Remote Sensing of Environment 191 (2017) 156–167

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Mapping per-pixel predicted accuracy of classified remote sensing images

Reza Khatami [a], Giorgos Mountrakis [a,*], Stephen V. Stehman [b]

[a] Department of Environmental Resources Engineering, State University of New York, College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, United States
[b] Department of Forest and Natural Resources Management, State University of New York, College of Environmental Science and Forestry, 1 Forestry Drive, Syracuse, NY 13210, United States

ARTICLE INFO

Article history:
Received 8 June 2016
Received in revised form 11 January 2017
Accepted 21 January 2017
Available online xxxx

Keywords:
Land-cover mapping
Classification accuracy assessment
Accuracy map
Local accuracy
Image classification
AUC

ABSTRACT

The traditional approach of map accuracy assessment based on an error matrix does not capture the spatial variation in classification accuracy. Here, per-pixel accuracy prediction methods are proposed based on interpolating accuracy values from a test sample in which the reference class of each sampled pixel has been determined. Different accuracy prediction methods were developed based on four factors: predictive domain (spatial versus spectral), interpolation function (constant, linear, Gaussian, and logistic), incorporation of class information (interpolating each class separately versus grouping them together), and sample size. Developing accuracy maps using the spectral domain is a new approach in contrast to previous efforts based on the spatial domain. Performance of the prediction methods was evaluated using 26 test blocks, with 10 km × 10 km dimensions, dispersed throughout the United States. Each block had complete coverage reference data manually extracted by interpreters and a land-cover map produced from Landsat imagery using a decision tree classification. The full scene maps were then compared to the corresponding reference maps to produce complete coverage accuracy information for each block. The predicted accuracy maps were produced from a sample of the reference data (i.e., the test dataset). The performance of the sample-based accuracy predictions was evaluated using the area under the curve (AUC) of the receiver operating characteristic. Relative to existing accuracy prediction methods, our proposed methods resulted in improvements of AUC of 0.15 or greater. Evaluation of the four factors comprising the accuracy prediction methods demonstrated that: i) interpolations should be done separately for each class instead of grouping all classes together; ii) if an all-classes approach is used, the spectral domain will result in substantially greater AUC than the spatial domain; iii) for the smaller sample size and per-class predictions, the spectral and spatial domain yielded similar AUC; iv) for the larger sample size (i.e., very dense spatial sample) and per-class predictions, the spatial domain yielded larger AUC; v) increasing the sample size improved accuracy predictions with a greater benefit accruing to the spatial domain; and vi) the function used for interpolation had the smallest effect on AUC. To conclude, the ability to produce per-pixel accuracy predictions yielding simple to understand accuracy maps opens up new possibilities for error propagation of remotely sensed products in a variety of disciplines.

1. Introduction

Land-cover maps provide critical information to environmental studies. With the advancement of remote sensing science, production of land-cover maps through classification of remotely sensed imagery has become a common practice. Numerous studies have been conducted investigating a broad range of classification processes to produce more accurate classified land-cover maps (Gómez et al., 2016; Khatami et al., 2016). However, land-cover maps include misclassifications and these errors can propagate in models used to study environmental processes and affect the reliability of model predictions (Castilla and Hay, 2007; Ge et al., 2007; McMahon, 2007; Straatsma et al., 2013). End users of land-cover maps could benefit from additional information expressing map quality including the magnitude and the spatial distribution of misclassifications so users can incorporate this information in their decisions, modeling, and data fusion tasks (DeFries and Los, 1999; Gahegan and Ehlers, 2000; Miller et al., 2007).

The conventional way to report land-cover map accuracy is through an error matrix estimated from a test dataset which is independent from the classification training process. Summary measures such as overall accuracy and class-specific measures such as user's and producer's accuracies are commonly estimated from the error matrix. However, spatial variation is likely present in accuracy of maps constructed through image classification, as misclassifications tend to be spatially

* Corresponding author.
E-mail addresses: sgkhatam@syr.edu (R. Khatami), gmountrakis@esf.edu (G. Mountrakis), svstehma@syr.edu (S.V. Stehman).

http://dx.doi.org/10.1016/j.rse.2017.01.025
0034-4257/

autocorrelated (Campbell, 1981; Chen and Wei, 2009; Congalton, 1988). Consequently, the error matrix summary measures do not offer information on the spatial distribution of classification error and may not represent subregion or local misclassification rates when they differ from global rates (Mcgwire and Fisher, 2001). The spatial propagation of error is of great importance for modeling environmental processes and ideally a land-cover map should be accompanied by a map of spatial distribution of error (Foody, 2002; Comber et al., 2012). Such a map would provide map users with local estimates of accuracy. Consequently, in this article we focus on *per-pixel* accuracy, where *per-pixel* accuracy is defined as the probability of correct classification of a pixel, assessment of land-cover maps created by classifying remotely sensed imagery.

Different approaches have been proposed to characterize quality of land-cover maps at the local scale. Many researchers have used classification outputs to quantify classification certainty. One approach is to use the probabilities of class memberships or similar measures of strength of prediction as indicators of classification certainty (or uncertainty as its complement). Examples include posterior probabilities from maximum likelihood classifier (Brown et al., 2009; Canters, 1997; Foody et al., 1992; Ge et al., 2009; Maselli et al., 1994), activation levels from neural networks (Brown et al., 2009; Foody, 2000; Gong et al., 1996), soft outputs of support vector machines (Giacco et al., 2010; Löw et al., 2013), fuzzy c-means (Ge et al., 2009; Prasad and Arora, 2014; Wang and Shi, 2013), ARTMAP (Carpenter et al., 1999; Liu et al., 2004), decision tree and random forest (Liu et al., 2004; Loosvelt et al., 2012; Peters et al., 2009) and boosting methods (McIver and Friedl, 2001). Probability of membership of the most probable class has been used as an indicator of certainty (Colditz et al., 2011; Giacco et al., 2010; Loosvelt et al., 2012; Löw et al., 2013; McIver and Friedl, 2001). The general idea is that for a given pixel the greater the probability of class membership for a given labeled class, the greater the certainty associated with that class. In addition, functions of all or a subset of the membership values of classes have been used to construct certainty measures instead of using only the membership value of the most probable class. Examples of these functions include the difference between first and second largest class membership values (Prasad and Arora, 2014), Shannon's entropy (Dehghan and Ghassemian, 2006; Loosvelt et al., 2012; Maselli et al., 1994; Wang and Shi, 2013), and α-quadratic entropy (Giacco et al., 2010; Löw et al., 2013; Pal and Bezdek, 1994). Entropy summarizes the information from membership values of all classes.

Certainty measures, such as those mentioned above, provide information on the spatial distribution of classification quality. However, certainty and accuracy of a classification represent different concepts, even though the two terms are often used interchangeably (Mountrakis and Xi, 2013). Certainty can be viewed as the degree of confidence or conviction that the classifier has assigned the correct class label, whereas accuracy is defined based on agreement between the assigned label and the true ground condition. Certainty measures provide valuable information about the classifier's certainty in class assignment and can be used by an analyst to improve the training process. However, the end users of land-cover maps would be more interested in accuracy. Even though in many cases there is a correlation between accuracy and certainty, accuracy assessment using an independent test dataset cannot be replaced by certainty determined from the training dataset: "The relationship between uncertainty and accuracy may not be simple, with some cases possibly allocated correctly but uncertainly while others were allocated with little uncertainty but erroneously" (Foody, 2005).

Researchers have also utilized empirical models to link classification accuracy, as a dependent variable, to different independent (predictor) variables, such as landscape contextual measures, topographic variables, object membership, and land-cover class (Burnicki, 2011; Carmel, 2004; Smith et al., 2002; Smith et al., 2003; Van Oort et al., 2004; Yu et al., 2008). Logistic regression is most commonly used for these models because the dependent variable is dichotomous (i.e., a pixel is classified correctly or not).

Another approach for characterizing map quality at the local scale involves spatial interpolation of classification accuracy of the test dataset (Comber, 2013; Comber et al., 2012; Foody, 2005; Kyriakidis and Dungan, 2001; Steele et al., 1998; Tsutsumida and Comber, 2015). Foody (2005) created a grid of spatially constrained confusion matrices using the nearby test pixels of each grid point. Then overall accuracy and user's and producer's accuracies of each class from the local confusion matrices are interpolated using an inverse squared distance function. Similarly, Comber (2013), Comber et al. (2012), and Tsutsumida and Comber (2015) used geographically weighted logistic regression to interpolate overall accuracy and user's and producer's accuracies for each class. Steele et al. (1998) first estimated misclassification probabilities for training data using a bootstrap approach and then these probabilities were spatially interpolated for the entire map using kriging. Generally, the assumption of spatial interpolation is that pixels that are close spatially would have similar accuracy. However, one issue that may affect spatial interpolation of classification accuracy is that accuracy extracted from test data may not be representative of nearby unsampled pixels if these pixels belong to a different class. For example, consider a lake with a vegetated island in the middle. Classification accuracy of water pixels, even though spatially close to the vegetation pixels, may not be associated with the accuracy of the vegetation class. In other words, proximity in space can translate to an accuracy relationship only if pixels are from the same class.

The objective of this research is to evaluate several new per-pixel accuracy prediction methods and to contrast the performance of these new methods with previously suggested methods. Of particular interest is use of the spectral domain as the domain for accuracy prediction, which to the best of our knowledge has not been investigated before. We also evaluate major factors that affect local classification accuracy prediction, including the interpolation function (e.g. linear versus nonlinear), the predictive domain (spatial versus spectral), incorporation of class-specific information, and the effect of the test dataset sample size. The proposed accuracy maps have several intended applications. An obvious application is simply the use of the accuracy map to provide local accuracy information for a specific subregion of interest to a map user. An accuracy map can help to decide if the accuracy of a subregion of interest meets the required accuracy for the intended application (Comber, 2013). The accuracy maps can also be used to improve an existing classification, for example by prioritizing where additional training data or auxiliary data may be needed to improve classifier performance in regions of the map that have smaller accuracy (Foody, 2005). Accuracy maps can also be used to enhance modeling processes by providing local description of accuracy (Leyk et al., 2005; Seibert and McDonnell, 2010; Verburg et al., 2011; Verburg et al., 2009; Zhang et al., 2010). If land-cover maps are used as input data to models, the per-pixel accuracy values can be used in data assimilation and fusion and numerical modeling to quantify the accuracy of these input maps in a spatially explicit manner to facilitate quantification of uncertainties in final products and model predictions (Ampe et al., 2012; Dieye et al., 2012; Loosvelt et al., 2014; Quaife et al., 2008; Wegehenkel et al., 2006).

2. Datasets

Data from the United States Geological Survey (USGS) Land-Cover Trends project (Loveland et al., 2002) were used as reference data in this research. The entire Trends dataset contains land-cover data for 2688 sample blocks randomly selected within the 84 U.S. Environmental Protection Agency (EPA) Level III ecological regions. In our research 26 blocks selected from the year 2011 Trends dataset were used (Fig. 1). These 26 blocks had 30 m spatial resolution and covered a 10 km × 10 km (333 pixels × 333 pixels) area. These blocks were purposely selected to span a broad geographic range of the US and a range of land-cover area distributions.

The reference land-cover for each Trends block was obtained using manual interpretation of multiple sources of imagery and based on a

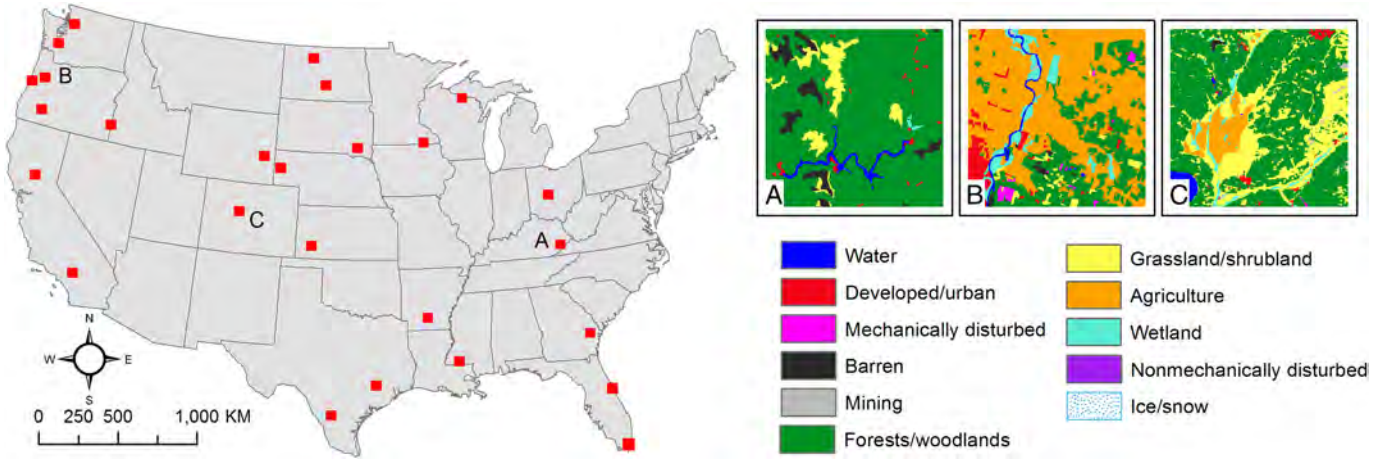

Fig. 1. Spatial distribution of the 26 Trends blocks. Actual block size was 10 km × 10 km, but blocks pictured are enlarged to enhance visualization.

modified Anderson (Anderson et al., 1976) Level I classification scheme that included the following 11 land-cover classes: water, developed/urban, mechanically disturbed, barren, mining, forests/woodlands, grassland/shrubland, agriculture, wetland, non-mechanically disturbed, and ice/snow (see http://landcovertrends.usgs.gov/main/classification.html for the actual definition of Trends classes, last accessed 2016). Landsat imagery was the primary source interpreted to determine the reference land-cover classification (Loveland et al., 2002; Sleeter et al., 2013) of each Trends block. For most blocks, interpreters also used aerial photographs, topographic maps, and Google Earth imagery to assist in the interpretation. All Trends blocks underwent a critique by the team of interpreters conducting the reference classification, and regions within the block could be re-classified following this team critique (Sleeter et al., 2013). While no reference dataset is without limitations, the use of many available sources of information to obtain the block land-cover classification and the extensive quality control process applied to each block means that the Trends data satisfy the criteria for reference data (GFOI, 2016, Section 5.1.5).

The land-cover maps to which the per-pixel accuracy prediction methods were applied were produced by a decision-tree classification of Landsat images (see Section 3.3). Landsat TM images from the 26 Trends blocks within the same year of the Trends data were used. The images were resampled to match with the Trends blocks. To remove any edge effect on resampled pixel values, one line of pixels was removed from all sides of the Landsat images and the Trends land-cover maps. During the classification process only the six reflective bands of Landsat images were used.

Throughout the subsequent text we will refer to the Trends data as "reference data" representing the best assessment of the ground condition of the pixel. Thus the Trends data represent complete coverage reference data for each block. Portion of these reference data are then used as: i)training data to develop the decision-tree classification of Landsat images that produces the land-cover map of each block and ii) as test data for accuracy assessment of these classifications (the training sample used to produce the land-cover map is always different from the accuracy assessment test sample). To distinguish the two uses of the Trends reference data, the phrase "training data" refers to data used to produce the classification that constitutes the land-cover map being assessed, and "test data" refers to data obtained to assess the accuracy of the map and to produce the per-pixel accuracy predictions. As noted, the training data and test data are obtained from different samples selected from each Trends block. The two steps of creating a land-cover map based on a sample of training data and assessing the accuracy of that map based on a sample of test data are standard elements of practice. Our evaluation of the performance of the accuracy prediction methods requires an additional step that is not present and in fact not possible in the conventional practice of accuracy assessment. This last step is to compare the predicted accuracy map to the actual accuracy map, where the actual ("true") accuracy map is defined by the agreement between the land-cover map produced by the decision-tree classifier and the census of reference data of each Trends block.

Our methods and analyses mimic the typical approach used in practice in that a land-cover classification is developed based on training data (that may contain classification errors) followed by an accuracy assessment that is conducted using independent test data (that also may contain classification errors). The predicted accuracy maps we produce from such an analysis are representations of accuracy based on agreement defined by comparison of the map to the reference classification. We should note that this, like any, accuracy assessment is dependent on the quality of the reference data. While we are confident that the Trends dataset is a reasonable depiction of the ground following the "good practices" identified in Olofsson et al. (2014) and GFOI (2016), readers should understand that no perfect reference dataset exists and that our work does not necessarily represent accuracy based on agreement with a gold-standard of "ground truth".

3. Methods

3.1. Factors of proposed accuracy interpolation methods

We examined the effect of four major factors in the performance of accuracy interpolation. These factors included *Predictive Domain*, *Interpolation Function*, *Class Incorporation*, and *Sample Size*. Detailed descriptions of these factors follow.

3.1.1. Predictive Domain

Two general predictive domains were tested, the *spatial* and *spectral* domains. The rationale for using the *spatial* domain was that pixels in close spatial proximity will exhibit similar accuracy. The use of the *spectral* domain for accuracy interpolation has not been tested before. The rationale for using the *spectral* domain was that classifiers operate in the *spectral* and not the *spatial* domain. Misclassifications typically occur at the class borders in the *spectral* domain. Furthermore, two pixels may be close spatially but have very different spectral signatures. While our testing used exclusively Landsat bands and we refer to this domain as *spectral*, in essence it is the input data predictive domain. For example, if texture or ancillary data are inputs to the classifier they would be part of this predictive domain.

3.1.2. Interpolation Function

Each pixel in the test dataset was assigned a binary value, 0 if misclassified and 1 if correctly classified. An interpolation process was necessary to propagate these test values to the entire image or the unsampled pixels, where an "unsampled pixel" is defined as any pixel

that is not included in the test dataset. Two general approaches to perform interpolation and/or quantify and model spatial autocorrelation are stochastic methods, for example kriging, and deterministic methods, such as kernel functions (Myers, 1994). Kernel functions have been shown to yield predictions comparable in quality to the stochastic methods (Scheuerer et al., 2013) and due to their simplicity they were selected in our work.

Four per-pixel *Interpolation Functions* were implemented including three kernel-based predictors and one surface fitting function. For the kernel-based predictors, the contribution of each neighbor test pixel in the accuracy prediction of a given unsampled pixel was determined based on its distance as determined by a weighting function. Then, the classification accuracy of the unsampled pixels was predicted by the weighted average of the binary classification values of neighboring test data. For all *Interpolation Functions* distances were calculated as Euclidean distance. The number N of the nearest test pixels used in the interpolation process expresses the degree of localization. A very small N translates to high degree of localization and a very large N can result in obtaining general map scale accuracy predictions similar to the global measures derived from the error matrix. An optimization process was used to determine the optimal N (see Section 3.3, Step 2). The four *Interpolation Functions* were the following:

(1) *Constant* kernel: The constant kernel assigns the same weight to all N of the nearest neighboring test pixels of the unsampled pixel.
(2) *Linear* kernel: This kernel uses a linear function of distance to assign weights to the nearby test pixels. The kernel is adaptive and formed for each unsampled pixel independently. The weights are computed based on the distances *h* from the unsampled pixel using Eq. (1) where *h_max* is the maximum distance between the unsampled pixel and its N nearest test pixels. In practice, the maximum distance is multiplied by 1.001 to avoid obtaining zero weights when all the nearest test pixels have the same distance to the unsampled pixel.

$$W_{Linear}(h) = 1 - \frac{h}{h\_max} \quad (1)$$

Gaussian kernel: *Gaussian* weights are calculated using Eq. (2) where *h* is the distance, constant *c* is equal to 0.1 and *h_max* is the maximum distance between the unsampled pixel and its N nearest test pixels.

$$W_{Gaussian}(h) = e^{-\frac{h^2}{c*h\_max}} \quad (2)$$

The 1.001 and 0.1 constants for *linear* and *Gaussian* kernels control the falling off speed of kernels and are kept fixed for all predictions. However, the shape of these kernels is dependent on the *h_max* value, therefore it is dynamically calculated for each unsampled pixel.

(3) *Logistic* regression: For each unsampled pixel, this method uses a separate predicted accuracy surface estimated via logistic regression applied to the N nearest test pixels around the unsampled pixel. Independent variables of the logistic model are coordinates in the predictive domain used for interpolation (either *spatial* or *spectral*) and the dependent variable is the binary accuracy value of each test pixel. The value of the fitted surface at the location of the unsampled pixel is the predicted classification accuracy.

3.1.3. Class Incorporation

The third component of the accuracy prediction process identified whether a single (*per-class*) or all classes (*all-classes*) should be used in the interpolation process. In *per-class* predictions, only the test data with the same classification label as the unsampled pixel were used to predict its accuracy. In *all-classes* interpolations, the test data from all classes contributed to the prediction.

3.1.4. Sample Size

Two test dataset *sample sizes* based on 0.5% and 2.5% of pixels per class were implemented to examine if *Sample Size* affects the performance of the interpolations. Each Trends block contains 109,561 pixels, so the 0.5% and 2.5% *Sample Sizes* translate to approximately 548 and 2740 test pixels respectively, with minor variations due to rounding.

Table 1 lists all accuracy prediction methods tested in this article. Sixteen methods (methods 1–16 in Table 1) were implemented based on all combinations of the *Predictive Domain*, (two options), *Interpolation Function* (four options), and *Class Incorporation* (two options). In addition, the proposed accuracy predictors were compared to three benchmark predictors with all predictions constructed from the same test sample data. The first benchmark, the Overall Accuracy (*OA*), was used as the predicted accuracy of all pixels (method 17). The second benchmark, the User's Accuracy (*UA*) of the assigned class (classified label), was used as the predicted accuracy of an unsampled pixel (method 18). Note that in practical applications, the reference label of an unsampled pixel is unknown and consequently it is not possible to model producer's accuracy. A third and final benchmark (method 19) was implemented using the concept of Spatially Constrained Confusion Matrices (*SCCM*) based on the work by (Foody, 2005). For the *SCCM* method, a 7×7 equidistant grid was created for each of the 26 blocks. Then for each grid point, the local overall accuracy was predicted using the spatially nearest 150 test sample pixels. Finally, Inverse Distance Weighting (IDW) spatial interpolation was used to interpolate the overall accuracies of the grid points to predict the accuracy of all map pixels.

3.2. Evaluating accuracy predictions using area under the receiver operator characteristic curve

The area under the receiver operating characteristic curve (AUC) (Bradley, 1997; Mas et al., 2013) was used to evaluate the accuracy predictions. AUC is a commonly used measure for assessing the performance of models constructed to predict binary outcomes. AUC can be interpreted as the probability that a greater predicted score is assigned to a randomly chosen positive case (e.g., correctly classified pixel) than to a randomly chosen negative case. To calculate AUC, model predictions for the test data are first discretized to 0 and 1 values based on multiple thresholds ranging from 0 to 1. For each threshold value, an error matrix is constructed by comparing the discretized model

Table 1
Per-pixel classification accuracy prediction methods evaluated.

No	Method abbreviation	Predictive domain	Interpolation function	Class incorporation
1	SpecConPer	Spectral	Constant kernel	Per-class
2	SpecLinPer	Spectral	Linear kernel	Per-class
3	SpecGauPer	Spectral	Gaussian kernel	Per-class
4	SpecLogPer	Spectral	Logistic regression	Per-class
5	SpatConPer	Spatial	Constant kernel	Per-class
6	SpatLinPer	Spatial	Linear kernel	Per-class
7	SpatGauPer	Spatial	Gaussian kernel	Per-class
8	SpatLogPer	Spatial	Logistic regression	Per-class
9	SpecConAll	Spectral	Constant kernel	All-classes
10	SpecLinAll	Spectral	Linear kernel	All-classes
11	SpecGauAll	Spectral	Gaussian kernel	All-classes
12	SpecLogAll	Spectral	Logistic regression	All-classes
13	SpatConAll	Spatial	Constant kernel	All-classes
14	SpatLinAll	Spatial	Linear kernel	All-classes
15	SpatGauAll	Spatial	Gaussian kernel	All-classes
16	SpatLogAll	Spatial	Logistic regression	All-classes
17	OA	–	Overall accuracy	All-classes
18	UA	–	User's accuracy	Per-class
19	SCCM	Spatial	SCCM and IDW	All-classes

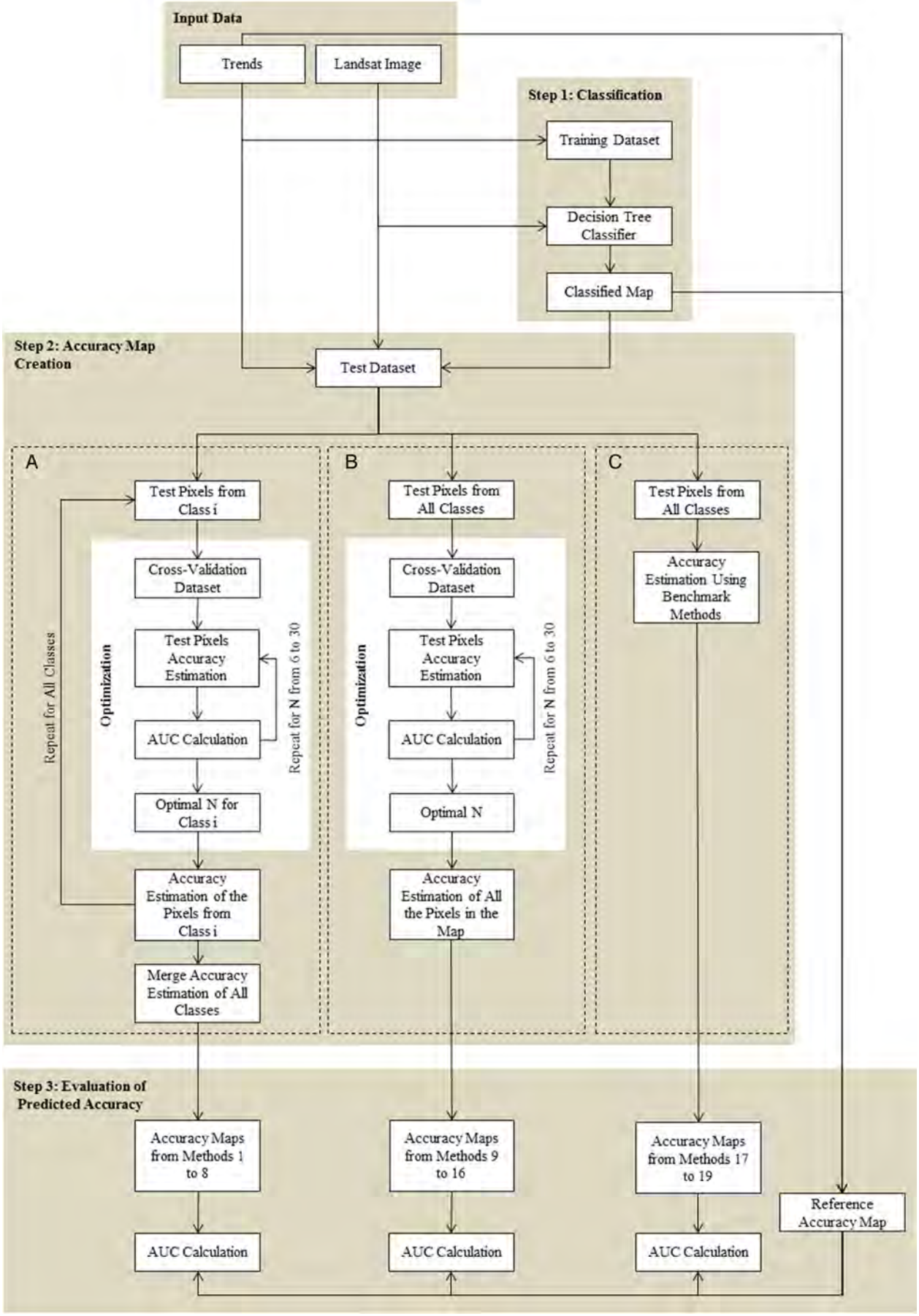

Fig. 2. The flowchart of image classification and accuracy map creation. Box A, B, and C show the accuracy map production process for methods 1 to 8 (*per-class* interpolations), 9 to 16 (*all-classes* interpolations), and benchmark methods (17 to 19).

predictions with the reference binary values. The true and false positive rates are calculated from each error matrix. Then, a graph of the true positive rate versus the false positive rate for all threshold values forms a curve, and the area under this curve is AUC. AUC values can theoretically range from 0 to 1 with larger values indicating better prediction of accuracy. If the model has good predictive performance (i.e., larger predicted values are assigned to the correctly classified pixels and smaller predicted values are assigned to the misclassified pixels), the curve would rise sharply and AUC would be close to 1. For a random model or a constant model, the true and false positive rates increase at almost the same rate and AUC would be close to 0.5 (for a constant model it would be exactly 0.5). AUC values <0.5 indicate that predictions are worse than random guessing. Thus, the practical range of AUC values is usually between 0.5 and 1.

3.3. Experimental design

Our experiment consisted of three general steps (Fig. 2). The first two steps represent the process that would be implemented in a practical application, which is creation of the land-cover map followed by implementation of an accuracy assessment and per-pixel prediction of accuracy from the test sample data. The land-cover map for each block was produced by classifying the corresponding Landsat image using a decision tree classifier applied to a training sample. Second, an independent test sample (i.e., independent from the training sample) was selected and the data from the test sample were used for all 19 accuracy prediction methods (Table 1) and an accuracy map was created for each method (Fig. 2). The third step in the experiment was to evaluate the performance of the accuracy predictions of the different methods. Each predicted accuracy map was compared to the "reference" accuracy map (i.e., the census of reference data available for each Trends block). The entire process was implemented independently for each of the 26 blocks. The three general steps are elaborated upon in the following.

Step 1 (Classification to Produce the Land-Cover Map):

- The training data input to the classifier was selected by stratified random sampling. Two percent of the pixels from each class were randomly selected from each Trends block as training data for that block.
- For each block the corresponding Landsat image was classified by a decision tree classifier to produce a land-cover map. Other classification processes could have been used; however, the focus of this research was on accuracy assessment rather than optimization or evaluation of the classifier. All per-pixel accuracy prediction methods for that block used this single classified product.

Step 2 (Accuracy Map Creation):

- The test sample dataset for each block was selected using stratified random sampling with the classification labels as strata. The same test dataset was used by all 19 accuracy prediction methods so comparisons among different methods were not confounded by differences in test datasets. Although these test data could be used to estimate an error matrix using the appropriate stratified estimators (e.g., Olofsson et al., 2014), our focus is on use of these data to create the complete coverage maps of predicted accuracy.

Boxes A, B, and C in Fig. 2 show the accuracy map production process for methods 1 to 8 (*per-class* interpolations), 9 to 16 (*all-classes* interpolations), and the three benchmark methods (17 to 19). The steps in the *per-class* interpolation process for methods 1 to 8 (box A in Fig. 2) were as follows:

- Only the test pixels from each map class were used to predict the classification accuracy of the unsampled pixels with the same classified label on the map.
- The optimal number N of nearest test pixels that were used to produce the accuracy predictions was determined by an optimization process. The optimal N was determined separately for each class and method using only the test pixels from the same class. To do so, ten-fold cross-validation was used to evaluate predictive performance for different choices of N. The test pixels from a given class were randomly partitioned to non-overlapping subsets (the same 10 subsets were used by methods 1 to 8).

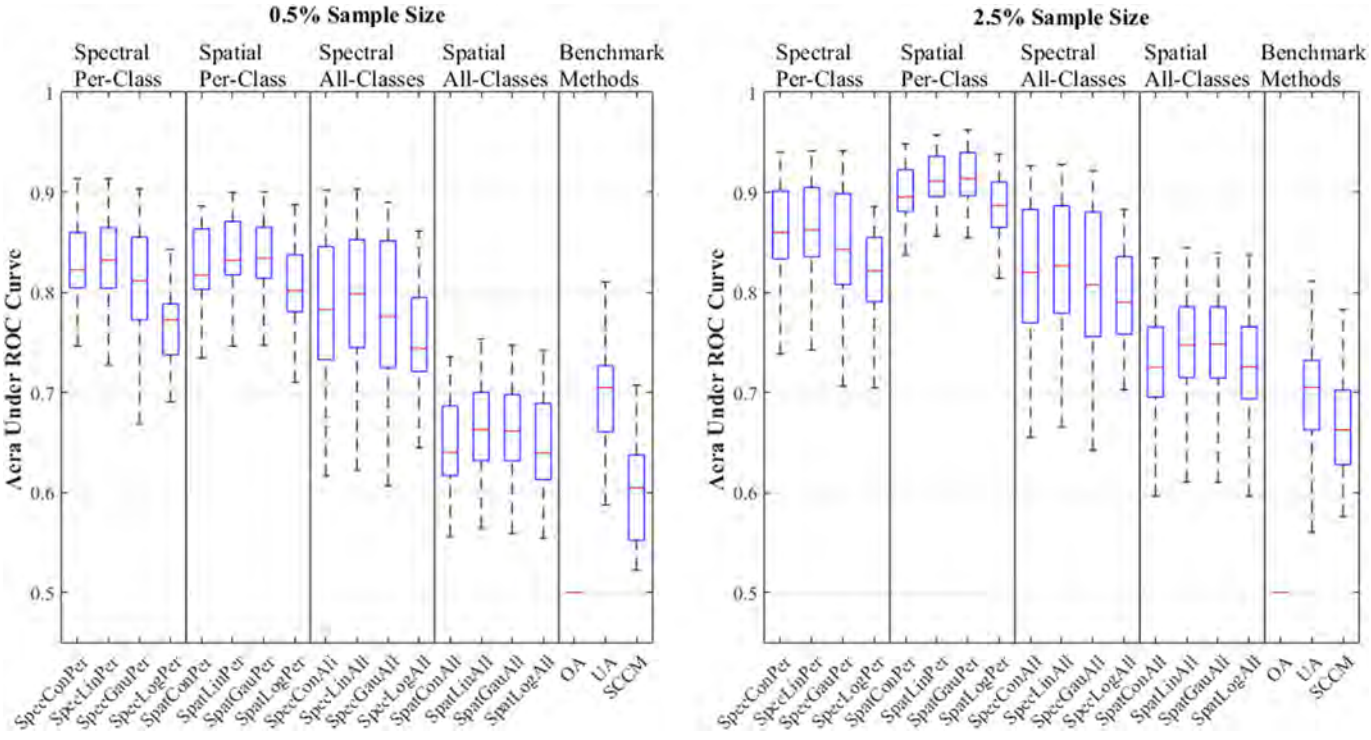

Fig. 3. Box-plots of AUC values of per-pixel accuracy prediction methods (larger AUC indicates more accurate prediction). Method abbreviations in this figure correspond to Table 1.

- For each of the cross-validation subsets, the data from the other 9 subsets were used to predict the accuracy of pixels in the target subset using a given prediction method and number of nearest neighbors.
- After predicting the accuracy of all test sample pixels of the given class, these predictions were compared to their corresponding reference accuracy values. AUC was used to evaluate the accuracy predictions of the cross-validation holdout sample.
- The search for the optimum number of neighbors N was initialized at 6 to have enough cases to fit the logistic models and terminated at 30 to constrain computing time. AUC was then calculated for each set of predictions for nearest neighbors from 6 to 30. If the number of test pixels in the 9 cross-validation subsets was smaller than 30 for a class, the optimization proceeded up to the smallest number of pixels among all 9 subsets. Also, if for a class the number of available test pixels was fewer than 6, then the average accuracy of the available test pixels was used as the accuracy prediction for all map pixels from that class.
- The number of neighbors that resulted in the largest AUC was selected as the optimum number of nearest test pixels (optimal N) for the given class and method.
- The sample test pixels from the target class, the given method, and the optimal N were used to predict the accuracy of all pixels mapped as the target class.
- The process of optimal N selection and accuracy prediction was repeated independently for each class using the given method.
- Finally, an accuracy map for the given method was created by merging predictions for all classes.

The process of accuracy map creation was repeated for all methods 1 to 8 using the same test dataset.

Box B in Fig. 2 shows the accuracy map production process for the *all-classes* interpolation methods 9 to 16. The same sample test dataset was used for these methods as in methods 1 through 8. However, for these *all-classes* methods the process was not partitioned by land-cover class. Consequently, all the test pixels regardless of their classified label contributed to accuracy prediction of unsampled pixels. For each of methods 9 to 16 all test pixels were used in the optimization process and to determine a single number as the optimal N for that method (in contrast to determining an optimum N for each class in the *per-class* method). Then, the classification accuracy map was constructed by the given method using the entire test dataset and the optimal N.

Box C in Fig. 2 describes the process for the three benchmark methods 17, 18 and 19. The same sample test dataset as methods 1–16 was used to create these three classification accuracy maps. First, an error matrix was constructed using the test dataset. For method 17, the estimated overall accuracy from the error matrix was used as

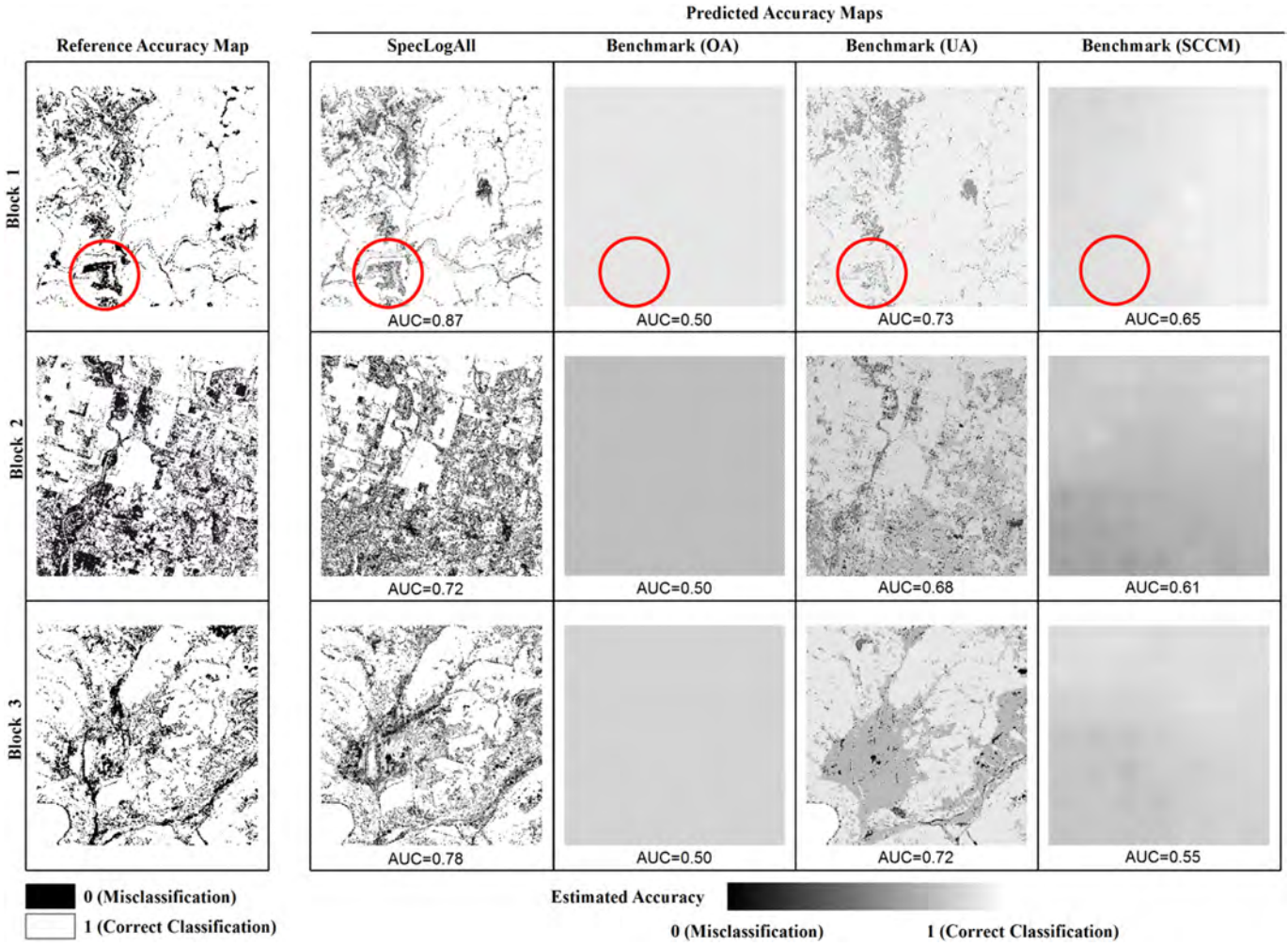

Fig. 4. Reference and predicted accuracy maps from the *SpecLogAll* (one of the least accurate new methods) and benchmark methods for three blocks (0.5% *Sample Size*). The region delineated by the red circle shows an example where the proposed new method (*SpecLogAll*) yields more accurate prediction of accuracy than the benchmark methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
ANOVA evaluating effect of four factors of per-pixel accuracy prediction methods on AUC.

Source of variation	Numerator degree of freedom	Denominator degree of freedom DF	Sum of squares	*p*-Value
Block	25	25	1.696	
Sample Size	1	25	0.738	<0.001
Whole-plot error	25		0.011	
Interpolation Function	3	750	0.136	<0.001
Predictive Domain	1	750	0.217	<0.001
Class Incorporation	1	750	2.093	<0.001
Interpolation Function * Predictive Domain	3	750	0.032	<0.001
Interpolation Function * Class Incorporation	3	750	0.009	0.03
Interpolation Function * Sample Size	3	750	0.002	0.62
Predictive Domain * Class Incorporation	1	750	0.860	<0.001
Predictive Domain * Sample Size	1	750	0.083	<0.001
Class Incorporation * Sample Size	1	750	0.000	0.71
Interpolation Function * Predictive Domain * Class Incorporation	3	750	0.000	0.95
Interpolation Function * Predictive Domain * Sample Size	3	750	0.001	0.71
Interpolation Function * Class Incorporation * Sample Size	3	750	0.001	0.79
Predictive Domain * Class Incorporation * Sample Size	1	750	0.001	0.32
Interpolation Function * Predictive Domain * Class Incorporation * Sample Size	3	750	0.000	0.96
Sub-plot error	750		0.746	

"*" denotes interaction between factors.

predicted accuracy of all pixels. The overall accuracy was estimated using stratified estimators because the test data were sampled using stratified random sampling (Olofsson et al., 2014). For method 18, the predicted accuracy was assigned the user's accuracy of the class to which the pixel was labeled. For method 19, a spatial interpolation process based on the concept of Spatially Constrained Confusion Matrices (*SCCM*) described in Section 3.1 was used to propagate the binary classification results from the test dataset to the entire map.

Step 3 (Evaluation of Predicted Accuracy):

- The reference accuracy map with binary values of 0 for misclassification and 1 for correct classification was created by comparing the land-cover product from step 1 to the reference map of each block.
- AUC was used to quantify the performance of the 19 accuracy prediction methods. Note that here an AUC value was calculated for the entire accuracy map (block) and this value was different from the AUC calculation used in Step 2 to determine the optimal number of nearest neighbors for interpolations based on the test dataset.

The classification and per-pixel accuracy predictions were applied to all 26 Trends blocks independently. Two test dataset *Sample Sizes* based on 0.5% and 2.5% of pixels per class were implemented. To account for variability in performance of prediction methods due to sampling variability, 10 independent test datasets were selected from each Trends block. Therefore, each map of a block (from step 1) was assessed for each of the 10 test sample datasets producing 10 accuracy maps for each accuracy prediction method. That is, after the land-cover map was produced for each block, steps 2 and 3 were repeated for two *Sample Sizes* and 10 times for each *Sample Size*. Box-plots of AUC values were used to summarize the performance of the different per-pixel accuracy prediction methods for the 26 Trends blocks. The AUC values used in the plots are the mean AUC from the 10 predicted accuracy maps based on the 10 independent test samples selected for each block (see appendix tables S1 and S2 for detailed AUC values).

Analysis of variance (ANOVA) was used to evaluate how the four different factors of the accuracy prediction methods affected AUC. The mean AUC value over the 10 test samples was used as the response variable. The ANOVA applied to the AUC data for methods 1 to 16 was a four-way factorial with *Predictive Domain*, *Interpolation Function*, *Class Incorporation*, and *Sample Size* as treatment factors with two levels of each factor except *Interpolation Function* that had four levels. The ANOVA was implemented for a split-plot experiment design with the *Sample Size* factor as the whole-plot treatment factor. *Predictive Domain*, *Interpolation Function*, and *Class Incorporation* were sub-plot treatment factors because these three factors were varied within each replication of a level of *Sample Size*. Each Trends block was regarded as a blocking factor at the whole-plot level of analysis. The AUC value for each block and each realization of the test sample selection was based on the census of reference data available for the block. ANOVA tables including the main effect and interaction tests are reported. Main effect tests indicate if there is statistically significant difference among levels of a given factor averaged over all levels of the other factors. Interaction tests indicate if the effect of a subset of factors depends on the levels of the other factors.

Table 3
Mean AUC values for combinations of factors affecting accuracy predictions.

(a)		Class Incorporation: Per-class	Class Incorporation: All-classes
Predictive Domain	Spectral	0.83[a]	0.79
	Spatial	0.86	0.70

(b)		Sample Size (%): 0.5	Sample Size (%): 2.5
Predictive Domain	Spectral	0.79	0.83
	Spatial	0.74	0.82

(c)		Interpolation Function: Constant	Linear	Gaussian	Logistic
Predictive Domain	Spectral	0.82	0.83	0.81	0.78
	Spatial	0.77	0.79	0.79	0.77

(d)		Interpolation Function: Constant	Linear	Gaussian	Logistic
Class Incorporation	Per-Class	0.85	0.86	0.85	0.82
	All-Classes	0.74	0.76	0.75	0.73

(e)		Sample Size (%): 0.5	Sample Size (%): 2.5
Class Incorporation	Per-class	0.82	0.87
	All-classes	0.71	0.77

(f)		Interpolation Function: Constant	Linear	Gaussian	Logistic
Sample Size (%)	0.50	0.77	0.78	0.77	0.74
	2.50	0.83	0.84	0.83	0.81

[a] The standard errors of the differences between all pairs of means were small ranging from 0.003 to 0.004.

4. Results

4.1. Comparison of new prediction methods with benchmark methods

The proposed new accuracy prediction methods generally outperformed the three benchmark methods as AUC values were generally greater for the new methods (Fig. 3). The *UA* benchmark had the largest AUC among the three benchmarks followed by *SCCM*. Overall accuracy (*OA*) had the smallest AUC among the benchmarks. As mentioned before, the AUC value for a constant model such as *OA* is 0.5. Accordingly, the AUC value for *OA* (method 17) was 0.5 as it assigned the same OA value as the accuracy value for every pixel. For the 2.5% *Sample Size*, all 16 proposed new accuracy prediction methods produced more accurate predictions than the benchmarks (Fig. 3). For the smaller *Sample Size* of 0.5%, the *UA* benchmark method produced more accurate predictions (greater AUC values) than the four *spatial all-classes* methods which had the smallest AUC values of the new prediction methods.

To visually illustrate the distinction between the accuracy maps produced by the new methods and maps generated from the three benchmark methods, we display the accuracy prediction maps generated by one of the least accurate new methods, *SpecLogAll* (*spectral_logistic_all-classes*). The spatial patterns of predicted accuracy produced by the *SpecLogAll* maps are much more similar to the patterns of the reference accuracy maps than are the predicted accuracy maps of the three benchmark methods (Fig. 4). *OA* maps are constant providing no information on spatial variation of accuracy. *UA* maps have constant values for each class and while the *UA* maps provide much more spatial detail than the *OA* and *SCCM* maps, the lack of information on within-class variation of accuracy fails to capture much of the true variation in accuracy. Lastly, the *SCCM* maps also show limited information on spatial variation of accuracy because they aggregate values from different classes. For example, in the lower left corner of Block 1 (region highlighted in red in Fig. 4), the reference accuracy map shows a region with very low accuracy (dark area) surrounded by an area of high accuracy. This pattern is clearly shown in the *SpecLogAll* accuracy map but it is not identified as clearly in the benchmark maps, with the *UA* predicted map being the best of the benchmarks at detecting this pattern. Moreover, the paired *t*-tests comparing mean AUC of the new prediction methods to each of the three benchmark methods were statistically significant (p-value < 0.001 for both *Sample Sizes*) confirming the advantage of *SpecLogAll* method and consequently all methods 1 through 12 over the benchmark methods.

4.2. Comparison of proposed new methods for predicting accuracy

In the comparison among methods, several general tendencies were evident from Fig. 3: 1) larger *Sample Size* improved the AUC values of most methods; 2) *per-class* accuracy prediction methods had greater AUC than *all-classes* methods; 3) the *spectral* domain outperformed the *spatial* domain for the *all-classes* methods; 4) the *spatial* domain had slightly greater AUC than the *spectral* domain for *per-class* methods

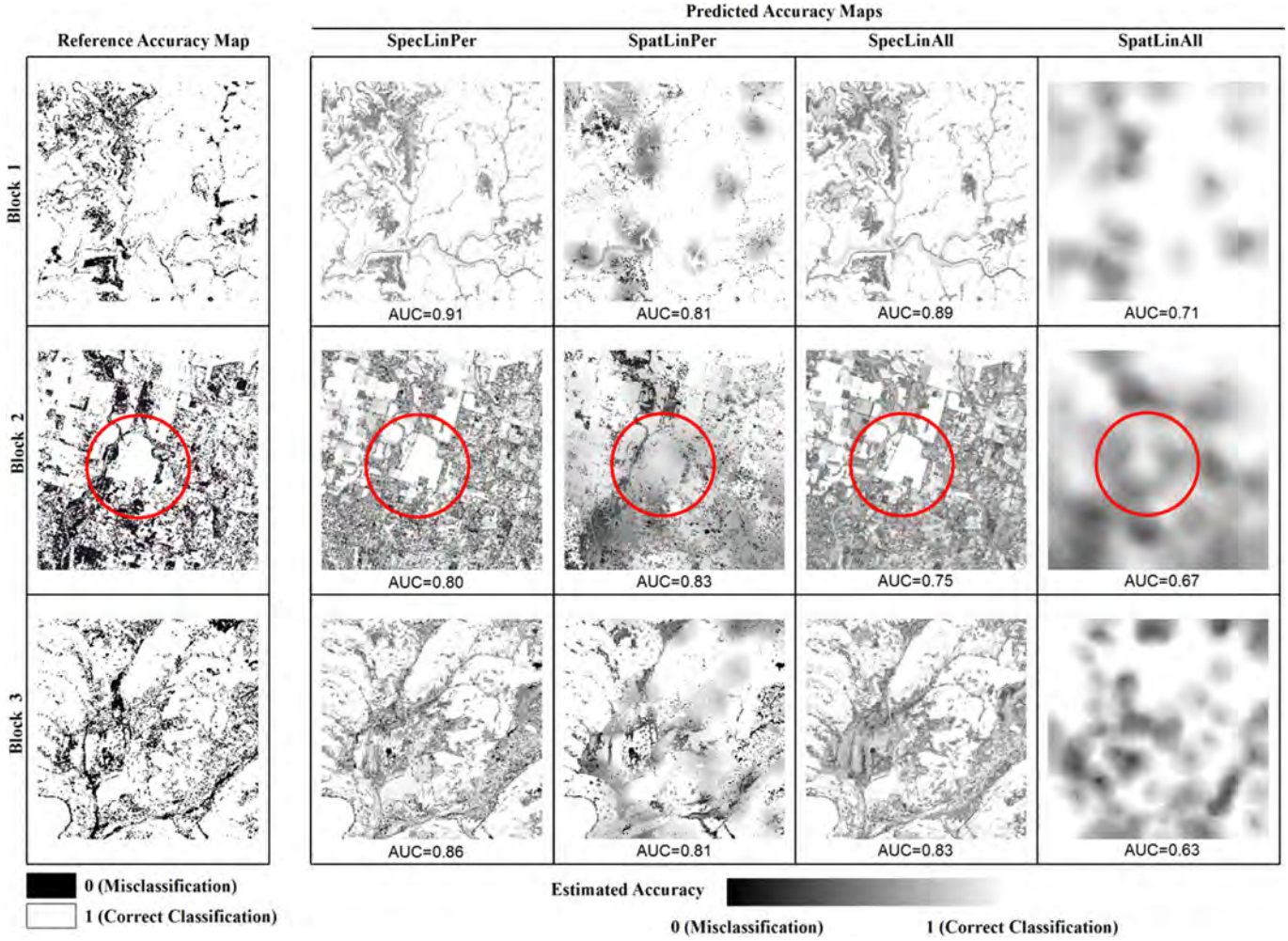

Fig. 5. Reference and predicted accuracy maps for three sample blocks using a *linear* interpolation and a *Sample Size* of 0.5%. The region delineated by the red circle shows an example where the *spectral* methods produced more accurate predictions than the *spatial* methods. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for the larger *Sample Size*, but there was little difference between *spectral* and *spatial* methods at the smaller *Sample Size*; and 5) the differences among the four *Interpolation Functions* were minor compared to other factors with the logistic function showing the worst predictive performance.

The impact of the different features of the accuracy prediction methods on AUC was further evaluated through the analysis of variance of the factorial design. The largest impacts were observed for *Sample Size*, *Class Incorporation*, and the interaction between *Predictive Domain* and *Class Incorporation* (Table 2). The four-way interaction (p-value = 0.96) and all three-way interactions (p-value > 0.30) were not statistically significant so none of the two-way interactions depended on the levels of the other factors (Table 2) and we can proceed to interpret two-way interactions. Four of the six two-way interactions were statistically significant (p-value <0.05) with the two exceptions being the interaction between *Sample Size* and *Interpolation Function* (p-value = 0.62) and the interaction between *Class Incorporation* and *Sample Size* (p-value = 0.71).

We further compared mean AUC for the different factors based on these interaction patterns. That is, because the three- and four-factor interactions were not significant, each two-factor table provided in Table 3 displays means averaged over the two factors not present (e.g., mean AUC values in Table 3a are reported as a two-way table by *Predictive Domain* and *Class Incorporation* averaged over *Sample Size* and *Interpolation Function*). Two-factor interactions were significant necessitating reporting the simple effect means presented in Table 3. The following statements summarize the effect of each of the four factors on AUC:

1) *Class Incorporation*: *Per-class* interpolations outperformed *all-classes* interpolations. *Per-class* interpolation increased AUC approximately 0.10 with respect to *all-classes* interpolation for all *Interpolation Functions* and *Sample Sizes* (Table 3d and e). The effect of *Class Incorporation* was substantially less for the *spectral* versus the *spatial* approach. Relative to *all-classes* interpolation *per-class* interpolation increased AUC by 0.04 for *spectral* interpolation versus an increase of 0.16 for *spatial* interpolation (Table 3a). This observation can be attributed to the ability of *spectral* domain to separate pixels from different classes. As a result, even when interpolation was performed for all classes together, mostly pixels from the same class contributed to the accuracy prediction of an unsampled pixel if the *spectral* domain was used. However, in the *spatial* domain the *all-classes* interpolation often resulted in confusion of accuracy of pixels from different classes while predicting the accuracy of a given pixel and there was consequently a large benefit from *per-class* interpolation compared to *all-classes* interpolation in this domain. Generally, *Class Incorporation* had the largest effect among all four factors (Table 3).

2) *Predictive Domain*: *Spectral* interpolation outperformed *spatial* interpolation by a mean of 0.09 for AUC when interpolating all classes together but *spatial* interpolation produced more accurate predictions than *spectral* interpolation by a mean of 0.03 when interpolating classes separately (Table 3a). Also, *spectral* interpolation outperformed *spatial* interpolation for both *Sample Sizes* but this advantage of *spectral* domain was not present for the larger *Sample Size* (Table 3b). Finally, *spectral* interpolation yielded greater AUC than *spatial* interpolation for different *Interpolation Functions* with the largest increase in mean AUC for *spectral* occurring for *constant* and *linear* kernels (Table 3c).

3) *Sample Size*: Larger *Sample Size* improved AUC for both *spectral* and *spatial* interpolations, but it had a larger effect on *spatial* interpolation (Table 3b). In other words, *spectral* interpolation would be less affected than *spatial* interpolation if *Sample Size* decreases from 2.5% to 0.5% (decrease in mean AUC of 0.04 for *spectral* interpolation versus a 0.08 decrease for *spatial* interpolation). Finally, increasing *Sample Size* from 0.5% to 2.5% increased AUC by about 0.06 for both *per-class* and *all-classes* interpolation and for all *Interpolation Methods* (Table 3e and f).

4) *Interpolation Function*: This factor had the smallest effect among the four factors as the differences between the mean AUC of different interpolation methods were commonly smaller than the differences between levels of other factors. Also, the *Interpolation Function* had the smallest sum of squares (0.136) among the four factors in Table 3. Generally, the *linear* kernel had the largest AUC and *logistic* regression had the smallest AUC. However, the differences among *Interpolation Functions* were practically small.

Accuracy maps for three example Trends blocks illustrate visually the similarity of the predicted accuracy maps from different methods relative to the reference accuracy map (Fig. 5). For the *spatial per-class* interpolation, the accuracy map exhibits continuity within classes and abrupt changes between classes. In the case of *spatial all-classes* interpolation, the accuracy continuity appears both within and between classes and the resulting map has the least similarity to the reference accuracy maps. Generally, *spectral* interpolations yield spatial patterns of predicted accuracy that are similar to the patterns of the reference accuracy maps whereas the *spatial* interpolations sometimes yield spatial patterns that do not correspond to those of the reference accuracy maps. For example, at the center of the Block 2 (highlighted with a red circle), there is a large correctly classified area. This area is assigned a high accuracy prediction value in spectrally interpolated accuracy maps. In contrast, this area is assigned some degree of misclassification when interpolation is done spatially even using a *per-class* interpolation. This can be attributed to the divergence of pixel reflectance from the same class. This is captured in the *spectral* interpolation; however, the *spatial* interpolation does not detect this variability due to close spatial proximity.

5. Discussion and conclusions

In this research per-pixel accuracy prediction methods were developed and evaluated. Spatially explicit representations of accuracy have been constructed in the past using separate interpolated maps for overall accuracy and user's and producer's accuracies of each class (Comber, 2013; Comber et al., 2012; Foody, 2005; Tsutsumida and Comber, 2015). A map with for example five classes would result in 11 such interpolated accuracy maps. Although each of these 11 maps will contain useful information, an advantage of our approach is that it encapsulates a large amount of accuracy information in a single predicted accuracy map. This simplifies subsequent usage of the predicted accuracy map. Moreover, while the previous researches used spatial domain for accuracy interpolation, spectral domain is used for the first time here.

The proposed new methods clearly outperformed the three benchmark methods included in this study. Three definitive conclusions regarding the proposed methods are: 1) Interpolations should be done separately for each class, with the advantages of *per-class* predictions attributable to their capacity to prevent confusion of accuracy from different classes, unlike *all-classes* predictions that are negatively affected by diverse accuracy rates among classes; 2) the method used for interpolation (*constant*, *linear*, *Gaussian*, or *logistic*) does not substantially impact the performance of the accuracy predictions; and 3) as expected, a larger sample size improves predictive accuracy. Each of these conclusions is discussed in additional detail in the following paragraphs.

From a practical perspective, the 2.5% test *Sample Size* represents a proportionally large sample size as in most applications the test sample size will be <0.5%. The Trends blocks offered an ideal set of test cases because of the intensive manual interpretation of the reference land-cover classes for the complete block allowed for comparison of the predicted accuracy to the known accuracy over the entire block. Given the limited area (100 km^2) of the Trends blocks, the sample density was necessarily high leading to test sample locations being closer in space than might be the case in some practical applications. Additional case study results for test sites covering larger spatial extents would be informative, but the challenge is to identify test datasets that could be used for this purpose.

The results comparing the *spectral* and *spatial* domains were less definitive as good predictive accuracy could be achieved using both domains. The sample density issue is clearly relevant in this regard, so

future work investigating less dense sampling intensities is needed to compare predictive performance when the sample pixels of the test dataset are not as close spatially as they were in the Trends blocks. The *spatial* methods offer a practical advantage in that they are independent of the classification's inputs. That is, the information required to implement a *spatial* method is the accuracy, the class label, and the spatial location of each test sample pixel, so the spatial methods completely separate the accuracy assessment from the classification process. In contrast, the *spectral* methods are dependent on the spectral inputs used in the classification. An advantage of the *spectral* method is that distances in spectral space are an intuitive measure expressing how likely it is for a pixel to be correctly classified.

To summarize, six major conclusions can be drawn from this study. First, the interpolation method used to predict accuracy at unsampled locations should be done separately for each class instead of grouping all classes together. Second, if an all-classes approach was used, interpolation using the spectral domain resulted in substantially greater AUC than interpolation in the spatial domain. Third, for the smaller sample size (0.5% sampling intensity) and per-class predictions, the spectral and spatial domain achieved similar AUC. Fourth, for the larger sample size yielding a very dense spatial sample, the spatial domain yielded greater AUC for per-class predictions relative to the AUC achieved for the spectral domain. Fifth, increasing the sample size improved accuracy predictions with a greater benefit accruing to the spatial domain. Lastly, the function used for interpolation had the smallest effect on AUC so choice of interpolation function is not critical to the process. The per-pixel accuracy prediction methods allow for producing wall-to-wall accuracy maps. These accuracy maps may serve to enhance applications of land-cover products by alerting map users to spatial variation of classification accuracy over the entire region mapped. In addition, our methods are agnostic to the classification algorithm used. This, together with the fact that the method is simple to develop and incorporate in existing software packages, indicates that the methodology is ready for operational use by the remote sensing community.

Acknowledgments

This work was supported by the USDA McIntire Stennis program, a SUNY ESF Graduate Assistantship and NASA's Land Cover Land Use Change Program (grant # NNX15AD42G). We thank Kristi Sayler and Mark Drummond (USGS) for providing the Trends data.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.rse.2017.01.025.

References

Ampe, E.M., Vanhamel, I., Salvadore, E., Dams, J., Bashir, I., Demarchi, L., et al., 2012. Impact of urban land-cover classification on groundwater recharge uncertainty. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 5, 1859–1867.
Anderson, J.R., Hardy, E.E., Roach, J.T., Witmer, R.E., 1976. A land use and land cover classification system for use with remote sensor data. US Geol. Surv. Prof. Pap. 964.
Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recogn. 30, 1145–1159.
Brown, K.M., Foody, G.M., Atkinson, P.M., 2009. Estimating per-pixel thematic uncertainty in remote sensing classifications. Int. J. Remote Sens. 30, 209–229.
Burnicki, A.C., 2011. Modeling the probability of misclassification in a map of land cover change. Photogramm. Eng. Remote Sens. 77, 39–50.
Campbell, J.B., 1981. Spatial correlation effects upon accuracy of supervised classification of land cover. Photogramm. Eng. Remote Sens. 47, 355–363.
Canters, F., 1997. Evaluating the uncertainty of area estimates derived from fuzzy land-cover classification. Photogramm. Eng. Remote Sens. 63, 403–414.
Carmel, Y., 2004. Characterizing location and classification error patterns in time-series thematic maps. IEEE Geosci. Remote Sens. Lett. 1, 11–14.
Carpenter, G.A., Gopal, S., Macober, S., Martens, S., Woodcock, C.E., Franklin, J., 1999. A neural network method for efficient vegetation mapping. Remote Sens. Environ. 70, 326–338.
Castilla, G., Hay, G.J., 2007. Uncertainties in land use data. Hydrol. Earth Syst. Sci. 11, 1857–1868.
Chen, D., Wei, H., 2009. The effect of spatial autocorrelation and class proportion on the accuracy measures from different sampling designs. ISPRS J. Photogramm. Remote Sens. 64, 140–150.
Colditz, R.R., Schmidt, M., Conrad, C., Hansen, M.C., Dech, S., 2011. Land cover classification with coarse spatial resolution data to derive continuous and discrete maps for complex regions. Remote Sens. Environ. 115, 3264–3275.
Comber, A., Fisher, P., Brunsdon, C., Khmag, A., 2012. Spatial analysis of remote sensing image classification accuracy. Remote Sens. Environ. 127, 237–246.
Comber, A.J., 2013. Geographically weighted methods for estimating local surfaces of overall, user and producer accuracies. Remote Sensing Letters 4, 373–380.
Congalton, R.G., 1988. Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. Photogramm. Eng. Remote Sens. 54, 587–592.
DeFries, R.S., Los, S.O., 1999. Implications of land-cover misclassification for parameter estimates in global land-surface models: an example from the simple biosphere model (SiB2). Photogramm. Eng. Remote Sens. 65, 1083–1088.
Dehghan, H., Ghassemian, H., 2006. Measurement of uncertainty by the entropy: application to the classification of MSS data. Int. J. Remote Sens. 27, 4005–4014.
Dieye, A.M., Roy, D.P., Hanan, N.P., Liu, S., Hansen, M., Touré, A., 2012. Sensitivity analysis of the GEMS soil organic carbon model to land cover land use classification uncertainties under different climate scenarios in Senegal. Biogeosciences 9, 631–648.
Foody, G.M., 2005. Local characterization of thematic classification accuracy through spatially constrained confusion matrices. Int. J. Remote Sens. 26, 1217–1228.
Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201.
Foody, G.M., 2000. Mapping land cover from remotely sensed data with a softened feedforward neural network classification. J. Intell. Robot. Syst. 29, 433–449.
Foody, G.M., Campbell, N.A., Trodd, N.M., Wood, T.F., 1992. Derivation and applications of probabilistic measures of class membership from the maximum-likelihood classification. Photogramm. Eng. Remote Sens. 58, 1335–1341.
Gahegan, M., Ehlers, M., 2000. A framework for the modelling of uncertainty between remote sensing and geographic information systems. ISPRS J. Photogramm. Remote Sens. 55, 176–188.
Ge, J., Qi, J., Lofgren, B.M., Moore, N., Torbick, N., Olson, J.M., 2007. Impacts of land use/cover classification accuracy on regional climate simulations. J. Geophys. Res. - Atmos. 112.
Ge, Y., Li, S., Lakhan, V.C., Lucieer, A., 2009. Exploring uncertainty in remotely sensed data with parallel coordinate plots. Int. J. Appl. Earth Obs. Geoinf. 11, 413–422.
GFOI, 2016. Integration of remote-sensing and ground-based observations for estimation of emissions and removals of greenhouse gases in forests. Methods and Guidance from the Global Forest Observations Initiative, Edition 2.0. Food and Agriculture Organization, Rome Available at:. https://www.reddcompass.org/download-the-mgd.
Giacco, F., Thiel, C., Pugliese, L., Scarpetta, S., Marinaro, M., 2010. Uncertainty analysis for the classification of multispectral satellite images using SVMs and SOMs. IEEE Trans. Geosci. Remote Sens. 48, 3769–3779.
Gómez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: a review. ISPRS J. Photogramm. Remote Sens. 116, 55–72.
Gong, P., Pu, R., Chen, J., 1996. Mapping ecological land systems and classification uncertainties from digital elevation and forest-cover data using neural networks. Photogramm. Eng. Remote Sens. 62, 1249–1260.
Khatami, R., Mountrakis, G., Stehman, S.V., 2016. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. Remote Sens. Environ. 177, 89–100.
Kyriakidis, P.C., Dungan, J.L., 2001. A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions. Environ. Ecol. Stat. 8, 311–330.
Leyk, S., Boesch, R., Weibel, R., 2005. A conceptual framework for uncertainty investigation in map-based land cover change modelling. Trans. GIS 9, 291–322.
Liu, W., Gopal, S., Woodcock, C.E., 2004. Uncertainty and confidence in land cover classification using a hybrid classifier approach. Photogramm. Eng. Remote Sens. 70, 963–971.
Loosvelt, L., De Baets, B., Pauwels, V.R.N., Verhoest, N.E.C., 2014. Assessing hydrologic prediction uncertainty resulting from soft land cover classification. J. Hydrol. 517, 411–424.
Loosvelt, L., Peters, J., Skriver, H., Lievens, H., Van Coillie, F.M., De Baets, B., et al., 2012. Random forests as a tool for estimating uncertainty at pixel-level in SAR image classification. Int. J. Appl. Earth Obs. Geoinf. 19, 173–184.
Loveland, T.R., Sohl, T.L., Stehman, S.V., Gallant, A.L., Sayler, K.L., Napton, D.E., 2002. A strategy for estimating the rates of recent United States land-cover changes. Photogramm. Eng. Remote Sens. 68, 1091–1099.
Löw, F., Michel, U., Dech, S., Conrad, C., 2013. Impact of feature selection on the accuracy and spatial uncertainty of per-field crop classification using Support Vector Machines. ISPRS J. Photogramm. Remote Sens. 85, 102–119.
Mas, J.-F., Filho, B.S., Pontius Jr., R.G., Gutiérrez, M.F., Rodrigues, H., 2013. A suite of tools for ROC analysis of spatial models. ISPRS Int. J. Geo-Inf. 2, 869–887.
Maselli, F., Conese, C., Petkov, L., 1994. Use of probability entropy for the estimation and graphical representation of the accuracy of maximum likelihood classifications. ISPRS J. Photogramm. Remote Sens. 49, 13–20.
Myers, D.E., 1994. Spatial interpolation: an overview. Geoderma 62, 17–28.
Mcgwire, K.C., Fisher, P., 2001. Spatially variable thematic accuracy: beyond the confusion matrix. In: Hunsaker, C.T., Goodchild, M.F., Friedl, M.A., Case, T.J. (Eds.), Spatial Uncertainty in Ecology: Implications for Remote Sensing and GIS Applications. Springer-Verlag, New York, pp. 308–329.
McIver, D.K., Friedl, M.A., 2001. Estimating pixel-scale land cover classification confidence using nonparametric machine learning methods. IEEE Trans. Geosci. Remote Sens. 39, 1959–1968.
McMahon, G., 2007. Consequences of land-cover misclassification in models of impervious surface. Photogramm. Eng. Remote Sens. 73, 1343–1353.
Miller, S.N., Phillip Guertin, D., Goodrich, D.C., 2007. Hydrologic modeling uncertainty resulting from land cover misclassification. J. Am. Water Resour. Assoc. 43, 1065–1075.

Mountrakis, G., Xi, B., 2013. Assessing reference dataset representativeness through confidence metrics based on information density. ISPRS J. Photogramm. Remote Sens. 78, 129–147.
Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57.
Pal, N.R., Bezdek, J.C., 1994. Measuring fuzzy uncertainty. IEEE Trans. Fuzzy Syst. 2, 107–118.
Peters, J., Verhoest, N.E.C., Samson, R., Van Meirvenne, M., Cockx, L., De Baets, B., 2009. Uncertainty propagation in vegetation distribution models based on ensemble classifiers. Ecol. Model. 220, 791–804.
Prasad, M.S.G., Arora, M.K., 2014. A simple measure of confidence for fuzzy land-cover classification from remote-sensing data. Int. J. Remote Sens. 35, 8122–8137.
Quaife, T., Quegan, S., Disney, M., Lewis, P., Lomas, M., Woodward, F.I., 2008. Impact of land cover uncertainties on estimates of biospheric carbon fluxes. Glob. Biogeochem. Cycles 22.
Scheuerer, M., Schaback, R., Schlather, M., 2013. Interpolation of spatial data - a stochastic or a deterministic problem? Eur. J. Appl. Math. 24, 601–629.
Seibert, J., McDonnell, J.J., 2010. Land-cover impacts on streamflow: a change-detection modelling approach that incorporates parameter uncertainty. Hydrol. Sci. J. 55, 316–332.
Sleeter, B.M., Sohl, T.L., Loveland, T.R., Auch, R.F., Acevedo, W., Drummond, M.A., et al., 2013. Land-cover change in the conterminous United States from 1973 to 2000. Glob. Environ. Chang. 23, 733–748.
Smith, J.H., Stehman, S.V., Wickham, J.D., Yang, L., 2003. Effects of landscape characteristics on land-cover class accuracy. Remote Sens. Environ. 84, 342–349.
Smith, J.H., Wickham, J.D., Stehman, S.V., Yang, L., 2002. Impacts of patch size and land-cover heterogeneity on thematic image classification accuracy. Photogramm. Eng. Remote Sens. 68, 65–70.
Steele, B.M., Chris Winne, J., Redmond, R.L., 1998. Estimation and mapping of misclassification probabilities for thematic land cover maps. Remote Sens. Environ. 66, 192–202.
Straatsma, M.W., van der Perk, M., Schipper, A.M., de Nooij, R.J.W., Leuven, R.S.E.W., Huthoff, F., et al., 2013. Uncertainty in hydromorphological and ecological modelling of lowland river floodplains resulting from land cover classification errors. Environ. Model. Softw. 42, 17–29.
Tsutsumida, N., Comber, A.J., 2015. Measures of spatio-temporal accuracy for time series land cover data. Int. J. Appl. Earth Obs. Geoinf. 41, 46–55.
Van Oort, P.A.J., Bregt, A.K., De Bruin, S., De Wit, A.J.W., Stein, A., 2004. Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database. Int. J. Geogr. Inf. Sci. 18, 611–626.
Verburg, P.H., Neumann, K., Nol, L., 2011. Challenges in using land use and land cover data for global change studies. Glob. Chang. Biol. 17, 974–989.
Verburg, P.H., van de Steeg, J., Veldkamp, A., Willemen, L., 2009. From land cover change to land function dynamics: a major challenge to improve land characterization. J. Environ. Manag. 90, 1327–1335.
Wang, Q., Shi, W., 2013. Unsupervised classification based on fuzzy c-means with uncertainty analysis. Remote Sens. Lett. 4, 1087–1096.
Wegehenkel, M., Heinrich, U., Uhlemann, S., Dunger, V., Matschullat, J., 2006. The impact of different spatial land cover data sets on the outputs of a hydrological model - a modelling exercise in the Ucker catchment, north-east Germany. Phys. Chem. Earth 31, 1075–1088.
Yu, Q., Gong, P., Tian, Y.Q., Pu, R., Yang, J., 2008. Factors affecting spatial variation of classification uncertainty in an image object-based vegetation mapping. Photogramm. Eng. Remote Sens. 74, 1007–1018.
Zhang, J., Zhou, Y.K., Li, R.Q., Zhou, Z.J., Zhang, L.Q., Shi, Q.D., et al., 2010. Accuracy assessments and uncertainty analysis of spatially explicit modeling for land use/cover change and urbanization: a case in Beijing metropolitan area. Sci. China Earth Sci. 53, 173–180.