dijk, P.M., 2006. Comparison of pixel-based and object-oriented image classification approaches-A case study in a coal fire area, Wuda, Inner Mongolia, China. Int. J. Remote Sens. 27 (18), 4039–4055.

Yang, J., Wang, Y., 2012. Classification of 10m-resolution SPOT data using a combined Bayesian network classifier-shape adaptive neighborhood method. ISPRS J. Photogram. Remote Sens. 72, 36–45.

Yu, B., Ostland, I.M., Gong, P., Pu, R.L., 1999. Penalized discriminant analysis of in situ hyperspectral data for conifer species recognition. IEEE Trans. Geosci. Remote Sens. 37 (5), 2569–2577.

Yu, Q., Gong, P., Tian, Y.Q., Pu, R.L., Yang, J., 2008. Factors affecting spatial variation of classification uncertainty in an image object-based vegetation mapping. Photogram. Eng. Remote Sens. 74 (8), 1007–1018.

Zhang, X.L., Feng, X.Z., Xiao, P.F., He, G.J., Zhu, L.J., 2015. Segmentation quality evaluation using region-based precision and recall measures for remote sensing images. ISPRS J. Photogram. Remote Sens. 102, 73–84.