Remote Sensing of Environment 199 (2017) 401–414

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

Predicting individual pixel error in remote sensing soft classification

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

Keywords:
Sub-pixel land-cover mapping
Classification accuracy assessment
Spectral unmixing
Error map
Local accuracy
Image classification

ABSTRACT

Accuracy assessment of remote sensing soft (sub-pixel) classifications is a challenging topic. Previous efforts have focused on constructing a soft classification error matrix and producing summary measures to describe overall and per-class map accuracy. However, these summary assessments do not provide information on the spatial distribution of the soft classification error as distributed at the individual pixel level. This is important because the map error of a given class may vary considerably over different regions. Spatial interpolation has been previously used for predicting soft classification error at the pixel level. Here, we propose two alternative domains for soft classification error interpolation, the spectral and mapped class proportion domains. In the spectral domain we interpolate errors in the classification feature space, whereas in the mapped class proportion domain interpolation takes place in a space with dimensions defined by the mapped class proportions (i.e., the output of the soft classification). The two newly proposed prediction methods (spectral domain and mapped class proportion domain), spatial interpolation, and a summary measure method were evaluated using 23 test regions, each 10 km × 10 km, distributed throughout the United States. These 10 km × 10 km blocks had complete coverage reference data (where the reference classification was determined by manual interpretation) and the predicted error maps were then evaluated by comparing them to these complete coverage reference error maps. Mean absolute error was used to quantify the agreement of the predicted error maps to the reference error maps. The spectral and mapped class proportion methods generally outperformed the spatial interpolation and the summary measure methods both in terms of smaller mean absolute error and visual similarity of predicted error maps to the reference error maps. The superiority of the new methods over spatial interpolation is an important result because spatial interpolation is a familiar method analysts would commonly consider for modeling spatial variation of classification error. The predicted soft classification error maps provide a straightforward visual assessment of the spatial patterns of error that can accompany the original classification products to enhance their value in subsequent analysis and modeling tasks. Furthermore, from the standpoint of implementation, our methods do not require additional datasets; the same test dataset currently used for confusion/error matrix construction can be used for our error interpolation methods.

1. Introduction

Classified land-cover maps have become one of the most important products of remote sensing science and industry enabling environmental and natural resources monitoring, modeling and management from local to global spatial extents. Land-cover maps are essential inputs for a broad range of applications such as forest and carbon monitoring (Carreiras et al., 2012; Dong et al., 2003; Eva et al., 2012; Réjou-Méchain et al., 2014); environmental change detection (Roy et al., 2014; Wulder et al., 2008); climate studies (Grimm et al., 2008; Seneviratne et al., 2010); and hydrological modeling (Khan et al., 2011; Nie et al., 2011; Sorooshian et al., 2014). Significant work has been done by the remote sensing community to improve the associated classification processes and to increase the accuracy of classified land-cover maps (Cihlar, 2000; Franklin & Wulder, 2002; Gómez et al., 2016; Khatami et al., 2016; Lu & Weng, 2007). Land-cover classification can be generally divided into two major categories, hard and soft classifications. In hard classifications, each pixel is assigned to a single class, whereas in soft classification, a pixel may belong to multiple classes and different levels of class membership or proportion are assigned. Soft classifications can potentially be very useful when a large number of mixed pixels exists in an image (Foody & Doan, 2007; Paneque-Gálvez

* 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.2017.07.028
Received 23 January 2017; Received in revised form 18 July 2017; Accepted 25 July 2017
Available online 04 August 2017
0034-4257/

et al., 2013; Tsutsumida et al., 2016), as for example when the scene is heterogeneous and the pixel size is larger than the size of the objects of interest.

Whether a hard or soft classification is implemented, it is important to quantify classification error. The accuracy assessment of a hard classification is typically reported through the error or confusion matrix and summary measures derived from it that describe the accuracy of the entire map or a class (Foody, 2002; Olofsson et al., 2014; Stehman & Czaplewski, 1998; Story & Congalton, 1986). In addition, per-pixel classification accuracy prediction methods for hard classification have been investigated to produce maps depicting the spatial distribution of classification accuracy (Comber et al., 2012; Comber, 2013; Foody, 2005; Khatami et al., 2017; Kyriakidis & Dungan, 2001; Steele et al., 1998; Tsutsumida & Comber, 2015) or classification confidence (Mountrakis & Xi, 2013). Khatami et al. (2017) also investigated factors affecting per-pixel accuracy interpolation of hard classifications.

Accuracy assessment of soft classifications is more challenging because the concept of the error matrices typically used in hard classifications cannot be directly applied for soft classifications. Efforts to construct error matrices for soft classification analogous to those applicable to a hard classification include fuzzy error matrix (Binaghi et al., 1999; Stehman et al., 2007) and soft classification error matrix (Latifovic & Olthof, 2004; Pontius Jr. & Cheuk, 2006). Generally, the objective is to build an error matrix for each test pixel based on the reference class proportions and mapped (from soft classification) class proportions. Error matrices for all test pixels can then be aggregated to produce a single estimated error matrix for the entire map. Summary measures such as overall, user's, and producer's accuracies can be estimated from the aggregated error matrix. However, because the spatial distribution of the reference and mapped class proportions within each test pixel is unknown, it is not possible to exactly determine the true overlap among reference and mapped classes and obtain the true error matrix for each test pixel. This issue is known as “sub-pixel area allocation problem” (Silván-Cárdenas & Wang, 2008). Many approaches or operators have been devised to allocate the overlap among reference and mapped class proportions to construct the error matrix of a given test pixel. Some of these methods include fuzzy minimum operator (Binaghi et al., 1999); composite operator (Pontius Jr. & Cheuk, 2006); product operator (Lewis & Brown, 2001); similarity index (Townsend, 2000); and confusion intervals (Silván-Cárdenas & Wang, 2008). Silván-Cárdenas & Wang (2008) introduced a series of characteristics for an ideal per-pixel confusion matrix and discussed whether different operators could result in error matrices that hold those characteristics.

Another group of methods employed for accuracy assessment of soft classifications is based on directly measuring the proximity, similarity, or correlation among the reference and mapped class proportions. Commonly, an accuracy summary measure is calculated using test data to describe how close the reference and mapped class proportion values are for the entire classification or for a given class. Some of these measures include Euclidian and city block distance (Foody, 1996; Foody & Arora, 1996); root mean squared error (RMSE) (Carpenter et al., 1999; Chen et al., 2010; Lu & Weng, 2006; Olthof & Fraser, 2007); correlation coefficient (Foody & Cox, 1994; Maselli et al., 1996); entropy (Finn, 1993; Maselli et al., 1994); cross-entropy (Foody, 1995); information closeness (Foody, 1996); weighted disagreement (Gómez et al., 2008); kappa coefficient (Homer et al., 2012); and Morisita's index (Ricotta, 2004). Similarity can also be assessed in the context of fuzzy logic (Foody, 1999; Gopal & Woodcock, 1994; Laba et al., 2002; Woodcock & Gopal, 2000).

The summary measures derived from the two aforementioned general categories of soft classification accuracy assessment are useful to describe the classification accuracy at the general map scale. However, the summary measures do not provide specific information about the spatial distribution of the classification error. This is an important issue because the classification accuracy would likely vary over different regions of the map (Campbell, 1981; Chen & Wei, 2009; Congalton, 1988) and the summary measures may not be useful when the local accuracy for an area of interest differs from the global accuracy (Mcgwire & Fisher, 2001).

Classification errors affect the reliability of subsequent map use for environmental analyses and modeling (Castilla & Hay, 2007; Ge et al., 2007; Jin et al., 2014; McMahon, 2007; Straatsma et al., 2013). Because classification accuracy varies over different map regions, subsequent models would also inherit this spatial accuracy variation. Thus, environmental modeling can be greatly enhanced if local estimates of classification accuracy or error are available (DeFries & Los, 1999; Gahegan & Ehlers, 2000; Miller et al., 2007). Consequently, in this research we focus on pixel-level error map construction for land-cover maps created by soft classification of remotely sensed imagery. Spatial interpolations have been previously suggested to create error maps for soft classifications (Comber, 2013; Foody, 2005). In this manuscript, we introduce spectral and mapped class proportion domains as the explanatory domain for error prediction, domains that to the best of our knowledge have not been previously explored for error interpolation of soft classifications. The performances of the spectral and mapped class proportion interpolation methods are compared to two benchmark methods, a map-level summary measure and a spatial interpolation method.

The rest of the manuscript is organized as follows. In Section 2, the datasets including the reference data and satellite images used to evaluate the soft classification error mapping methods along with input data preprocessing are presented. The details of the four error prediction methods are explained in Section 3. In Section 4, the research experimental design and the overall workflow is elaborated, including details of the four main steps: (i) input data preprocessing, (ii) land-cover soft classification, (iii) classification error map predictions (using the methods discussed in Section 3), and (iv) evaluation of the predicted error maps. In Section 5, results of evaluation of the error prediction methods are discussed. Evaluations are based on (i) quantitative analysis using mean absolute error (MAE) as a measure to quantify prediction error and (ii) visual investigation and comparison of the reference and predicted error maps. Discussion and conclusions are presented in Section 6.

2. Datasets used to evaluate methods for predicting per-pixel error

The performances of error prediction methods were investigated using reference data from the United States Geological Survey (USGS) Land-Cover Trends project (Loveland et al., 2002). Twenty-three Trends blocks (Fig. 1) were used to provide a diverse set of examples to evaluate the error prediction methods. Each block represents a special case study. The 2011 Trends reference data for each block were obtained using manual interpretation of all pixels in the block providing a census of reference data at a 30 m pixel size. Each pixel was assigned a single class (hard classification) based on a modified Anderson (Anderson et al., 1976) Level I classification scheme including 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 specific class definitions, last accessed April 2017). Each of the 23 blocks covered a 10 km × 10 km (333 pixels × 333 pixels) area (blocks are enlarged to enhance visualization in Fig. 1). Because the Trends reference data represent a hard classification, a recoding and resampling process was applied to convert these data to a soft classification for use as reference data as discussed below.

To evaluate the error prediction methods it was necessary to produce land-cover soft classification maps. The land-cover classification for each of the 23 Trends blocks was implemented by applying a spectral unmixing method using six reflective bands (excluding thermal) from Landsat TM images for 2011. This method required that the number of target classes not be larger than the number of spectral

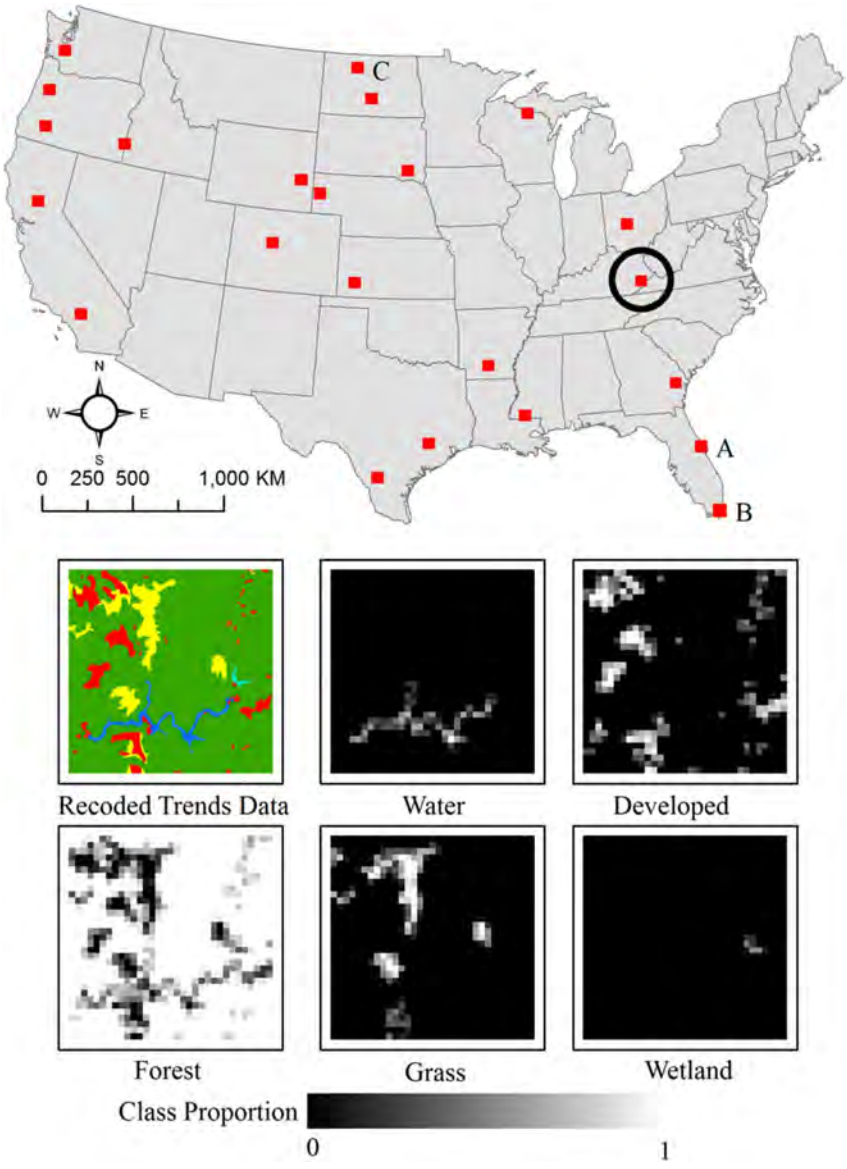

Fig. 1. Top: Spatial distribution of the 10 km × 10 km Trends blocks. Bottom: Recoded Trends data (30 m pixel size) for the circled sample block and the corresponding coarse resolution (300 m pixel size) class proportions. Blocks A, B, and C are highlighted for special attention later in the text.

bands plus one. Consequently, the number of classes was reduced from 11 to 6 by assigning mechanically disturbed, barren, mining, and non-mechanically disturbed classes to the developed/urban class. This scheme of recoding was applied because the spectral signatures of the four recoded classes have the most similarity to that of the developed/urban class among all other classes. In addition, the four recoded classes typically represented only a small percent of block area ranging from a mean (over the 23 blocks) of 1.7% for the mechanically disturbed class to a mean of 0.01% for the non-mechanically disturbed class. The ice/snow class does not exist in these 23 blocks.

As mentioned, in the Trends data each 30 m pixel is assigned to a single class. In order to obtain soft classification reference data, a coarser resolution reference dataset was created from each recoded Trends block. The coarse resolution dataset was created by aggregating the 30 m × 30 m pixels of the Trends blocks to a 10 pixel × 10 pixel scale resulting in a 300 m × 300 m pixel size dataset. For each 300 m pixel the proportion of each of the six land-cover classes was calculated by counting the number of 30 m pixels from the same class in the recoded Trends data covered by the 300 m pixel. Similarly, Landsat images were resampled to 300 m pixel size. The spectral value of each 300 m pixel, at each band, was calculated as the average spectral value of the one hundred 30 m pixels it covered. Soft classifications and pixel-level class proportion error predictions were exclusively implemented on the coarse resolution dataset.

The reference data from each Trends block were used for three purposes: i) a sample of pixels was selected and the Trends reference data were used as training data to implement a soft classification (details are described later), ii) a second sample of pixels was selected independently of the training sample and the Trends reference data were used as test data to implement the error prediction methods and produce the error maps, and iii) lastly, all pixels of the given block were used as a census of reference data to evaluate the predicted error maps produced from the test sample.

3. Methods for predicting pixel-level class proportion error

Generally a test dataset, derived from a sample of reference data is required to evaluate the performance of a classification and to quantify the errors of the mapped class proportions of area. For pixel i we define the error of class c, err_{ic}, as follows:

$$err_{ic} = p_{ic} - \hat{p}_{ic},\ c = 1,\ \ldots, C \tag{1}$$

where, p_{ic} and $\hat{p}_{ic}$ are the reference and mapped (from soft classification) class proportions of class c in pixel i, and C is the number of

classes. The reference and mapped proportions of a given class can be real numbers between 0 and 1, where 0 is complete absence and 1 is complete presence of the class in a given pixel (the mapped class proportion values are bounded within 0 and 1 by applying full additivity and non-negativity constraints during soft classification, which is discussed later). Therefore, err_{ic} values can range from -1 to $+1$ with positive and negative values corresponding to underestimation and overestimation of the proportion of a class for a pixel, respectively. In practice, these error values, err_{ic}, are known for the test sample pixels. Interpolation methods are then applied using these known error values to predict error for all pixels in the map that were not included in the test sample. Four error prediction methods are examined, the constant method based on mean error of a class, spatial interpolation, and two newly proposed interpolations in the spectral and mapped class proportion domains. These methods are elaborated below.

3.1. Interpolation domains

Three interpolation approaches based on three different explanatory domains (i.e. feature spaces) were investigated in this research to propagate the observed class proportion errors from the test sample pixels to the unsampled pixels. The explanatory domains included spatial, spectral, and mapped class proportion domains. The spatial domain is the two-dimensional spatial distribution of pixels on the ground. This domain has been previously investigated (Comber, 2013; Foody, 2005) for soft classification error interpolation and will serve as our first benchmark. The underlying assumption for the spatial interpolation of error is that pixels in close spatial proximity will have similar error rates. In this manuscript we examine two new domains, the spectral domain and the mapped class proportion domain. The spectral domain is based on the L-dimensional spectral space, where L is the number of bands of the image (or alternatively any features used in the classification process). The coordinates of a pixel in this space correspond to its spectral values for different bands. Spectral interpolation was used because classification takes place in the spectral domain and it can be expected that pixels close in spectral domain will have similar error rates. The last domain was the C-dimensional mapped class proportion domain, where C is the number of classes. Basically, in soft classification every pixel is decomposed into proportions of different classes. Therefore, every classified pixel can be considered as a point in the C-dimensional mapped class proportion space. For interpolations in this domain, the underlying assumption is that soft classification errors for an unsampled pixel will be close to errors of known test pixels that have similar class composition.

3.2. Interpolation function

A linear kernel function was used for interpolation of err_{ic} values by all three interpolation methods. Error interpolation was conducted separately for each class using the same test sample pixels. In the interpolation function, the weighted average of err_{ic} values of an unsampled pixel's nearby test pixels was assigned as its error for class c. The linear kernel function was used to assign weights to the nearby test pixels (proximity can be defined in any of the three domains) of a given unsampled pixel while predicting its class proportion error. The weights were computed based on the Euclidian distances between test pixels and the unsampled pixel using Eq. (2):

$$w_{ijc}(h) = 1 - \frac{h_{ij}}{hmax_{ic}} \tag{2}$$

where w_{ijc} is the weight of the jth nearest test sample pixel to the unsampled pixel i for class c, h_{ij} is Euclidian distance between the unsampled pixel i and its jth nearest test sample pixel, and h_max_{ic} is the maximum distance between the unsampled pixel i and its K nearest test pixels. For any interpolation, the choice of K, the number of nearest test pixels used for interpolation of an unsampled pixel, is an important factor that strongly impacts local predictions. A very small K could result in predictions with high variance whereas a very large K could mask local variations of the quantity of interest. A cross-validation process was used to find the optimal number of neighbors (*optimal_K*), and this process was implemented independently for each interpolation method and class (details described later). To avoid obtaining zero weights when all of the nearest test pixels have the same distance to the unsampled pixel the maximum distance was multiplied by 1.001. The slope of the kernel defines the rate at which the weights decrease as the distance increases. Because the h_max_{ic} would be different for each unsampled pixel the slope would adaptively change for each unsampled pixel based on the density of test pixels within the local neighborhood of the unsampled pixel. Note that interpolations of error were done separately for each class and $optimal\_K_c$ can vary for different classes. Consequently, h_max_{ic} and accordingly w_{ijc} for an unsampled pixel i can vary for different classes.

After the weight values were calculated, the proportion error $\widehat{err}_{ic}$ of class c for unsampled pixel i was predicted using the following equation, which is a weighted average of the known class proportion error of the nearby test sample pixels:

$$\widehat{err}_{ic} = \frac{\sum_{j=1}^{optimalK_c} w_{ijc} \times err_{jc}}{\sum_{j=1}^{optimalK_c} w_{ijc}} \tag{3}$$

All three interpolation methods used the above interpolation process (Eqs. 2 and 3). Interpolations were done independently in each of the three interpolation domains.

3.3. Constant benchmark

In addition to the spatial interpolation method that acted as our first benchmark, a second benchmark method was defined as the average of the class proportion error values (i.e., mean over/under-estimation error) of a given class over all test pixels. Thus, the mean error was the constant value used as the predicted class proportion error of that class for all pixels in the map:

$$\widehat{err}_{ic} = \frac{1}{n} \sum_{j=1}^{n} err_{jc},\ c = 1, \ldots, C \tag{4}$$

where n is the number of test pixels and err_{jc} is the reference class proportion estimation error of class c of the jth test pixel. This method would be the equivalent of using a summary measure to describe the accuracy of a given class for the entire map. For example, one class might be on average 5% over-estimated over the entire map. If no local prediction of error is done, the best available estimation of class proportion error of the given class for all pixel would be 5% over-estimation.

4. Experimental design

In this section, the steps of implementing the soft classifications, the error predictions, and evaluation of predictions are elaborated. Our experiment was composed of four general steps (Fig. 2). First, for each block, the Trends classes were recoded to the six classes and coarse resolution data (300 m pixel) were created for both the Trends data and the corresponding Landsat image. Second, soft classification of the coarse resolution Landsat image was conducted using the spectral unmixing method (Guerschman et al., 2009; Iordache et al., 2011; Keshava & Mustard, 2002). This resulted in six class proportion maps (or fewer if some classes did not exist in a block), covering the entire block at the 300 m pixel size and corresponding to the six classes in the classification scheme. Then, a test dataset was selected from the coarse resolution Trends data. This test dataset would represent the sample of reference data used to assess map accuracy in a practical application.

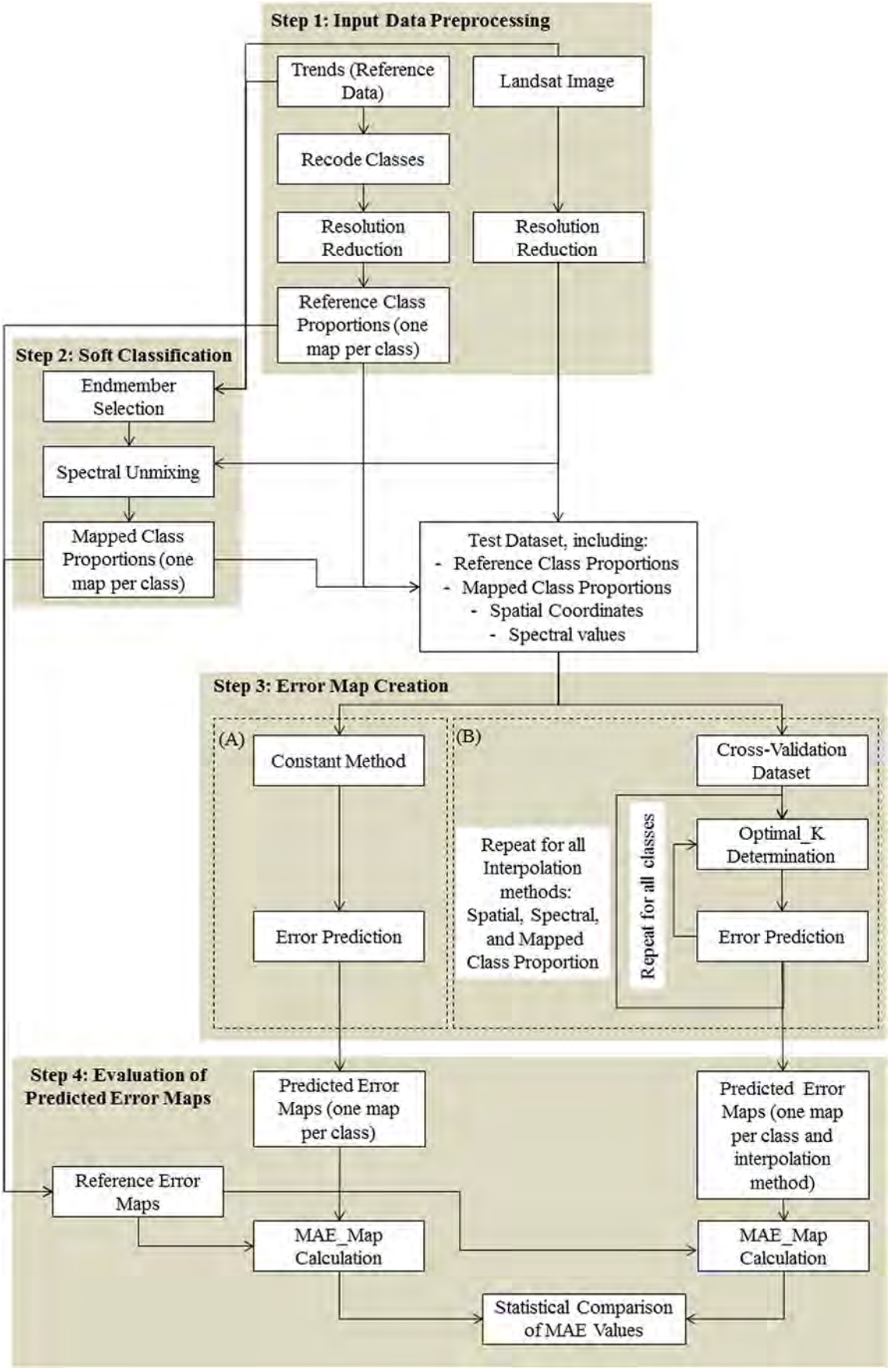

Fig. 2. The overall workflow of image soft classification and error map creation.

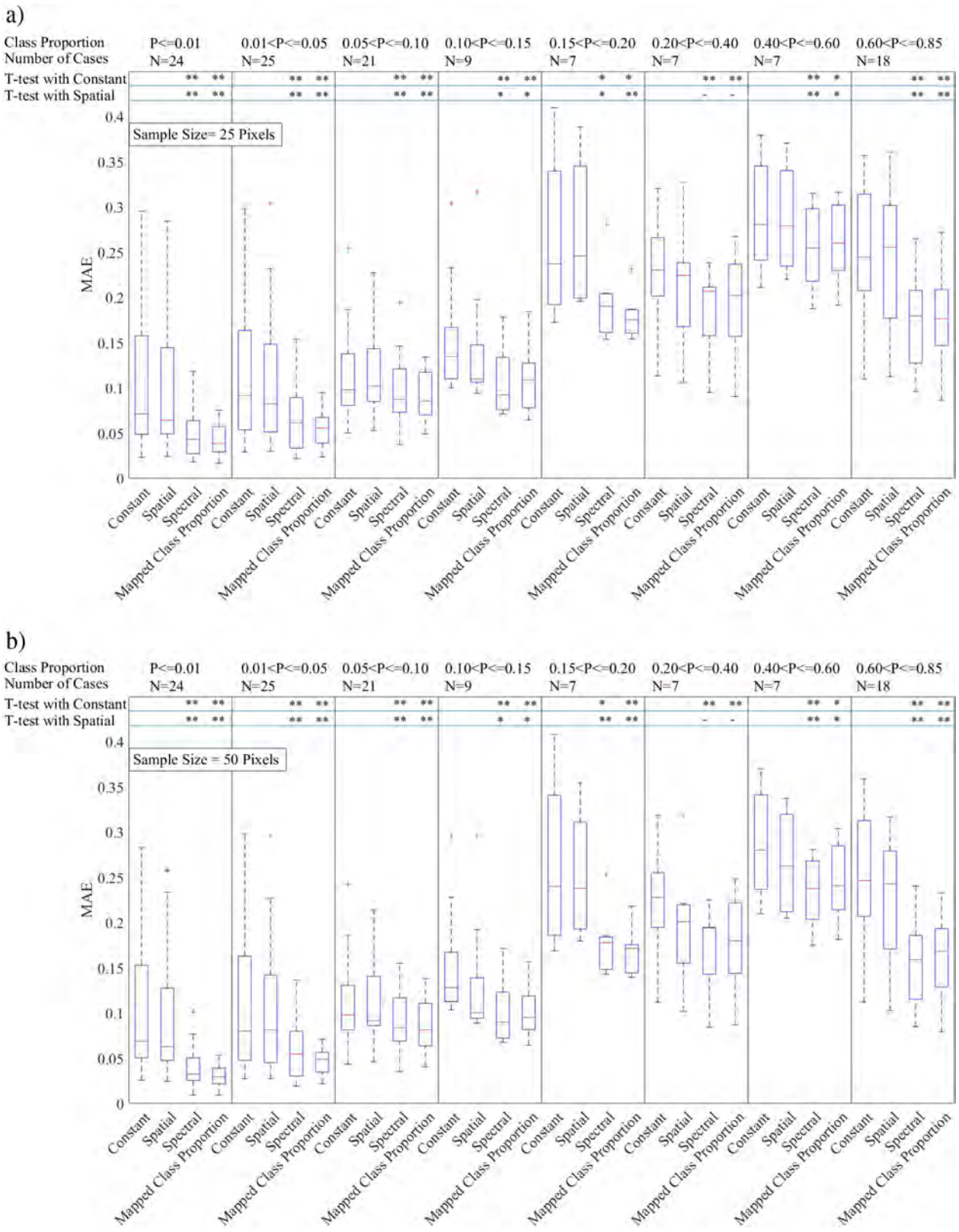

Fig. 3. Boxplots of MAE values for all classes in all Trends blocks grouped by reference class proportion P (25 pixel test sample size). First and second rows of symbols at the top of each plot show the paired *t*-test results comparing the new (spectral domain and mapped class proportion domain) interpolations to the constant and spatial interpolation benchmark methods (* = 0.05, ** = 0.01 significance levels, - indicates not statistically significant at the 0.05 level).

The four error prediction methods were implemented using the test sample data for that block and a predicted error map was created for each class by each prediction method. In addition, the "reference" error map for each class was constructed by subtracting the mapped class proportion (from the soft classification) from the reference class proportion (from the coarse resolution Trends data) for all pixels in the block. These "reference" error maps show the exact amount of pixel-level class proportion error for all classes and were used to validate the

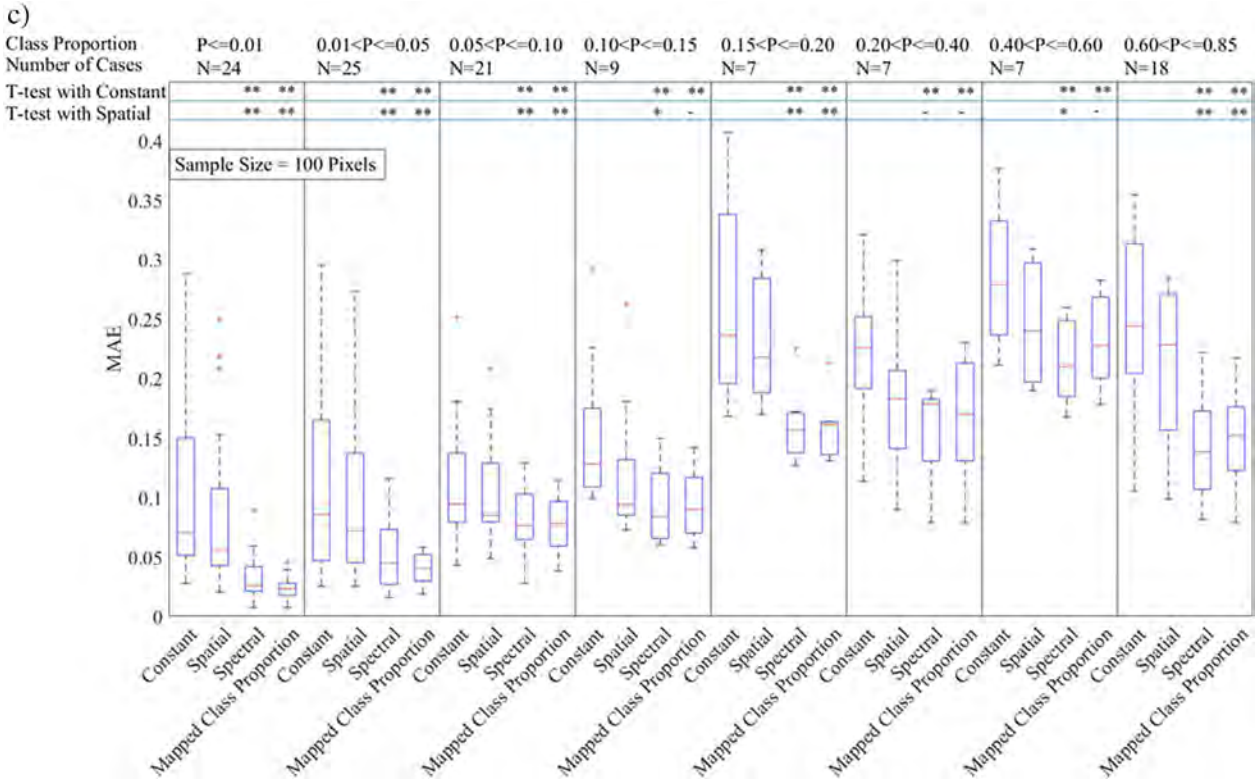

Fig. 3. (*continued*)

performance of the predicted error maps created by the four methods from the test sample. The entire process was implemented independently for each block. The four general steps are further elaborated below:

4.1. Step 1 (input data preprocessing)

- The Trends data were recoded from 11 classes to 6 classes (see Section 2).
- The pixel size of both the recoded Trends data and Landsat image were increased to 300 m (see Section 2). For each pixel of the coarse resolution Trends data the reference proportion of each of the six land-cover classes was calculated based on the number of pixels from that class in the original 30 m Trends data covered by the given 300 m pixel. This would result in a "reference" class proportion map for each land-cover class present in a block.

4.2. Step 2 (soft classification)

- The spectral unmixing method was used for soft classification of Landsat images.
- A key requirement of the spectral unmixing method is identifying the spectra of the endmembers or pure pixels from each class. The original 30 m Trends and Landsat images were used for endmember extraction. First, for each class all 5 pixel × 5 pixel homogenous areas (i.e. covered with a single class) were determined from the 30 m Trends data. Then, five of these homogeneous areas were selected at random for each class. Finally, the spectra of the central pixel of these selected homogeneous areas were determined from the corresponding 30 m resolution Landsat image. The average spectra of the five selected homogeneous pixels were used as the spectra of the endmember of that class during the spectral unmixing process. Because the soft classification was conducted at 300 m resolution, 30 m resolution data could be considered as an appropriate source for endmember extraction.
- A constrained least squares method, based on minimizing the squared error, was used to solve for the p_{ic} values for each pixel of the reduced resolution Landsat image. The output of the least squares solution, $\widehat{p}_i$, was the $C \times 1$ vector of mapped proportions of different classes in pixel i. Two constraints were considered for the least squares solutions: full additivity and non-negativity. Full additivity, also known as orthogonality, is a constraint that ensures that the mapped proportions of all classes for a given pixel sum to one ($\sum_{c=1}^{C} \widehat{p}_{ic} = 1$). The non-negativity constraint ensures that the mapped proportion values cannot be negative ($\widehat{p}_{ic} \geq 0$, $c = 1, \ldots, C$). Other classification processes (Bioucas-Dias et al., 2012; Mertens et al., 2003; Plaza et al., 2009; Tatem et al., 2002; Xu et al., 2005) and endmember selection techniques (Miao & Qi, 2007; Nascimento & Dias, 2005; Plaza et al., 2004) could be used; however, the focus of this research was on accuracy assessment rather than optimization or evaluation of the classification process.
- The soft classification resulted in a class proportion map for each land-cover class present in a block.

4.3. Step 3 (error map creation)

After the soft classification, a test dataset was selected using simple random sampling. Because the "reference" and "mapped" class proportions of these sampled test pixels were known, the actual (true) values of err_{ic}, the class proportion errors (Eq. 1), were also known for these test pixels. Step 3 (error map creation) was conducted independently for each land-cover class using the same test dataset. Boxes A and B in Fig. 2 show the error map production process for the constant benchmark and the interpolation (spatial, spectral, and mapped class proportion domains) methods, respectively.

4.3.1. Step 3-A (constant benchmark method)

- The mean value of class proportion error over all sampled test pixels was calculated for each class. The mean error value of each class was

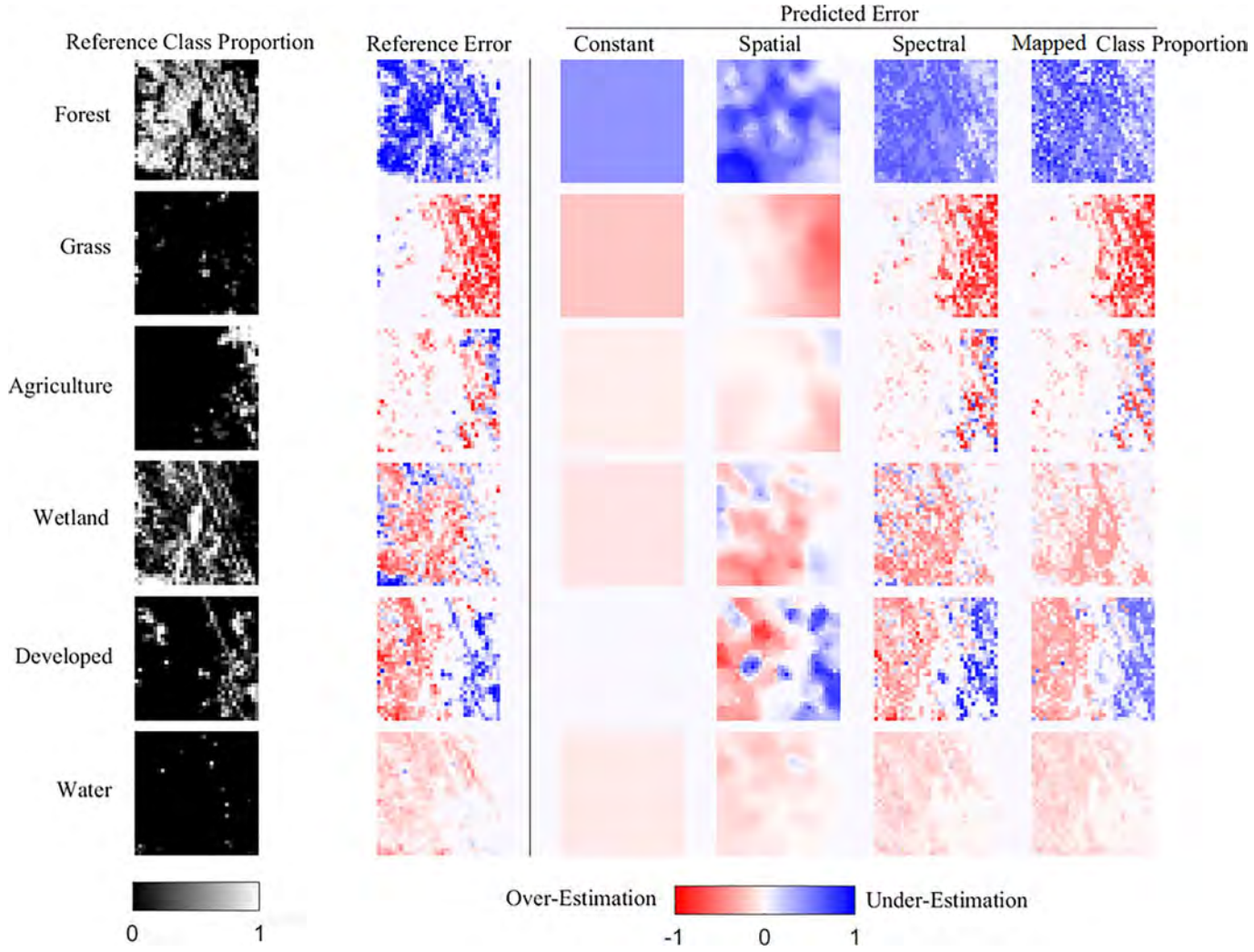

Fig. 4. Reference class proportion, reference error, and predicted error maps (from 100 pixel test sample size) for six classes of the Trends block A in Fig. 1.

assigned to all pixels in the map as the predicted class proportion error of that class.

4.3.2. Step 3-B (interpolation methods)

Three interpolation methods based on three different explanatory domains (i.e., spatial, spectral, and mapped class proportion domains) were implemented. For each class and interpolation method, a ten-fold cross-validation optimization process, as elaborated below, was used to determine the optimal number of nearest neighbor test pixels (*optimal_*K_c) which was used during error interpolation of the given class with the given method. Values of K_c from 1 to 20 were tested to determine the *optimal_*K_c:

- First, 10 non-overlapping subsets were created by randomly dividing the test pixels. The same 10 subsets were used by all interpolation methods.
- The class proportion errors were predicted for pixels in each of the cross-validation subsets using the data from the other 9 subsets and the given interpolation method for each possible number of nearest neighbors, K = 1, 2, …, 20.
- Then, the predicted class proportion errors (for the given class) of all test sample pixels were compared to their corresponding reference values. Mean absolute error (MAE) between the reference and predicted errors (Eq. 5) was used to evaluate the error predictions:

$$MAETest_c := \frac{1}{n}\sum_{i=1}^{n} |err_{ic} - \widehat{err}_{ic}| \quad (5)$$

where n is the number of test pixels.

- The MAE was calculated for each set of predictions based on different values of K from 1 to 20.
- The *optimal_*K_c, for the given class and method, was selected as the number of neighbors that resulted in the lowest MAE.
- After *optimal_*K_c determination, the class proportion error of the target class for all pixels in the block was predicted using the sample test pixels and the given method based on the *optimal_*K_c. This resulted in an error map for each class predicted by the given interpolation method from the test sample data.
- The *optimal_*K_c determination and interpolation was repeated for each class using each of the three interpolation methods. Note that the *optimal_*K_c can vary for different classes and interpolation domains.

4.4. Step 4 (evaluation of predicted error maps)

- First, the "reference" error map of each class with values from − 1 to + 1 was created by subtracting the mapped class proportions (determined from the soft classification, see step 2) from the reference class proportions (determined from the coarse resolution Trends data, see step 1). The reference error map provides a census of

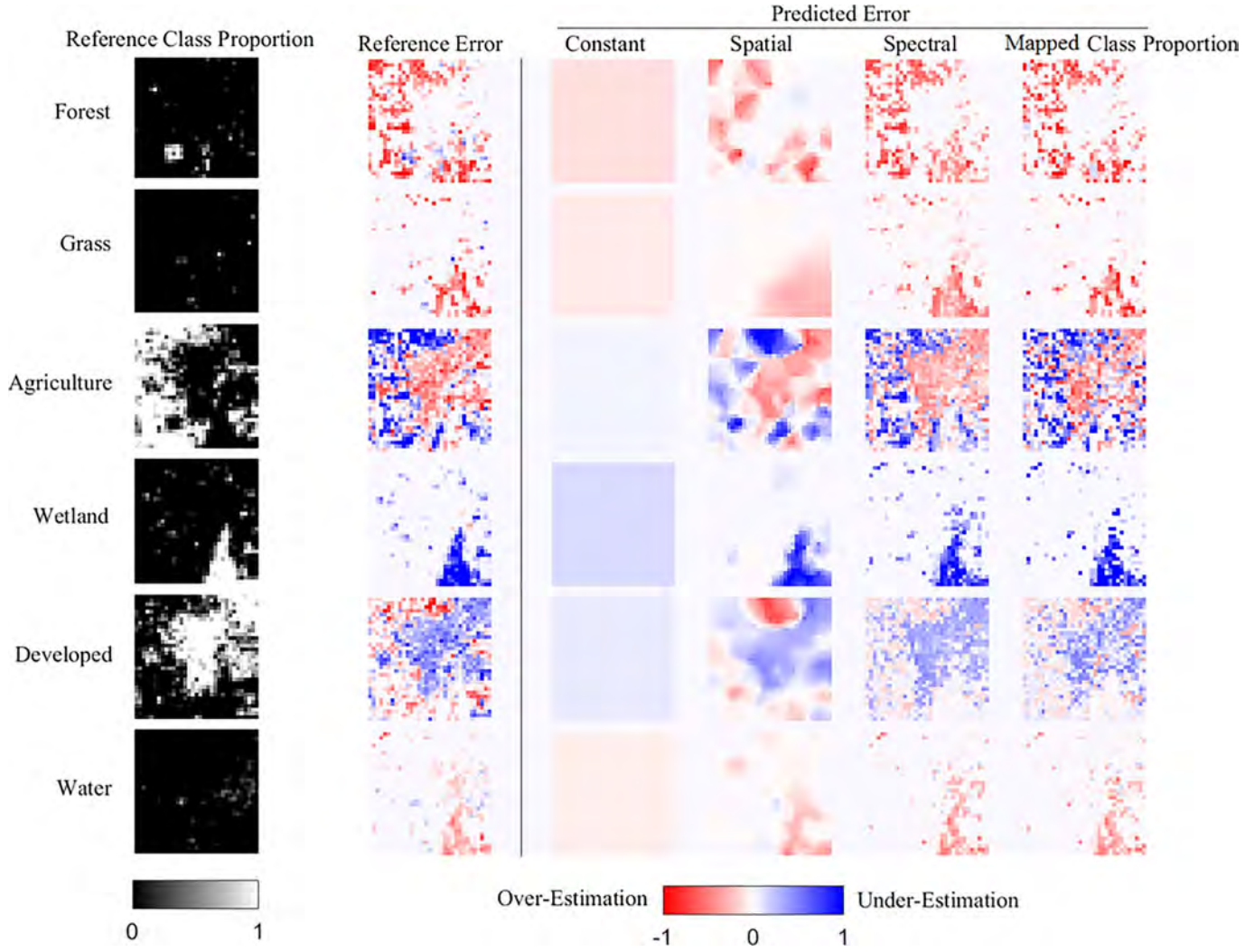

Fig. 5. Reference class proportion, reference error, and predicted error maps (from 100 pixel test sample size) for six classes of the sample Trends block B in Fig. 1.

reference error values of the given class for the entire block. Ideally, the predicted error map of each class should be similar to its corresponding reference error map.

- Then, for each class, the reference error map was compared to the predicted error maps created from the four different methods, where each method used the same test sample data for each block. The MAE between reference and predicted error maps was used to evaluate the predictions from each method for each class (Eq. 6). Note that these MAE values were calculated using the reference and predicted error of all pixels contained in a block and are used for validation of predictions. These MAE values were different from the MAE values calculated during *optimal_K*$_c$ determination in Eq. 5. The MAE values used to evaluate performance of each error prediction method were calculated as follows:

$$MAEMap_{c\cdot\cdot} = \frac{1}{N}\sum_{i=1}^{N} |err_{ic} - \widehat{err}_{ic}| \tag{6}$$

where N is the total number of pixels in the Trends block.

The entire workflow in Fig. 2 was applied to all 23 Trends blocks independently. Three test dataset sample sizes of 25, 50, and 100 pixels were implemented for error map prediction. In addition, the sampling of the test datasets from each Trends block was repeated 10 times, resulting in 10 different sample test datasets for each sample size, to account for variability in performance of prediction methods due to sampling variability. Therefore, each soft classification of a block (from step 2) was assessed using each of the 10 test sample datasets, producing 10 error maps for each class, prediction method, and sample size. In other words after completion of steps 1 and 2, steps 3 and 4 were repeated for three sample sizes and 10 times for each sample size for each block. The Mean Absolute Error (MAE) values of each method for each class present in a block were averaged over the 10 test samples for each test sample size. Some sample blocks did not have all six classes so in such cases MAE for those classes was not computed.

To compare the performance of the new interpolation methods (i.e., interpolations in spectral and mapped class proportion domains) with those of the two existing benchmark methods (i.e., constant and spatial interpolation methods), paired *t*-tests were conducted between MAE values of each pair of new and existing methods. The null hypothesis of the paired *t*-tests was equality of mean MAE values of a new method and a benchmark method and the alternative hypothesis was that the mean MAE of the new method was smaller than that of the benchmark method.

5. Results

Fig. 3 shows the boxplots of MAE values for the four error prediction methods from all classes in all 23 blocks and for all three test sample sizes (see appendix Table S1 for mean MAE values). The boxplots group results based on the reference class proportion (denoted as P) presented at the top of each boxplot. For example, the first four boxplots on the

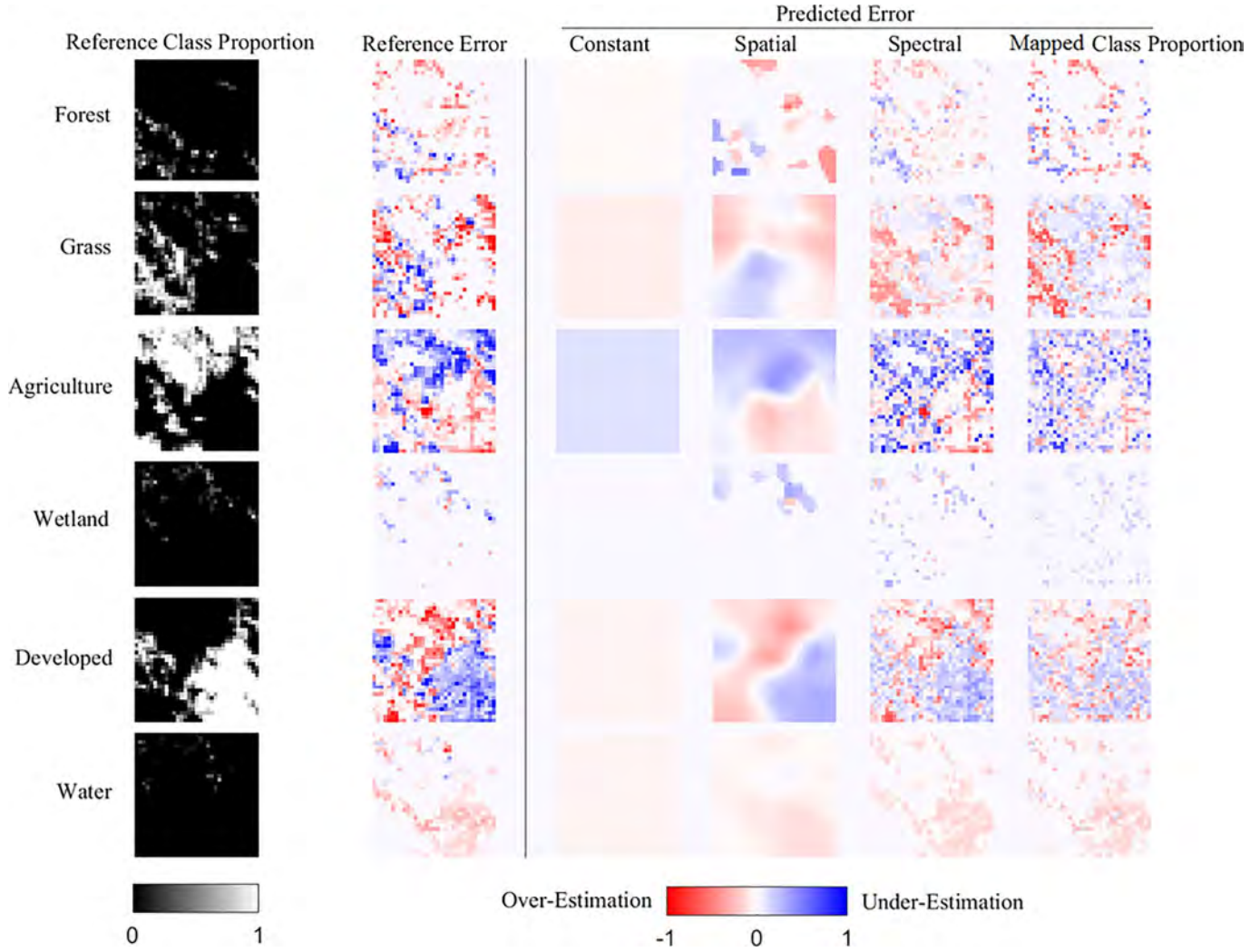

Fig. 6. Reference class proportion, reference error, and predicted error maps (from 100 pixel test sample size) for six classes of the Trends block C in Fig. 1.

left, separated with a vertical solid line from the next four boxplots, show the results from classes that had a reference proportion P smaller than 0.01 (or 1%) of their given block. Each boxplot displays the median, the 25th and 75th percentiles, and the "whiskers" of the boxplot extend to minimum and maximum MAE values excluding outliers. Outliers are depicted with red crosses, where outliers are defined as values that are 1.5 times the interquartile range (i.e. the difference between 75th and 25th percentiles) smaller than the 25th percentile or greater than the 75th percentile. The first and second rows of symbols on the top of each boxplot in Fig. 3 show the paired *t*-test results comparing the new interpolation methods (i.e., interpolations in spectral and mapped class proportion domains) to the two existing benchmark methods (i.e., constant and spatial interpolation methods), respectively.

Generally, errors of prediction increase as the reference class proportion P increases, reaching a maximum at about $P = 0.5$ (or 50%) and then decreasing when $P > 0.5$. This pattern exists for all four error prediction methods. In the vast majority of cases, the new interpolation methods outperform the two benchmark methods. The improvement in MAE achieved by the new methods over the benchmark methods is generally greater for common classes (i.e., the larger values of P on the right side of the figure). In addition, the variability of the MAE values decreases for the new methods for very low class proportion groups ($P < 0.01$) and is substantially smaller than the variability of the benchmark methods. The performance of the two new methods is very similar with a slight advantage for the spectral method, especially for the two larger test sample sizes. Between the two benchmark methods spatial interpolation clearly outperforms the constant method.

The effect of sample size of the test data is that MAE decreases as sample size increases for all methods except the constant benchmark method (Fig. 3, and see appendix Table S1 for mean MAE values). For example, for the group of the most common classes ($0.60 < P <= 0.85$), MAE decreases by around 0.03 for the three interpolation methods as sample size increases from 25 to 100 whereas MAE of the constant method decreases by 0.003 (Table S1). Change in sample size generally did not affect the relative performance of the different interpolation methods, as the relative advantage of the spectral and mapped class proportion domains was constant for all sample sizes.

The predicted error maps can also be compared to the reference error maps based on visual inspection. To illustrate typical results, Fig. 4–6 present reference error maps and the predicted error maps for three of the Trends blocks. Ideally, a predicted error map should be similar to its corresponding reference map. Error maps from the constant method use the estimated mean error of the class for the entire block and do not provide any information on the spatial distribution of classification error. The spatial interpolation method captures the large-scale patterns of classification error but fails to detect local variation when error values vary substantially in space. For example, the spatial interpolation for Grass in Fig. 4, Water in Fig. 5, and Developed in Fig. 6 capture the large scale variation of classification error. However, for classes such as Developed in Fig. 5 or Agriculture in Fig. 6 spatial

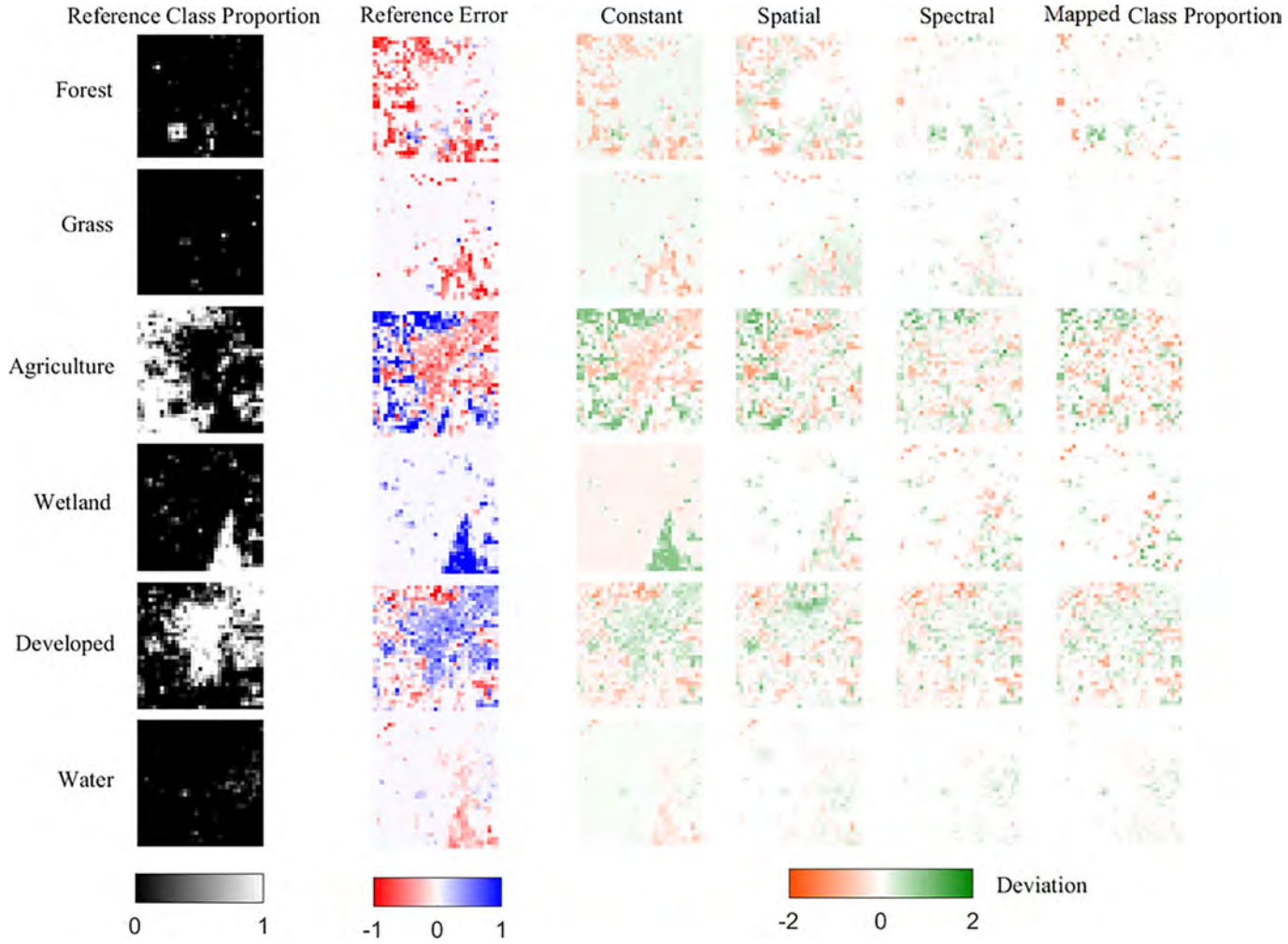

Fig. 7. Reference class proportion, reference error, and deviation of predicted error from reference error maps (from 100 pixel test sample size) for six classes of a sample Trends block (block B in Fig. 1).

interpolation fails to capture fine resolution variation of classification error due to a high degree of mixing of over/under-estimation errors.

Relative to the benchmark methods, the predictions from the new spectral and mapped class proportion methods are more similar to the reference error maps as the new methods better depict the fine resolution variation of classification error. For example, according to the reference error map of the Agriculture in Fig. 4, there is a mixture of over and under-estimation errors on the right side of the block. The two new methods capture this pattern to some extent but the spatial interpolation method fails to identify this fine resolution variation. Other examples are Forest in Fig. 5 and Water in Fig. 6 in which classification error, which is mainly overestimation, is clustered in some regions. The predicted patterns from the new methods are very similar to their corresponding reference error maps while in the spatial interpolations those variations are smoothed out.

The reference error maps in Fig. 4–6 clearly show that misclassification rates can vary substantially over different regions of a classified map. For many classes shown in Fig. 4–6 the class proportion is over-estimated in some regions and under-estimated in other regions. The Developed class in Fig. 4 shows an example of the importance of error mapping. According to the constant method the mean error is small which would imply very good classification, but the spatial variation of error is considerable. The class proportion is over-estimated on the left side of the block and under-estimated on the right side of the block. By taking into account where the Developed regions actually are located based on the corresponding reference class proportion map, one can see that the variation in over/under-estimation errors exists both between Developed and non-Developed regions (between-class variation) and between different areas of Developed regions (within-class variation). Another example would be the Agriculture class in Fig. 5 where even though the mean error is small, the class is substantially under-estimated for regions where Agriculture is actually present and over-estimated for regions where Agriculture is absent.

6. Discussion and conclusions

Numerous studies are dependent on accurate land-cover maps, for example climate modeling (Anav et al., 2010; Brovkin et al., 2013; Cuo et al., 2011; Lawrence et al., 2012); biodiversity studies (Bremer & Farley, 2010; Joseph et al., 2009; Rojas et al., 2013; Zimmermann et al., 2010); biomass estimation (Avitabile et al., 2012; Fang et al., 2013; Zheng et al., 2004); carbon measurement and modeling (Achard et al., 2004; Jung et al., 2006; Quaife et al., 2008; Ramankutty et al., 2007; Riley et al., 1997); and food security studies (Brown, 2016; Moore et al., 2012; Vermeulen et al., 2012). Figs. 4–6 clearly show that soft classification error varies over space. Summary accuracy measures fall short on communicating this spatial variability, whereas our proposed individual pixel error maps provide the desired visualization of errors needed to comprehend spatial patterns of error. The individual pixel error maps also enable implementation of advanced error propagation techniques in interdisciplinary studies. The new error mapping methods proposed in this research are

straightforward to implement. In practice creation of the error maps requires only the sample of reference data that would be collected in any accuracy assessment. Therefore, the same test dataset currently used for error matrix construction and summary measures estimation can be used for our error interpolation method.

Spatial interpolations have been previously used for soft classification error mapping (Comber, 2013; Foody, 2005). In this research we examined two new explanatory domains, spectral and mapped class proportion domains. Spatial interpolation performed similar to or better than the constant method (Fig. 3), but spatial interpolation failed to capture fine resolution variation of error in many cases (Fig. 4–6). The new interpolations in the spectral and mapped class proportion domains outperformed both constant and spatial interpolation methods. Predictions by the new methods had smaller MAE (Fig. 3) and yielded more similarity to the reference error maps (Fig. 4–6) than the benchmark methods. The advantage of new methods over spatial interpolation is an important observation because spatial interpolation is familiar to many practitioners and may be the first method an analyst considers for error interpolation.

The advantage of the spectral domain over the spatial domain for error interpolation can be attributed to its capacity to distinguish different classes that could have different error rates. The spectral domain can also capture the within-class error variation to some extent (for example, see the Developed class in Fig. 4 in which error rates are different for the east and west sides of the map for pixels with large reference proportions of the Developed class). The fine resolution variation of classification error is often obscured when spatial interpolation is used. Generally, because classification is conducted in the spectral domain errors would be expected to be more correlated in that domain than in the spatial domain. Therefore, interpolation in the spectral domain is recommended as the first choice for error interpolation of soft classifications.

Regarding the mapped class proportion domain, its performance was comparable to that of the spectral domain. This observation can be attributed to the fact that the mapped class proportion domain can include much of the information from the spectral domain. Basically, a soft classification can be seen as a continuous transformation function that maps pixels from an L- dimensional spectral domain to a C-dimensional mapped class proportion domain. Consequently, pixels with similar spectral values will have similar mapped class compositions and vice versa. Therefore, the mapped class proportion domain could include a substantial portion of information contained in the spectral domain. However, the magnitude of information that is carried from the spectral domain to the mapped class proportion domain is dependent on spectral input dimensionality. That is, when transforming from a high-dimensionality domain to a low-dimensionality domain some portion of information might be lost. In this research, both the spectral and mapped class proportion domains were six-dimensional. Further research is required to examine predictions when the number of spectral bands is much larger than the number of classes. In such cases, the mapped class proportion domain might contain limited information relative to that of the spectral domain due to smaller dimensionality. Moreover, the correlation between the spectral domain and the mapped class proportion domain can be affected by the soft classifier (linear versus nonlinear) which might affect error predictions.

Finally, one limitation of the spectral interpolation is that besides the classified maps and test data, the method requires having access to the image(s) (i.e., spectral domain) used for classification. Consequently, a major practical advantage of the mapped class proportion domain is that the interpolation in this domain does not require any information from the classification input space (e.g. imagery data). This is because soft classification maps supply the mapped class proportion values for all pixels and therefore, any user of the soft classification maps can create the error maps using interpolations in the mapped class proportion domain without having access to the original classified image. In addition, the reference class proportion values are only required for the test sample data which is the requirement of any classification accuracy method. When spectral domain information is not available, based on the results of this research the mapped class proportion domain is preferred over the spatial domain for soft classification error interpolation. In this research the same linear kernel interpolation method was used for all the three domains to create a fair comparison between the different domains (i.e., not confounded by different interpolation functions). Future research extending these comparisons to different interpolation functions such as kriging would be a recommended next step in evaluating the different explanatory domains.

For any prediction modeling task, the spatial distribution of prediction error is important. A desired property of the prediction error would be constant variance. In the context of error map prediction, model errors are actually deviations of the predicted error maps from the reference error map (i.e., “error of error prediction”). Fig. 7 shows the distribution of these deviations for the sample block in Fig. 5. Note that because the reference and predicted errors can have values from -1 to 1, the deviations can have values between -2 to 2, where -2 would be the case where reference error is -1(100% over-estimation) and predicted error is 1 (100% under-estimation). A value of 2 for deviation would be the case where reference error is 1(100% under-estimation) and predicted error is -1 (100% over-estimation). The deviation maps for the four methods were obtained by subtracting their predicted error maps from the corresponding reference error maps. Ideally, these deviations should be distributed randomly in spatial domain with constant variance and mean equal to zero. Randomness is more evident for the new methods while spatial clustering can be seen in the constant and spatial benchmark method maps (Fig. 7).

To summarize, map users can utilize error maps in different ways. For example, the error maps can contribute to deciding if a land-cover map is accurate enough for an intended application within a region of interest. The error maps can also be used to diagnose and improve a land-cover soft classification thorough identification of locations where improvement in land-cover proportion estimation is necessary. In addition, error maps can provide map producers with the information on classes that are confused by the classifier and the location of confusions. For example, based on spatial pattern of error distribution in Fig. 5, it can be concluded that Grass and Water classes are confused with the Wetland class (this can be observed in the reference error maps and the error maps from the new methods, but it is not clear in the benchmark methods). Improvements may be achieved by revising the classification process or adding training data from classes/areas with larger error. In addition, if the spectral signature of a given class varies at different areas of the scene, due to changes in, for example, topography or vegetation health/species, this may result in spatial variation in classification error. These variations may be captured by the error maps and be considered in the classification process. Lastly, the local land-cover error information provided by the error maps may be used for advanced data fusion, assimilation, and modeling. The new methods introduced in this research can improve applications of error maps by producing more accurate error predictions relative to previous methods.

Further research would be necessary to examine if the error maps can be directly used to improve soft classifications. Because the error maps are per-pixel and per-class estimates of soft classification errors, one might consider subtracting those error maps from the corresponding land-cover soft classification maps to improve them. However, to do so, it is necessary to make sure that the error maps are of acceptable quality. This can be achieved by having a large enough test sample size and evaluating the quality of error prediction, for example through estimating cross-validation or out-of-bag error while producing the error maps. If the precision of error prediction is relatively small, compared to the classification error itself, the error maps can be used to adjust the mapped class proportions, however further testing from an independent reference dataset may be needed. Otherwise, the error maps should be used to explore the relative

distribution of soft classifications error, i.e. over/under-estimation, rather than to initiate adjustments of soft classifications.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.rse.2017.07.028.

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.

References

Achard, F., Eva, H.D., Mayaux, P., Stibig, H., Belward, A., 2004. Improved estimates of net carbon emissions from land cover change in the tropics for the 1990s. Glob. Biogeochem. Cycles 18, 1–11 GB2008.
Anav, A., Ruti, P.M., Artale, V., Valentini, R., 2010. Modelling the effects of land-cover changes on surface climate in the Mediterranean region. Clim. Res. 41, 91–104.
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.
Avitabile, V., Baccini, A., Friedl, M.A., Schmullius, C., 2012. Capabilities and limitations of Landsat and land cover data for aboveground woody biomass estimation of Uganda. Remote Sens. Environ. 117, 366–380.
Binaghi, E., Brivio, P.A., Ghezzi, P., Rampini, A., 1999. A fuzzy set-based accuracy assessment of soft classification. Pattern Recogn. Lett. 20, 935–948.
Bioucas-Dias, J.M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., et al., 2012. Hyperspectral unmixing overview: geometrical, statistical, and sparse regression-based approaches. IEEE J. Select. Topic. Appl. Earth Observat. Remote Sens. 5, 354–379.
Bremer, L.L., Farley, K.A., 2010. Does plantation forestry restore biodiversity or create green deserts? A synthesis of the effects of land-use transitions on plant species richness. Biodivers. Conserv. 19, 3893–3915.
Brovkin, V., Boysen, L., Arora, V.K., Boisier, J.P., Cadule, P., Chini, L., et al., 2013. Effect of anthropogenic land-use and land-cover changes on climate and land carbon storage in CMIP5 projections for the twenty-first century. J. Clim. 26, 6859–6881.
Brown, M.E., 2016. Remote sensing technology and land use analysis in food security assessment. J. Land Use Sci. 1–19.
Campbell, J.B., 1981. Spatial correlation effects upon accuracy of supervised classification of land cover. Photogramm. Eng. Remote. Sens. 47, 355–363.
Carpenter, G.A., Gopal, S., MacOmber, S., Martens, S., Woodcock, C.E., 1999. A neural network method for mixture estimation for vegetation mapping. Remote Sens. Environ. 70, 138–152.
Carreiras, J.M.B., Vasconcelos, M.J., Lucas, R.M., 2012. Understanding the relationship between aboveground biomass and ALOS PALSAR data in the forests of Guinea-Bissau (West Africa). Remote Sens. Environ. 121, 426–442.
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.
Chen, J., Zhu, X., Imura, H., Chen, X., 2010. Consistency of accuracy assessment indices for soft classification: simulation analysis. ISPRS J. Photogramm. Remote Sens. 65, 156–164.
Cihlar, J., 2000. Land cover mapping of large areas from satellites: status and research priorities. Int. J. Remote Sens. 21, 1093–1114.
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 Sens. Lett. 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.
Cuo, L., Beyene, T.K., Voisin, N., Su, F., Lettenmaier, D.P., Alberti, M., et al., 2011. Effects of mid-twenty-first century climate and land cover change on the hydrology of the Puget sound basin, Washington. Hydrol. Process. 25, 1729–1753.
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.
Dong, J., Kaufmann, R.K., Myneni, R.B., Tucker, C.J., Kauppi, P.E., Liski, J., et al., 2003. Remote sensing estimates of boreal and temperate forest woody biomass: carbon pools, sources, and sinks. Remote Sens. Environ. 84, 393–410.
Eva, H.D., Achard, F., Beuchle, R., de Miranda, E., Carboni, S., Seliger, R., et al., 2012. Forest cover changes in tropical south and central America from 1990 to 2005 and related carbon emissions and removals. Remote Sens. 4, 1369–1391.
Fang, H., Li, W., Myneni, R.B., 2013. The impact of potential land cover misclassification on modis leaf area index (LAI) estimation: a statistical perspective. Remote Sens. 5, 830–844.
Finn, J.T., 1993. Use of the average mutual information index in evaluating classification error and consistency. Int. J. Geogr. Inf. Syst. 7, 349–366.
Foody, G.M., 1995. Cross-entropy for the evaluation of the accuracy of a fuzzy land cover classification with fuzzy ground data. ISPRS J. Photogramm. Remote Sens. 50, 2–12.
Foody, G.M., 1996. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. Int. J. Remote Sens. 17, 1317–1340.
Foody, G.M., 1999. The continuum of classification fuzziness in thematic mapping. Photogramm. Eng. Remote. Sens. 65, 443–451.
Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sens. Environ. 80, 185–201.
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., Cox, D.P., 1994. Sub-pixel land cover composition estimation using a linear mixture model and fuzzy membership functions. Int. J. Remote Sens. 15, 517–520.
Foody, G.M., Arora, M.K., 1996. Incorporating mixed pixels in the training, allocation and testing stages of supervised classifications. Pattern Recogn. Lett. 17, 1389–1398.
Foody, G.M., Doan, H.T.X., 2007. Variability in soft classification prediction and its implications for sub-pixel scale change detection and super resolution mapping. Photogramm. Eng. Remote. Sens. 73, 923–933.
Franklin, S.E., Wulder, M.A., 2002. Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. Prog. Phys. Geogr. 26, 173–205.
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.
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.
Gómez, D., Biging, G., Montero, J., 2008. Accuracy statistics for judging soft classification. Int. J. Remote Sens. 29, 693–709.
Gopal, S., Woodcock, C., 1994. Theory and methods for accuracy assessment of thematic maps using fuzzy sets. Photogramm. Eng. Remote Sens. 60, 181–188.
Grimm, N.B., Faeth, S.H., Golubiewski, N.E., Redman, C.L., Wu, J., Bai, X., et al., 2008. Global change and the ecology of cities. Science 319, 756–760.
Guerschman, J.P., Hill, M.J., Renzullo, L.J., Barrett, D.J., Marks, A.S., Botha, E.J., 2009. Estimating fractional cover of photosynthetic vegetation, non-photosynthetic vegetation and bare soil in the Australian tropical savanna region upscaling the EO-1 Hyperion and MODIS sensors. Remote Sens. Environ. 113, 928–945.
Homer, C.G., Aldridge, C.L., Meyer, D.K., Schell, S.J., 2012. Multi-scale remote sensing sagebrush characterization with regression trees over Wyoming, USA: laying a foundation for monitoring. Int. J. Appl. Earth Obs. Geoinf. 14, 233–244.
Iordache, M., Bioucas-Dias, J.M., Plaza, A., 2011. Sparse unmixing of hyperspectral data. IEEE Trans. Geosci. Remote Sens. 49, 2014–2039.
Jin, H., Stehman, S.V., Mountrakis, G., 2014. Assessing the impact of training sample selection on accuracy of an urban classification: a case study in Denver, Colorado. Int. J. Remote Sens. 35, 2067–2081.
Joseph, S., Blackburn, G.A., Gharai, B., Sudhakar, S., Thomas, A.P., Murthy, M.S.R., 2009. Monitoring conservation effectiveness in a global biodiversity hotspot: the contribution of land cover change assessment. Environ. Monit. Assess. 158, 169–179.
Jung, M., Henkel, K., Herold, M., Churkina, G., 2006. Exploiting synergies of global land cover products for carbon cycle modeling. Remote Sens. Environ. 101, 534–553.
Keshava, N., Mustard, J.F., 2002. Spectral unmixing. IEEE Signal Process. Mag. 19, 44–57.
Khan, S.I., Hong, Y., Wang, J., Yilmaz, K.K., Gourley, J.J., Adler, R.F., et al., 2011. Satellite remote sensing and hydrologic modeling for flood inundation mapping in Lake Victoria basin: implications for hydrologic prediction in ungauged basins. IEEE Trans. Geosci. Remote Sens. 49, 85–95.
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.
Khatami, R., Mountrakis, G., Stehman, S.V., 2017. Mapping per-pixel predicted accuracy of classified remote sensing images. Remote Sens. Environ. 191, 156–167.
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.
Laba, M., Gregory, S.K., Braden, J., Ogurcak, D., Hill, E., Fegraus, E., et al., 2002. Conventional and fuzzy accuracy assessment of the New York gap analysis project land cover map. Remote Sens. Environ. 81, 443–455.
Latifovic, R., Olthof, I., 2004. Accuracy assessment using sub-pixel fractional error matrices of global land cover products derived from satellite data. Remote Sens. Environ. 90, 153–165.
Lawrence, P.J., Feddema, J.J., Bonan, G.B., Meehl, G.A., O'Neill, B.C., Oleson, K.W., et al., 2012. Simulating the biogeochemical and biogeophysical impacts of transient land cover change and wood harvest in the community climate system model (CCSM4) from 1850 to 2100. J. Clim. 25, 3071–3095.
Lewis, H.G., Brown, M., 2001. A generalized confusion matrix for assessing area estimates from remotely sensed data. Int. J. Remote Sens. 22, 3223–3235.
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.
Lu, D., Weng, Q., 2006. Use of impervious surface in urban land-use classification. Remote Sens. Environ. 102, 146–160.
Lu, D., Weng, Q., 2007. A survey of image classification methods and techniques for improving classification performance. Int. J. Remote Sens. 28, 823–870.
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.
Maselli, F., Rodolfi, A., Conese, C., 1996. Fuzzy classification of spatially degraded thematic mapper data for the estimation of sub-pixel components. Int. J. Remote Sens.

17, 537–551.
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.
McMahon, G., 2007. Consequences of land-cover misclassification in models of impervious surface. Photogramm. Eng. Remote. Sens. 73, 1343–1353.
Mertens, K.C., Verbeke, L.P.C., Ducheyne, E.I., De Wulf, R.R., 2003. Using genetic algorithms in sub-pixel mapping. Int. J. Remote Sens. 24, 4241–4247.
Miao, L., Qi, H., 2007. Endmember extraction from highly mixed data using minimum volume constrained nonnegative matrix factorization. IEEE Trans. Geosci. Remote Sens. 45, 765–777.
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.
Moore, N., Alagarswamy, G., Pijanowski, B., Thornton, P., Lofgren, B., Olson, J., et al., 2012. East African food security as influenced by future climate change and land use change at local to regional scales. Clim. Chang. 110, 823–844.
Mountrakis, G., Xi, B., 2013. Assessing reference dataset representativeness through confidence metrics based on information density. ISPRS J. Photogramm. Remote Sens. 78, 129–147.
Nascimento, J.M.P., Dias, J.M.B., 2005. Vertex component analysis: a fast algorithm to unmix hyperspectral data. IEEE Trans. Geosci. Remote Sens. 43, 898–910.
Nie, W., Yuan, Y., Kepner, W., Nash, M.S., Jackson, M., Erickson, C., 2011. Assessing impacts of Landuse and Landcover changes on hydrology for the upper San Pedro watershed. J. Hydrol. 407, 105–114.
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.
Olthof, I., Fraser, R.H., 2007. Mapping northern land cover fractions using Landsat ETM +. Remote Sens. Environ. 107, 496–509.
Paneque-Gálvez, J., Mas, J., Moré, G., Cristóbal, J., Orta-Martínez, M., Luz, A.C., et al., 2013. Enhanced land use/cover classification of heterogeneous tropical landscapes using support vector machines and textural homogeneity. Int. J. Appl. Earth Obs. Geoinf. 23, 372–383.
Plaza, A., Martínez, P., Pérez, R., Plaza, J., 2004. A quantitative and comparative analysis of endmember extraction algorithms from hyperspectral data. IEEE Trans. Geosci. Remote Sens. 42, 650–663.
Plaza, A., Benediktsson, J.A., Boardman, J.W., Brazile, J., Bruzzone, L., Camps-Valls, G., et al., 2009. Recent advances in techniques for hyperspectral image processing. Remote Sens. Environ. 113, S110–S122.
Pontius Jr., R.G., Cheuk, M.L., 2006. A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. Int. J. Geogr. Inf. Sci. 20, 1–30.
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.
Ramankutty, N., Gibbs, H.K., Achard, F., Defries, R., Foley, J.A., Houghton, R.A., 2007. Challenges to estimating carbon emissions from tropical deforestation. Glob. Chang. Biol. 13, 51–66.
Réjou-Méchain, M., Muller-Landau, H.C., Detto, M., Thomas, S.C., Le Toan, T., Saatchi, S.S., et al., 2014. Local spatial structure of forest biomass and its consequences for remote sensing of carbon stocks. Biogeosciences 11, 6827–6840.
Ricotta, C., 2004. Evaluating the classification accuracy of fuzzy thematic maps with a simple parametric measure. Int. J. Remote Sens. 25, 2169–2176.
Riley, R.H., Phillips, D.L., Schuft, M.J., Garcia, M.C., 1997. Resolution and error in measuring land-cover change: effects on estimating net carbon release from Mexican terrestrial ecosystems. Int. J. Remote Sens. 18, 121–137.
Rojas, C., Pino, J., Basnou, C., Vivanco, M., 2013. Assessing land-use and -cover changes in relation to geographic factors and urban planning in the metropolitan area of Concepción (Chile). Implications for biodiversity conservation. Appl. Geogr. 39, 93–103.
Roy, D.P., Wulder, M.A., Loveland, T.R., W., C.E., Allen, R.G., Anderson, M.C., et al., 2014. Landsat-8: science and product vision for terrestrial global change research. Remote Sens. Environ. 145, 154–172.
Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., et al., 2010. Investigating soil moisture-climate interactions in a changing climate: a review. Earth Sci. Rev. 99, 125–161.
Silván-Cárdenas, J.L., Wang, L., 2008. Sub-pixel confusion-uncertainty matrix for assessing soft classifications. Remote Sens. Environ. 112, 1081–1095.
Sorooshian, S., Aghakouchak, A., Li, J., 2014. Influence of irrigation on land hydrological processes over California. J. Geophys. Res. D: Atmos. 119, 13137–13152.
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.
Stehman, S.V., Czaplewski, R.L., 1998. Design and analysis for thematic map accuracy assessment: fundamental principles. Remote Sens. Environ. 64, 331–344.
Stehman, S.V., Arora, M.K., Kasetkasem, T., Varshney, P.K., 2007. Estimation of fuzzy error matrix accuracy measures under stratified random sampling. Photogramm. Eng. Remote. Sens. 73, 165–173.
Story, M., Congalton, R.G., 1986. Accuracy assessment: a user's perspective. Photogramm. Eng. Remote Sens. 52, 397–399.
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.
Tatem, A.J., Lewis, H.G., Atkinson, P.M., Nixon, M.S., 2002. Super-resolution land cover pattern prediction using a Hopfield neural network. Remote Sens. Environ. 79, 1–14.
Townsend, P.A., 2000. A quantitative fuzzy approach to assess mapped vegetataion classifications for ecological applications. Remote Sens. Environ. 72, 253–267.
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.
Tsutsumida, N., Comber, A., Barrett, K., Saizen, I., Rustiadi, E., 2016. Sub-pixel classification of MODIS EVI for annual mappings of impervious surface areas. Remote Sens. 8.
Vermeulen, S.J., Campbell, B.M., Ingram, J.S.I., 2012. Climate change and food systems. Annu. Rev. Environ. Resour. 37, 195–222.
Woodcock, C.E., Gopal, S., 2000. Fuzzy set theory and thematic maps: accuracy assessment and area estimation. Int. J. Geogr. Inf. Sci. 14, 153–172.
Wulder, M.A., White, J.C., Goward, S.N., Masek, J.G., Irons, J.R., Herold, M., et al., 2008. Landsat continuity: issues and opportunities for land cover monitoring. Remote Sens. Environ. 112, 955–969.
Xu, M., Watanachaturaporn, P., Varshney, P.K., Arora, M.K., 2005. Decision tree regression for soft classification of remote sensing data. Remote Sens. Environ. 97, 322–336.
Zheng, D., Rademacher, J., Chen, J., Crow, T., Bresee, M., Le Moine, J., et al., 2004. Estimating aboveground biomass using Landsat 7 ETM + data across a managed landscape in northern Wisconsin, USA. Remote Sens. Environ. 93, 402–411.
Zimmermann, P., Tasser, E., Leitinger, G., Tappeiner, U., 2010. Effects of land-use and land-cover pattern on landscape-scale biodiversity in the European alps. Agric. Ecosyst. Environ. 139, 13–22.