Remote Sensing of Environment 177 (2016) 89–100

Contents lists available at ScienceDirect

Remote Sensing of Environment

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

A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research

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 14 September 2015
Received in revised form 6 January 2016
Accepted 12 February 2016
Available online 21 February 2016

Keywords:
Remote sensing
Machine learning
Environmental monitoring
Land cover mapping
Multi-time/angle imagery
Texture
Indices
Support vector machines
Classification accuracy

ABSTRACT

Classification of remotely sensed imagery for land-cover mapping purposes has attracted significant attention from researchers and practitioners. Numerous studies conducted over several decades have investigated a broad array of input data and classification methods. However, this vast assemblage of research results has not been synthesized to provide coherent guidance on the relative performance of different classification processes for generating land cover products. To address this problem, we completed a statistical meta-analysis of the past 15 years of research on supervised per-pixel image classification published in five high-impact remote sensing journals. The two general factors evaluated were classification algorithms and input data manipulation as these are factors that can be controlled by analysts to improve classification accuracy. The meta-analysis revealed that inclusion of texture information yielded the greatest improvement in overall accuracy of land-cover classification with an average increase of 12.1%. This increase in accuracy can be attributed to the additional spatial context information provided by including texture. Inclusion of ancillary data, multi-angle and time images also provided significant improvement in classification overall accuracy, with 8.5%, 8.0%, and 6.9% of average improvements, respectively. In contrast, other manipulation of spectral information such as index creation (e.g. Normalized Difference Vegetation Index) and feature extraction (e.g. Principal Components Analysis) offered much smaller improvements in accuracy. In terms of classification algorithms, support vector machines achieved the greatest accuracy, followed by neural network methods. The random forest classifier performed considerably better than the traditional decision tree classifier. Maximum likelihood classifiers, often used as benchmarking algorithms, offered low accuracy. Our findings will help guide practitioners to decide which classification to implement and also provide direction to researchers regarding comparative studies that will further solidify our understanding of different classification processes. However, these general guidelines do not preclude an analyst from incorporating personal preferences or considering specific algorithmic benefits that may be pertinent to a particular application.

1. Introduction

Remote sensing science offers a unique environmental monitoring capability that covers extensive geographical areas in a cost efficient manner while capturing irreplaceable information on the Earth's land, atmosphere and oceans. Remote sensing products play an integral role in numerous applications, for example carbon emission monitoring (Birdsey et al., 2013; DeFries et al., 2002; Myneni et al., 2001; Schwalm et al., 2012), forest monitoring (Asner et al., 2006; Gong et al., 2013; Hansen et al., 2008, 2013; Myneni et al., 2007; Potapov et al., 2015; Townshend et al., 2012), medical science and epidemiology studies (Evans et al., 2013; Gilbert et al., 2008; Liu & Weng, 2012; Lobitz et al., 2000) land change detection (Giustarini et al., 2013; Grekousis, Mountrakis, & Kavouras, 2015; Hussain, Chen, Cheng, Wei, & Stanley, 2013; Lambin & Meyfroidt, 2011; Rindfuss, Walsh, Turner, Fox, & Mishra, 2004), natural hazard assessment (Fialko, Sandwell, Simons, & Rosen, 2005; Khatami & Mountrakis, 2012), agriculture and water/wetland monitoring (Alcantara, Kuemmerle, Prishchepov, & Radeloff, 2012; Anderson, Allen, Morse, & Kustas, 2012; Hong et al., 2012; Ogilvie et al., 2015), climate dynamics (Keegan, Albert, McConnell, & Baker, 2014; Knyazikhin et al., 2013; McMenamin, Hadly, & Wright, 2008; Syed, Famiglietti, Chambers, Willis, & Hilburn, 2010), and biodiversity studies (Asner et al., 2009; Mendenhall, Sekercioglu, Brenes, Ehrlich, & Daily, 2011; Nagendra & Gadgil, 1999; Skidmore et al., 2015).

* 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.2016.02.028
0034-4257/

Remote sensing image classification is the process that converts remotely sensed imagery to usable products. In this manuscript we focus on classification processes for creating land-cover maps. Land-cover mapping using satellite or airborne imagery has increased exponentially over the past decades, partially due to improved data availability and accessibility (Yu et al., 2014). Land-cover mapping is a complicated process with numerous factors influencing the quality of the final product. An image analyst has to select from a plethora of options including image type, classification algorithm, training/validation data, input features, pre- and post-processing techniques, ancillary data, and target classes. To make these decisions image analysts are typically drawing on their individual experience and expertise as opposed to the collective knowledge of the field.

The remote sensing community has undertaken considerable efforts to improve land-cover map accuracy. The majority of published research papers demonstrate the validity of their suggested improvements by comparing the accuracy of the proposed classification processes with that of an existing process. Due to the considerable work associated with reference data creation and the limited scope of most studies, the accuracy results reported in these studies are commonly limited to single sites with testing performed on reference data from a single image. Such comparisons are too limited to infer general guidelines for selecting a suitable process to produce highly accurate maps (Stehman, 2006). Moreover, in many cases different studies report conflicting results even when comparing similar classification methods, and inferring general recommendations from these individual studies in isolation is challenging. Consequently, questions such as "*Which classification process is the most promising among a set of processes*?" and "*What is the expected improvement in accuracy*?" have not been answered despite extensive work on classification methods. The objective of our research is to synthesize the collective knowledge of the remote sensing community, as represented by results in peer-reviewed journal articles, to identify which classification processes offer the most promising improvements in accuracy of supervised pixel-based land-cover classification. The analysis is focused on two general factors that can be controlled by analysts to improve classification accuracy, classification algorithms and input data manipulation.

Past review articles have provided useful descriptive summaries of methods and procedures of image classification. For example, Lu and Weng (2007) and Weng (2012) discussed details of major image classification approaches and their main steps, classification accuracy improvement techniques and issues affecting classification performance. Cihlar (2000) and Franklin and Wulder (2002) investigated mapping strategies used in large area land-cover classification and related issues such as multi-time/angle and multiple sensors, geometric processing, and radiometric scaling. Smits, Dellepiane, and Schowengerdt (1999) introduced a quality assessment protocol for land-cover mapping in the context of project requirements and economic cost of error. Wilkinson (2005) utilized scatter plots to investigate the relationship between classification accuracy and date of publication, number of classes, dimensions of feature space, spatial resolution, size of study area and neural network classifiers. Yu et al. (2014) also used scatter plots to investigate the relationship between classification accuracy and date of publication, size of study area, and classification system complexity and report the estimated average overall accuracy and its corresponding standard error for different sensors and classification algorithms. These previous review studies focused on descriptive results and direct comparisons were not targeted to attribute accuracy improvements to individual features of the classification process. A distinguishing feature of the meta-analysis that we implemented is that we synthesized the results of those studies that provided direct one-to-one comparisons of different classification processes. Consequently, our matched-pairs analysis controls for potential confounding factors such as different sites, legends, landscape complexities and reference data that would complicate our ability to fairly compare performance of different classification processes.

2. Protocol for selecting sample of articles

The comparisons among classification processes presented in this work were extracted from peer-reviewed articles published between 1998 and 2012 in five high-impact remote sensing journals: Remote Sensing of Environment, ISPRS Journal of Photogrammetry and Remote Sensing, IEEE Transactions on Geoscience and Remote Sensing, International Journal of Remote Sensing, and Photogrammetric Engineering and Remote Sensing. To identify recent findings while keeping the workload at manageable levels, our search focused on articles within the fifteen-year period, from 1998 to 2012. Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher et al., 2009) was followed for article selection. Fig. 1 describes the selection process (see appendix Table S1 for a detailed PRISMA statement). The following criteria were applied to select relevant articles:

(1) Articles were limited to land-cover mapping. Articles using non-spatial images, images not covering the earth's surface, or simulated data were not included.
(2) Only articles containing supervised per-pixel classification techniques were included. "Soft" classification techniques, where land-cover proportions were estimated for each pixel, with hardened results (i.e. assigning a single label to each pixel) were included.
(3) Articles were required to contain two or more classification processes of the same image(s) using the same training dataset where the only differentiating factor was that either two different classification algorithms were used or an input data enhancement method was added to the first classification process. This allowed isolation of any effect in overall accuracy to a single contributing factor. This is a very important point that was critical in article selection process. Differences in classification tasks including different sites or images, target classes, landscape complexities, and reference data can affect performance of the classification processes. If the classifiers were not applied to the same case study it would be difficult to determine if the differences in overall accuracies were due to the performance of the classification processes or because they had been applied to two different classification scenarios with different levels of difficulty. Consequently, the selected articles and analyses were limited to those studies that compared two or more classification processes based on the same case study (i.e., same image(s), training and test data, and target classes).
(4) Articles included a quantitative accuracy assessment that reported overall accuracy (OA). OA was selected over other accuracy measures because it is most frequently reported and thus would result in a larger sample size of articles.
(5) Accuracy assessment results were based on reference data that were independent of data used in the training phase of the classification.
(6) Accuracy assessment results were based on per-pixel comparisons between the map labels and the reference labels.

The Scopus database reported 15,913 articles published by the five aforementioned journals over the 1998–2012 year period. An automated general query was designed to remove most of the unrelated articles and to extract articles that were more likely to satisfy the selection criteria. Multiple queries were tested by trial and error on some randomly sampled journal issues. Queries were applied on article title, abstract, and keywords. Recall of queries and number of articles returned were considered to determine the appropriate query. Recall was defined as the percent of the articles that could be used in the research returned by the query. The final query was as shown in Table 1 where "OR" operator was applied among expressions inside each column.

The automated query scanned the 15,913 articles and returned 2410 articles. These 2410 articles were then manually examined and 266

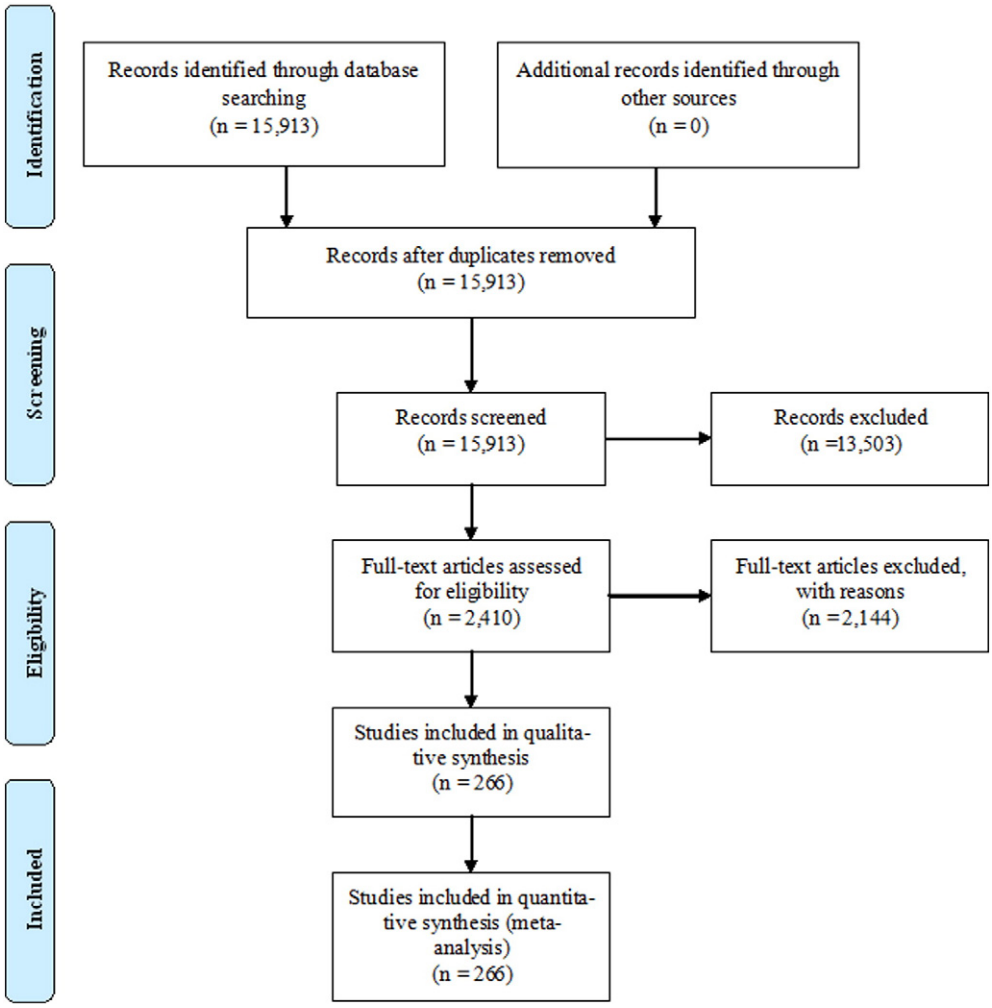

Fig. 1. PRISMA flow diagram for manuscript selection.

were found to satisfy the selection criteria. Of these 266 articles, 157 articles included comparisons of classification algorithms and 162 articles included comparisons of data enhancement methods (articles may include both categories).

A database of comparisons was constructed based on the 266 articles. Each row of the database corresponded to one extracted comparison among classification processes in one article. A comparison was included if two or more classifications were evaluated in the same case study, which the classifications differed by only one factor (either different classification algorithms were used or an input data enhancement method was added to the first classification process). For example, if in one article the same image was classified once with a maximum likelihood classifier and once with a neural network classifier, these two classifications formed one row (i.e., one comparison) in the database. It is possible that an article included several comparisons, as for example, when several classification processes were applied to different case studies. In addition, if an article included several versions of the same general classification process used to classify the same input data, only the version with the highest OA was used in the meta-analysis. For example, a spectral band selection study could investigate different combinations of bands or a study of a neural network classifier could include different numbers of hidden layers.

Table 1
Article search query design. Query: [A OR (B AND C)] AND D AND NOT E.

A	B	C	D	E
Accuracy	Comparison	Results	Class	Regression
Accurate	Improvement	Result	Classes	Unsupervised
Accuracies	Improve		Classification	Subpixel
Certainty	Improves		Classifier	"Sub-pixel"
"Error matrix"			Classifications	
"Error matrices"			Classifiers	
"Confusion matrix"			Mapping	
"Confusion matrices"			Training	

3. Features of the classification process evaluated

This research focused on two primary features of the classification process. For the analysis of different classification algorithm families, the classifications belonging to the same classifier family were grouped together. For example, all Support Vector Machines (SVM) (Cortes & Vapnik, 1995; Mountrakis, Im, & Ogole, 2011) variants comprised the SVM category. This aggregation of techniques created sufficient sample sizes of articles to represent the classifier families while preserving the major algorithmic differences. The second major component of the analysis was the assessment of the procedures used to enhance the input data to the classification algorithm. The input data enhancement categories included the following:

– *Texture*: This category included comparisons where in the second classification process texture layers were added as additional explanatory variables to the first classification. Incorporation of texture is a popular technique for classification improvement (Coburn & Roberts, 2004; Ji, 2000; Shackelford & Davis, 2003). Texture is

related to frequency of tonal change in an image and represents the degree of roughness of land-cover materials (Franklin & Wulder, 2002). Usually, a texture value for a pixel is calculated using a mathematical function incorporating tonal values inside a window centered at that pixel. Selection of appropriate window size is critical to texture extraction. The ability of texture to discriminate different land covers depends on the image spatial resolution and structural differences within and between land-cover classes (Chen, Stow, & Gong, 2004).

- *Ancillary data*: This category included comparisons where in the second classification process ancillary data or information extracted from these data were incorporated directly as additional classification explanatory variables to the first classification. Use of ancillary data to pre-process an image did not fall under this category. Numerous ancillary datasets can enhance classification performance. Common examples include topographic data such as digital elevation models, slope, aspect layers (Carpenter et al., 1999; Chang et al., 2007; Schmidt et al., 2004; Sesnie, Gessler, Finegan, & Thessler, 2008), geological layers (Sluiter & Pebesma, 2010), data from active sensors such as synthetic aperture radar (Held, Ticehurst, Lymburner, & Williams, 2003; Kuplich, Freitas, & Soares, 2000) or LiDAR (Geerling, Labrador-Garcia, Clevers, Ragas, & Smits, 2007), and data from passive sensors (Su, Chopping, Rango, Martonchik, & Peters, 2007a; Watts, Powell, Lawrence, & Hilker, 2011).
- *Multi-time imagery*: This category compared single image classification to the classification of fusion of images captured at different times from the same site (Brown De Colstoun et al., 2003; Carrão, Gonçalves, & Caetano, 2008; Del Frate et al., 2003; Guerschman, Paruelo, Di Bella, Giallorenzi, & Pacin, 2003; Sedano, Gong, & Ferrão, 2005).
- *Multi-angle imagery*: This category included comparisons of single image classification to classification of fusion of images captured from different perspectives from the same site at approximately the same date (Duca & Del Frate, 2008; Heikkinen, Korpela, Tokola, Honkavaara, & Parkkinen, 2011; Su, Chopping, Rango, Martonchik, & Peters, 2007b).
- *Image pre-processing*: This category included comparisons in which in the second classification process an image pre-processing was added to the first classification process. Image pre-processing is a common way to improve classification accuracy. Radiometric correction is a very commonly used pre-processing technique. Atmospheric correction (Chavez, 1996; Fukushima et al., 1998; Huang, Gong, Clinton, & Hui, 2008; Richter, 1997; Song, Woodcock, Seto, Lenney, & Macomber, 2001) is a radiometric correction process that reduces atmosphere effects from observed radiance. Spatial low pass or smoothing filters are additional methods to decrease intra-class variability and random noise (Rio & Lozano-García, 2000; Tottrup, 2004). Examples of other radiometric corrections include continuum removal (Clark & Roush, 1984; Youngentob et al., 2011) and sun distance and elevation correction (Lillesand et al., 2004, Chapter 7). Pan sharpening is an example of pre-processing (Taylor, Kumar, & Reid, 2010) along with geometric corrections to normalize for topographic effects (Colby & Keating, 1998; Ranson, Sun, Kharuk, & Kovacs, 2001; Ricchetti, 2000).
- *Spectral indices*: This category included comparisons where in the second classification process spectral indices were added as additional explanatory variables to the first classification. Spectral indices are frequently used in remote sensing image classification (Dash et al., 2007; Dymond, Mladenoff, & Radeloff, 2002; Mountrakis, Watts, Luo, & Wang, 2009; Tsai & Philpot, 2002) due to their ability to target specific land cover classes and to reduce the effect of variable band representation (e.g., due to illumination or atmospheric differences). Indices are usually calculated through arithmetic combinations of different spectral bands. In this research any combination of original bands based on a pre-determined formula was considered an index. The underlying principle is that indices can discriminate target classes if an appropriate combination of bands is selected. The Normalized Difference Vegetation Index (NDVI) (Carlson & Ripley, 1997), that is the normalized difference between near infrared and red bands, can be applied to delineate vegetated areas. NDVI was the most commonly observed index in the reviewed articles. Other examples are tasseled cap statistics, other vegetation indices, chromaticity filter, and spectral derivatives.
- *Feature extraction*: This category included comparison of classification of original image(s) bands to the classification of features extracted from those image(s). Feature extraction and band selection methods convert the feature space of the original image to a more efficacious one in which target classes are more separable. The main difference between feature extraction and the spectral indices category is that for feature extraction the combination of bands or the formulas to extract new features are not pre-determined. Generally, feature extraction is done for two purposes, to remove correlation among different bands, thereby allowing the classifiers to take advantage of the image's true dimensionality, and to accelerate the classification process by decreasing data volume. Feature extraction is usually done either by selecting the most discriminating bands from existing bands (Chan & Paelinckx, 2008; Clark, Roberts, & Clark, 2005; Thenkabail, Enclona, Ashton, & Van Der Meer, 2004), or by transforming data to a different, more efficient feature space. Linear transformation methods such as Principal Component Analysis (PCA) (Clark et al., 2005; Fauvel, Benediktsson, Chanussot, & Sveinsson, 2008) and Minimum Noise Fraction (MNF) (Mundt et al., 2005; Yang, Everitt, & Johnson, 2009) were the most common techniques of dimensionality reduction observed in the reviewed articles.

This general categorization of enhancement methods made it possible to have a sufficiently large sample of articles to conduct statistical tests. Other general enhancement methods such as post-classification processing and classifier fusion were not included in the analyses due to their small sample size.

4. Meta-analysis of classification processes

Our meta-analysis investigated independently two features of the classification process: 1) the image classification algorithms and 2) enhancement methods for the input data to the classification algorithm. The difference between the overall accuracies of two classification processes compared in an article was used as the statistical measure of the effect size evaluated by the meta-analysis. For the analysis of different classification algorithm families, this effect size equals the difference between overall accuracies of the two classifier families. Comparisons between two “classifier families” were based on articles in which results were reported directly comparing two different classifiers using the same input dataset (e.g., the same image was classified once by a maximum likelihood classifier and then by a neural network classifier using identical training and validation datasets). Because different articles examined different sets of classifiers, it was not possible to define one classifier as a benchmark for all comparisons. Consequently, all possible pairwise comparisons of classifier families were evaluated. To distinguish the two classifiers of each pairwise comparison in the description of results, the classifier with the smaller mean OA (less accurate) is called the “first” classifier and the other more accurate classifier of the pair is identified as the “second” classifier.

For the assessment of input data enhancement categories, effect size equals the difference between overall accuracies of the process without enhancement (i.e., the baseline classification) and the process with the enhancement category. Again, comparisons were based on the sample of published articles in which for each study both the with- and without-enhancement classifiers were subject to identical training and validation datasets (e.g., the same image is classified by a fixed classifier once without pre-processing and once with pre-processing). A positive effect size implies an improvement in overall accuracy through addition

of the specified enhancement method relative to the baseline classification without the enhancement.

Several comparisons of different classification processes may be contained within the same article and in such cases it would be questionable to consider each of these comparisons as independent in the meta-analysis. Typically comparisons within the same article shared common input data and analyst(s). For those articles containing several comparisons, all comparisons from the article that belonged to the same category were replaced by their mean effect size (i.e., the difference in overall accuracy). For example, to estimate mean effect size obtained by "Texture" for the high spatial resolution sub-category (Table 4), all the comparisons for this sub-category were extracted, mean effect sizes within each article were estimated, and then the mean effect size over all articles was calculated. This approach is standard practice in meta-analysis in which a single measure from each study is contributed to each analysis.

Wilcoxon signed-rank median tests were used for both pairwise comparison of classifier families and examination of the isolated effect

Table 2
Pairwise comparison of overall accuracy (OA) of image classification algorithm families.

Classifier		Number of articles or sample size (number of articles with Second OA > First OA)	Mean OA of first classifier (%)	Effect size (%) (Second OA − First OA)				P-value of median test
First (less accurate)	Second (more accurate)			Mean (standard error)	First quartile	Median	Third quartile	
ML	SVM	30 (28)	75.3	6.6 (1.5)	1.9	5.0	7.6	<0.01
KNN	SVM	13 (11)	72.5	1.3 (3.9)	1.1	3.0	9.2	0.03
DA	SVM	9 (8)	78.2	5.3 (2.3)	1.6	2.9	12.4	0.06
DT	SVM	15 (13)	76.7	4.6 (2.2)	0.9	2.0	6.3	0.01
RF	SVM	9 (7)	74.2	1.8 (0.6)	0.1	1.3	3.6	0.04
NN	SVM	22 (14)	83.4	2.3 (1.3)	−0.2	0.8	3.0	0.04
MD	NN	13 (13)	70.1	12.5 (2.4)	6.7	11.0	19.7	<0.01
KNN	NN	11 (7)	79.9	0.8 (2.7)	−2.5	3.2	7.0	0.58
ML	NN	46 (36)	76.8	3.9 (1.0)	0.5	3.0	6.6	<0.01
DT	NN	15 (12)	77.4	2.7 (0.8)	0.5	1.6	5.5	0.01
DA	NN	6 (4)	81.0	2.3 (3.4)	−1.2	1.9	5.8	0.41
ML	KNN	17 (12)	78.7	1.3 (1.5)	−3.7	3.0	5.2	0.41
DT	KNN	9 (6)	66.9	2.4 (2.9)	−2.1	2.0	5.6	0.43
DA	KNN	6 (4)	75.7	1.5 (2.1)	−1.9	3.0	4.9	0.44
MD	KNN	5 (3)	73.4	8.8 (6.5)	−1.0	2.6	15.9	0.31
SAM	ML	10 (6)	74.4	6.9 (6.5)	−2.1	4.6	16.8	0.30
MD	ML	24 (19)	71.9	7.0 (2.6)	0.9	3.4	10.0	<0.01
PP	ML	5 (5)	58.7	21.8 (4.1)	16.9	23.9	28.1	0.06
ML	DT	15 (9)	78.2	1.1 (1.2)	−2.5	1.1	4.1	0.23
DT	RF	9 (9)	79.4	4.0 (0.8)	2.0	3.1	5.0	<0.01
DT	DA	9 (5)	77.3	0.5 (2.7)	−2.9	0.1	3.8	0.98
PP	MD	5 (5)	58.7	14.7 (3.2)	11.0	18.4	19.1	0.06
Comparisons with number of articles <5								
ML	IB	4 (3)	90.9	0.5 (3.0)	*	3.1	*	0.88
ML	DA	4 (3)	73.2	2.5 (1.8)	*	3.9	*	0.25
PP	NN	4 (4)	54.1	29.6 (8.2)	*	29.2	*	0.13
SS	SVM	4 (3)	78.3	1.3 (1.6)	*	2.7	*	0.88
MD	SAM	4 (2)	75.0	16.7 (15.4)	*	6.1	*	0.63
FZ	ML	3 (2)	72.7	6.0 (4.9)	*	9.5	*	0.50
ML	RF	3 (1)	79.0	0.5 (2.9)	*	−0.2	*	1.00
IS	NN	3 (2)	79.5	1.0 (3.1)	*	0.5	*	0.75
NN	RF	3 (1)	75.3	0.2 (2.8)	*	−2.0	*	1.00
KNN	RF	3 (1)	72.0	1.6 (3.2)	*	−0.3	*	1.00
MD	SVM	3 (3)	67.0	22.7 (12.0)	*	11.3	*	0.25
ML	SS	2 (1)	81.1	1.7 (2.9)	*	1.7	*	1.00
ML	IS	2 (2)	84.3	3.7 (1.3)	*	3.7	*	0.50
PP	KNN	2 (2)	69.0	16.9 (4.2)	*	16.9	*	0.50
SVM	SAM	2 (1)	84.8	0.1 (2.9)	*	0.1	*	1.00
PP	SVM	2 (2)	69.2	21.5 (7.6)	*	21.5	*	0.50
MD	DT	2 (2)	59.1	10.9 (2.1)	*	10.9	*	0.50
IS	DT	2 (1)	73.3	0.2 (5.2)	*	0.2	*	1.00
DA	RF	2 (2)	72.3	1.6 (0.5)	*	1.6	*	0.50
ML	GA	1 (1)	78.1	5.3 (–)	*	5.3	*	–
NN	GA	1 (1)	66.4	17.0 (–)	*	17.0	*	–
SAM	NN	1 (1)	69.0	17.0 (–)	*	17.0	*	–
NN	FZ	1 (1)	94.4	3.7 (–)	*	3.7	*	–
KNN	GA	1 (1)	83.1	0.2 (–)	*	0.2	*	–
KNN	SS	1 (1)	63.8	0.7 (–)	*	0.7	*	–
KNN	IS	1 (1)	81.1	11.0 (–)	*	11.0	*	–
MD	DA	1 (1)	41.0	30.0 (–)	*	30.0	*	–
MD	FZ	1 (1)	93.8	4.3 (–)	*	4.3	*	–
MD	IS	1 (1)	78.4	13.6 (–)	*	13.6	*	–
SAM	DA	1 (1)	38.5	25.3 (–)	*	25.3	*	–
PP	IS	1 (1)	60.0	32.0 (–)	*	32.0	*	–

H_0: Median OA of first classifier = Median OA of second classifier, H_a: Median OA of first classifier ≠ Median OA of second classifier. Abbreviations: Decision Tree (DT), Discriminant Analysis (DA), Fuzzy (FZ), Genetic Algorithm (GA), Immune System (IS), Index-Based (IB), K-Nearest Neighbor (KNN), Maximum Likelihood (ML), Minimum Distance (MD), Neural Network (NN), Parallelepiped (PP), Random Forest (RF), Spectral Angle Mapper (SAM), Subspace (SS), and Support Vector Machines (SVM). Results are shown for all comparisons even with the number of sampled articles <5 to provide an indication of results but with the recognition that there are too few studies to permit reliable assessment. The "–"indicates that information cannot be calculated. Because of small sample size the first and third quartiles were not estimated for comparisons in which the number of sampled articles was below 5 and these cases are shown as "*".

of each input data enhancement category. For pairwise comparison of classifier families, the null hypothesis was "equality of classifiers" and the alternative hypothesis was "inequality of classifiers". For the assessment of input data enhancement categories, the null hypothesis was "zero effect size" and the alternative hypothesis was "positive effect size". The alternative hypothesis was chosen to be one-sided to examine if a given enhancement category improved classification accuracy. For both cases, number of articles (sample size), mean OA of the first (less accurate) classification or baseline classification, mean, median, first and third quartiles of the effect size (difference in OA between two classification processes), standard error of mean effect size, and p-values of the Wilcoxon signed-rank median test are reported. A positive effect size represents the mean improvement in overall accuracy achieved by the second classifier (Table 2) or by the enhancement method (Tables 3 and 4). In addition, the number of articles or studies with mean effect size greater than zero is reported. This number corresponds to the article total where the second process performed better than the first process.

To investigate whether the effect size of a given enhancement category varied for different sensor conditions, sensors were categorized into Space-borne, Air-borne, and Active. Furthermore, Space-borne sensors were sub-categorized based on their spatial and spectral resolution. Spatial resolution was split into High (<10 m), Medium (10 m to 100 m), and Low (>100 m). Spectral resolution was categorized into High (more than 10 bands), Medium (between 5 to 10 bands), and Low (fewer than 5 bands) (Table 4).

Comparisons between different input data enhancement categories required accounting for the fact that the mean effect size of each enhancement method depended on the mean overall accuracy of the baseline classification for the case studies investigating that method. For example, if the case studies of one enhancement method had a mean OA of 90% for the baseline classification whereas another method had a mean OA of 70% for its baseline classification, the enhancement method with the 90% baseline mean OA would have more difficulty improving accuracy relative to this baseline. Because different enhancement methods were not implemented on the same set of case studies, comparison of effect sizes for different enhancement methods must be adjusted for different mean baseline overall accuracies. Analysis of Covariance (ANCOVA) using overall accuracy of the baseline classification as the covariate provides the needed adjustment. ANCOVA compares mean effect size for each enhancement method adjusted to a common baseline OA. Because the comparison may vary depending on the choice of OA to which the adjustment is made, we report the results of the ANCOVA for three levels of baseline overall accuracy, 70%, 80%, and 90% (Table 5). Thus the comparison of adjusted means evaluates differences in mean effect sizes if the first (baseline) classification has overall accuracy of 70%, 80%, and 90% (Table 5). Pairwise comparisons of input data enhancement categories based on the adjusted mean effect sizes were evaluated using the Tukey–Kramer test to control experimentwise type I error rate at $\alpha = 0.05$. Note that this ANCOVA approach was not needed in the comparison of classification algorithm families (Table 2) because of the matched-pair feature of those comparisons.

5. Results

5.1. Image classification algorithm

Because different articles compared different sets of classification algorithms, our meta-analysis of classifier families was conducted in a pairwise fashion. Table 2 presents the comparisons between pairs of classifier families. In Table 2, the effect size is equal to the difference between overall accuracies of two classifier families. A positive effect size implies an improvement in overall accuracy by the second classifier relative to the first classifier.

SVM algorithms outperformed every other classifier family in the head-to-head pairwise comparisons (Table 2). We elaborate on the details of the comparisons involving the SVM algorithms to illustrate how the results of the meta-analysis can be applied. SVM was directly compared to Maximum Likelihood (ML) (Strahler, 1980) in 30 case studies, and for 28 of those case studies SVM had higher overall accuracy than ML (Table 2). The conditions represented by these 30 case studies span a variety of different types of imagery, reference data, sites, legends, etc., but there is a clear tendency for SVM to outperform ML over this set of studies. The effect size was greater than 1.9% in 75% of the case studies (i.e., first quartile for improvement in OA was 1.9 indicating that OA of SVM exceeded OA of ML by 1.9% or more), and in 25% of the case studies the effect size was 7.6% or greater (based on the third quartile effect size of 7.6%). The Table 2 results thus provide quantitative information on the general tendencies of relative performance of the two classifier families, SVM and ML, in terms of how often the second family (SVM) outperformed the first family (ML) of the pair as effect size quantifies the potential magnitude of improvement. Clearly these results are not necessarily predictive of the relative performance of SVM and ML in any specific application. Any specific application is tantamount to a single case study and the relative performance would be specific to the unique features of that application. To further explore individual case studies, we have provided a table listing all case studies for each pairwise comparison with the effect size and citation provided. For example, in the example of comparing SVM to ML, we may want to further examine the two case studies in which SVM did not outperform ML or we may want to examine the 25% of the case studies for which SVM had an effect size of 7.6% or greater. The appendix table provides the information needed to identify the articles that have this information and the details of the conditions of these case studies can be extracted from the original sources.

As a second illustration of the use of Table 2, we consider the comparison of SVM with Neural Network (NN) (Rumelhart, Hinton, & Williams, 1986). Unlike SVM versus ML in which SVM had higher OA in 28 of 30 case studies, SVM had higher OA than NN in only 14 out of 22 case studies. The improvement in OA achieved by SVM relative to NN was generally small with a mean effect size of 2.3%. The median effect size was only 0.8%, and the effect size of SVM exceeded 3.0% in only 25% of the case studies. Although these results suggest that SVM generally tended to outperform NN, these two classifier families were much more similar in performance than were SVM and ML. Once again

Table 3
Summary and test results of input data enhancement categories.

Input data enhancement category	Number of articles or sample size (number of articles with Second OA > First OA)	Mean OA of first classification (%)	Effect size (%) (Second OA − First OA)				P-value of median test
			Mean (standard error)	First quartile	Median	Third quartile	
Texture	31 (31)	71.2	12.1 (1.8)	4.2	8.5	19.8	<0.01
Ancillary data	57 (52)	71.4	8.5 (1.1)	3.3	6.0	13.3	<0.01
Multi-angle imagery	5 (5)	66.7	8.0 (2.6)	3.6	6.0	13.1	0.03
Multi-time imagery	16 (14)	73.3	6.9 (1.3)	4.1	7.0	10.4	<0.01
Image pre-processing	28 (24)	74.0	4.8 (1.0)	0.9	3.9	7.5	<0.01
Spectral indices	8 (6)	73.5	2.4 (1.3)	0.0	0.8	4.7	0.04
Feature extraction	47 (28)	79.5	−0.2 (1.5)	−2.4	0.8	3.3	0.21

H_0: Median OA effect size = 0, H_a: Median OA effect size > 0.

Table 4
Summary and test results of input data enhancement categories partitioned based on sensor characteristics.

Input data enhancement category	Sensor characteristic	Number of articles or sample size (number of articles with Second OA > First OA)	Mean OA of first classification (%)	Effect size (%) (Second OA − First OA)				P-value of median test
				Mean (standard error)	First quartile	Median	Third quartile	
Texture	High spatial	11 (11)	77.6	9.6 (2.2)	5.1	8.0	13.4	<0.01
	Medium spatial	4 (4)	78.5	6.9 (3.3)	*	4.5	*	0.06
	Low spatial	– (–)	–	–	–	–	–	–
	Aerial	10 (10)	74.7	12.5 (3.9)	3.2	8.1	20.5	<0.01
	Active	4 (4)	41.6	21.0 (8.8)	*	19.4	*	0.06
	Low spectral	11 (11)	77.3	9.5 (2.8)	2.5	7.2	10.5	<0.01
	Medium spectral	6 (6)	73.1	11.5 (2.4)	5.8	11.3	15.0	0.02
	High spectral	1 (1)	85.7	3.3 (–)	*	3.3	*	–
Ancillary data	High spatial	6 (5)	73.7	4.1 (3.9)	2.5	3.2	7.1	0.16
	Medium spatial	27 (25)	72.1	7.4 (1.3)	2.8	5.5	10.1	<0.01
	Low spatial	3 (2)	84.2	1.7 (2.3)	*	2.8	*	0.38
	Aerial	17 (16)	70.0	8.9 (2.2)	4.4	6.2	13.3	<0.01
	Active	6 (5)	61.7	17.4 (3.7)	15.8	20.8	22.1	0.03
	Low spectral	13 (12)	76.5	5.9 (1.9)	3.5	4.6	7.7	0.01
	Medium spectral	23 (20)	70.6	8.0 (1.8)	0.9	5.9	11.4	<0.01
	High spectral	4 (4)	77.9	2.7 (0.7)	*	2.9	*	0.06
Multi-angle imagery	High spatial	– (–)	–	–	–	–	–	–
	Medium spatial	1 (1)	86.7	6.0 (–)	*	6.0	*	–
	Low spatial	3 (3)	57.1	9.9 (4.2)	*	12.3	*	0.13
	Aerial	1 (1)	75.7	4.1 (–)	*	4.1	*	–
	Active	– (–)	–	–	–	–	–	–
	Low spectral	2 (2)	50.4	14 (1.7)	*	14.0	*	0.25
	Medium spectral	1 (1)	70.3	1.8 (–)	*	1.8	*	–
	High spectral	1 (1)	86.7	6.0 (–)	*	6.0	*	–
Multi-time imagery	High spatial	– (–)	–	–	–	–	–	–
	Medium spatial	9 (9)	71.0	8.9 (1.0)	6.9	7.6	12.1	<0.01
	Low spatial	5 (4)	76.4	5.0 (2.9)	1.3	3.1	7.9	0.06
	Aerial	– (–)	–	–	–	–	–	–
	Active	2 (2)	57.0	10.7 (3.9)	*	10.7	*	0.25
	Low spectral	3 (2)	68.3	2.4 (8.1)	*	7.6	*	0.63
	Medium spectral	12 (11)	77.2	7.6 (1.3)	5.5	7.2	10.6	<0.01
	High spectral	1 (1)	62.3	2.0 (–)	*	2.0	*	–
Image pre-processing	High spatial	5 (3)	81.8	0.8 (2.2)	−1.7	0.7	4.6	0.44
	Medium spatial	9 (9)	70.5	7.3 (1.6)	3.3	5.0	12.0	<0.01
	Low spatial	– (–)	–	–	–	–	–	–
	Aerial	6 (6)	80.2	3.9 (1.6)	0.6	2.7	8.0	0.02
	Active	2 (2)	60.5	10.5 (6.9)	*	10.5	*	0.25
	Low spectral	6 (5)	67.7	2.4 (2.4)	0.7	1.7	5.8	0.16
	Medium spectral	11 (9)	69.2	5.9 (1.6)	1.5	4.8	11.2	<0.01
	High spectral	2 (1)	90.5	2.1 (2.1)	*	2.1	*	0.50
Spectral indices	High spatial	2 (2)	73.0	4.8 (4.5)	*	4.8	*	0.25
	Medium spatial	4 (3)	80.9	2.3 (1.5)	*	1.8	*	0.13
	Low spatial	1 (1)	68.9	1.3 (–)	*	1.3	*	–
	Aerial	1 (0)	49.6	−1.3 (–)	*	−1.3	*	–
	Active	– (–)	–	–	–	–	–	–
	Low spectral	3 (2)	83.5	4.1 (2.8)	*	3.3	*	0.25
	Medium spectral	3 (3)	73.0	2.2 (1.9)	*	0.3	*	0.13
	High spectral	1 (1)	68.9	1.3 (–)	*	1.3	*	–
Feature extraction	High spatial	2 (1)	81.2	0.1 (2.1)	*	0.1	*	0.50
	Medium spatial	6 (1)	77.6	−2.5 (2.8)	−3.3	−2.3	−0.1	0.89
	Low spatial	– (–)	–	–	–	–	–	–
	Aerial	31 (21)	80.9	0.8 (1.4)	−1.0	1.5	3.3	0.06
	Active	5 (5)	67.2	7.8 (5.5)	1.2	2.9	11.0	0.03
	Low spectral	1 (1)	79.0	7.0 (–)	*	7.0	*	–
	Medium spectral	4 (0)	78.1	−1.7 (0.6)	*	−2.0	*	1.00
	High spectral	3 (0)	79.8	−9.0 (3.1)	*	−9.6	*	1.00

H_0: Median OA effect size = 0, H_a: Median OA effect size > 0. The "–" means that information cannot be calculated or no sample is available. Because of small sample size the first and third quartiles were not estimated for comparisons in which the number of sampled articles was below 5 and these cases are shown as "*".

more specific information regarding the conditions contributing to differences in performance can be extracted from the case studies comparing SVM to NN that were used to derive the general tendencies of performance. For example, SVM's advantage over NN slightly increased with higher spatial resolution and higher spectral resolution (Fig. 2).

Table 5
Comparison of adjusted effect sizes (difference in overall accuracy) (%) for input data enhancement categories.

First classification OA (%)	Texture	Ancillary data	Multi-time imagery	Multi-angle imagery	Image pre-processing	Spectral indices	Feature extraction
70	12.5^{a}	9.0^{ab}	7.8^{abc}	7.1^{abc}	5.1^{bc}	2.5^{bc}	1.7^{c}
80	8.4^{d}	5.5^{d}	5.1^{de}	4.5^{de}	4.3^{de}	2.1^{de}	-0.3^{e}
90	4.3^{f}	2.0^{f}	2.4^{f}	1.9^{f}	3.5^{f}	1.8^{f}	-2.3^{f}

Within a row, means sharing a common superscript are not statistically significantly different (Tukey–Kramer pairwise comparisons controlling experimentwise type I error rate at $\alpha = 0.05$).

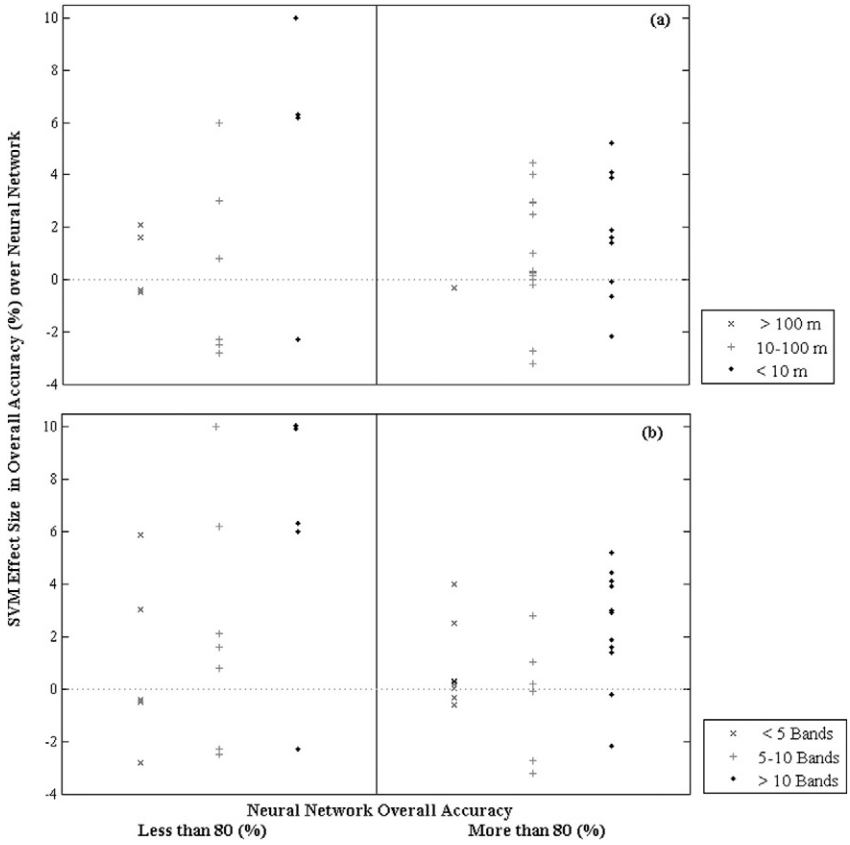

Fig. 2. Comparison of overall accuracy of SVM and NN by (a) spatial resolution (b) spectral resolution (number of bands). Note: Data points with positive effect size greater than 10% are plotted at 10% to maintain an informative scale for effect size.

A few other major results gleaned from the pairwise comparisons of classifier algorithm families are highlighted. The Random Forest classifier (RF) (Breiman, 2001), which is considered a Decision Tree (DT) (Brodley & Friedl, 1997) extension, performed better than DT in all nine articles in which they were compared with mean improvement in OA of 4.0%. Despite its simplicity, K-Nearest Neighbor (KNN) (Cover & Hart, 1967) generally performed as well or better than some of the more complex classifiers such as ML and DT. KNN was better than ML in 12 out of 17 comparisons (mean improvement of KNN of 1.3%) and KNN was better than DT in 6 out of 9 articles (mean improvement of 2.4%). This makes KNN a viable option when there are limitations of computational resources. Even though Minimum Distance has a similar concept to KNN, it did not work as well as KNN. The number of articles used for each pairwise comparison of classifier families shown in Table 2 provides a good indication of the strength of evidence for each comparison and identifies potential gaps in evidence in terms of absence of direct comparisons.

To illustrate results for some of the more important comparisons of classifier algorithm families, Fig. 3 displays scatterplots of the sample data for pairwise comparison of five main classifiers, ML, NN, KNN, SVM and DT. This figure illustrates the distribution of the overall accuracies for these pairs of classifiers and indicates the number of articles where one classifier works better than another. It can also help to interpret the magnitude of improvement for each sample article while considering the other classifier's accuracy.

5.2. Input data enhancement

The second part of our analysis investigated improvements in classification overall accuracy through different input data enhancement methods. Tables 3 and 4 provide the summary and test results of effect size of categories and sub-categories respectively. Some articles did not report spatial or spectral resolution and because many articles included multiple datasets, the number of samples of improvement categories in Table 3 is not equal to the sum of the number of samples of their subcategories in Table 4.

Except for feature extraction, all other data enhancement methods demonstrated statistically significant improvements in overall accuracy ($\alpha = 0.05$). The largest improvements resulted from inclusion of texture (mean of 12.1%), especially when spatial resolution increased (Table 4). All 31 sampled articles of inclusion of texture showed improvements in overall accuracy after adding textural information to their baseline classification processes. Inclusion of ancillary data, multi-angle imagery and multi-time imagery was the next most favorable input data enhancement methods yielding mean improvements of 8.5%, 8.0%, and 6.9% respectively. The three enhancement methods with the lowest effect size did not offer new sources of information. Pre-processing techniques are designed to enhance the quality of the original image and could enhance spectral quality of the signal. Inclusion of spectral indices and feature extraction achieved minor OA gains.

Table 5 shows the adjusted least squares mean effect sizes and pairwise Tukey–Kramer comparisons from the ANCOVA at three predetermined levels of baseline classification OA. While Table 3 highlights comparisons between each input data enhancement category with its corresponding baseline classification, Table 5 focuses on comparisons between input data enhancement methods. The estimated effect size from inclusion of texture was the highest by a wide margin. The mean effect size achieved by including texture was 12.5% when OA of the baseline classification was 70%, and the effect size was still considerable (4.3%) even for high accuracy baseline classifications with 90% OA. Other methods also demonstrated effect sizes comparable to those documented in Table 3.

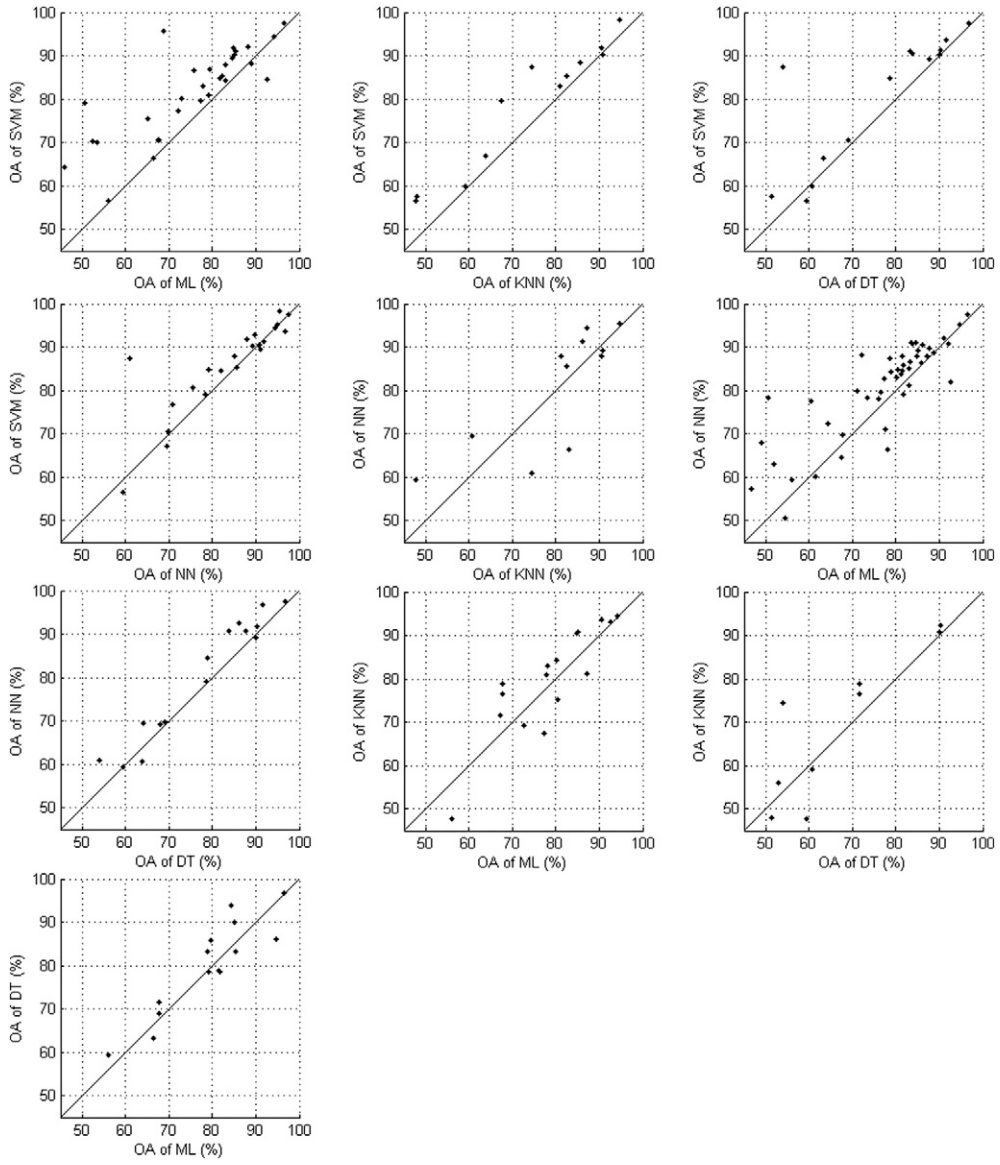

Fig. 3. Comparison of selected pairs of classifiers based on overall accuracy (OA). Classifiers: Decision Tree (DT), K-Nearest Neighbor (KNN), Maximum Likelihood (ML), Neural Network (NN), and Support Vector Machines (SVM).

6. Conclusion and discussion

Satellite monitoring capabilities are essential to understand large scale environmental changes affecting climate, biodiversity and humans. Vast resources have been invested in satellite development, launch and product creation, and thousands of articles have evaluated image classification methods for producing land-cover maps. To date the results of these studies have not been synthesized to generate conclusive guidelines for selecting a land-cover classification process. Our work bridges this gap via a meta-analysis of peer-reviewed studies to statistically quantify improvement in accuracy achieved by different input data enhancements and classification algorithm choices. By identifying the current state-of-the-art, we provide pragmatic guidance, derived from years of published research, aiming to accelerate future processing and algorithmic developments that will yield further improvements in land-cover classification accuracy.

The main practical application of our results is to help researchers decide which improvement methods are most promising. Researchers can prioritize their efforts by obtaining a clearer idea of expected improvements achieved by different methods. The highest improvement resulted from inclusion of texture with a mean increase in overall accuracy of 12.1% (Table 3). The increase in overall accuracy achieved by including texture is attributed to the additional spatial context information provided. These easy to compute texture metrics should be a primary consideration for image classification of land cover. Ancillary data, multi-angle images, and multi-time images, with 8.5%, 8.0%, and 6.9% of mean improvements in overall accuracy respectively, also have the potential to complement existing spectral information, for example through topographic information (Franklin, Peddle, Dechka, & Stenhouse, 2002), by capturing the anisotropy of land surface reflectance, or by capturing different temporal rates of phenological changes in vegetation mapping (Peterson, Price, & Martinko, 2002). Inclusion of spectral indices and feature extraction resulted in minor OA gains. These last two techniques aim to enhance the original classification feature space; however, this enhancement can usually be achieved with advanced highly adaptable classifiers such as SVM and NN.

Generally, inclusion of additional explanatory variables that can differentiate target classes is more promising than enhancement of the original variables. However, the selection of suitable improvement methods depends on other factors such as availability of data and tools, analyst's familiarity with different methods, and time and budget constraints. For instance, although feature extraction had the smallest

expected improvement, it does not require additional datasets. Also, feature extraction options are readily available in remote sensing software and this may lead an analyst to prefer feature extraction before other methods.

Land-cover mapping is a complicated process with numerous factors influencing the quality of the final product. Some of these factors such as landscape complexity, target classes, and scale are usually predetermined by the project's requirements. In this research we focused on the analysis of some of the factors that can be controlled by the analyst to improve classification accuracy. However, these results are not intended to offer a predetermined optimal classification process for a particular case study. Our results portraying the general relative performance of different methods can be used by analysts to select options that have been demonstrated to produce better classifications. The general findings of this synthesis of past research are not meant to substitute for analyst preferences or to account for specific algorithmic benefits that may be relevant for a particular application. Our goal is not to advocate for automated choice of a classification process but to illuminate the aggregated experience from previous comparison studies of classification processes. For example, even though it has been demonstrated that SVM outperforms many other classifiers, if an analyst knows that the quality of the training dataset is not suitable for SVM, then the analyst can use this additional knowledge to rule out SVM methods. It is also important to document these synthesis findings to establish a general standard of comparison. For example, while there is tacit understanding that SVMs improve accuracy, there are still numerous manuscripts published using ML classification as a benchmark. One of the goals of this study is to include some of the general findings in the benchmarking process. While our general findings are not intended to identify optimal performance for a specific application, these results offer a common basis for comparison and discussion of general tendencies of relative performance. To facilitate use of our data for more application-specific purposes, we have included a spreadsheet as an appendix where readers can identify specific manuscripts incorporated in each comparison and examine the details of the features of each case study comparisons.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.rse.2016.02.028.

Acknowledgments

Work was supported by the USDA McIntire Stennis program, a SUNY ESF Graduate Assistantship and the ASPRS National Committee on Academic Engagement. The authors thank Mr. Yuguang Li and Mr. Sheng Yang for assisting in the manuscript database creation. Detailed comments from the anonymous reviewers improved considerably the readability of the manuscript.

References

Alcantara, C., Kuemmerle, T., Prishchepov, A. V., & Radeloff, V. C. (2012). Mapping abandoned agriculture with multi-temporal MODIS satellite data. *Remote Sensing of Environment*, *124*, 334–347.

Anderson, M. C., Allen, R. G., Morse, A., & Kustas, W. P. (2012). Use of Landsat thermal imagery in monitoring evapotranspiration and managing water resources. *Remote Sensing of Environment*, *122*, 50–65.

Asner, G. P., Broadbent, E. N., Oliveira, P. J. C., Keller, M., Knapp, D. E., & Silva, J. M. M. (2006). Condition and fate of logged forests in the Brazilian Amazon. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(34), 12947–12950.

Asner, G. P., Levick, S. R., Kennedy-Bowdoin, T., Knapp, D. E., Emerson, R., Jacobson, J., et al. (2009). Large-scale impacts of herbivores on the structural diversity of African savannas. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(12), 4947–4952.

Birdsey, R., Angeles-Perez, G., Kurz, W. A., Lister, A., Olguin, M., Pan, Y., et al. (2013). Approaches to monitoring changes in carbon stocks for REDD+. *Carbon Management*, *4*, 519–537.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brodley, C. E., & Friedl, M. A. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, *61*(3), 399–409.

Brown De Colstoun, E. C., Story, M. H., Thompson, C., Commisso, K., Smith, T. G., & Irons, J. R. (2003). National park vegetation mapping using multitemporal landsat 7 data and a decision tree classifier. *Remote Sensing of Environment*, *85*(3), 316–327.

Carlson, T. N., & Ripley, D. A. (1997). On the relation between NDVI, fractional vegetation cover, and leaf area index. *Remote Sensing of Environment*, *62*(3), 241–252.

Carpenter, G. A., Gopal, S., Macober, S., Martens, S., Woodcock, C. E., & Franklin, J. (1999). A neural network method for efficient vegetation mapping. *Remote Sensing of Environment*, *70*(3), 326–338.

Carrão, H., Gonçalves, P., & Caetano, M. (2008). Contribution of multispectral and multitemporal information from MODIS images to land cover classification. *Remote Sensing of Environment*, *112*(3), 986–997.

Chan, J. C. W., & Paelinckx, D. (2008). Evaluation of random forest and adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*, *112*(6), 2999–3011.

Chang, Y. L., Liang, L. S., Han, C. C., Fang, J. P., Liang, W. Y., & Chen, K. S. (2007). Multisource data fusion for landslide classification using generalized positive boolean functions. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(6), 1697–1708.

Chavez, P. S., Jr. (1996). Image-based atmospheric corrections — Revisited and improved. *Photogrammetric Engineering and Remote Sensing*, *62*(9), 1025–1036.

Chen, D., Stow, D. A., & Gong, P. (2004). Examining the effect of spatial resolution and texture window size on classification accuracy: An urban environment case. *International Journal of Remote Sensing*, *25*(11), 2177–2192.

Cihlar, J. (2000). Land cover mapping of large areas from satellites: Status and research priorities. *International Journal of Remote Sensing*, *21*(6–7), 1093–1114.

Clark, R. N., & Roush, T. L. (1984). Reflectance spectroscopy: Quantitative analysis techniques for remote sensing applications. *Journal of Geophysical Research*, *89*(B7), 6329–6340.

Clark, M. L., Roberts, D. A., & Clark, D. B. (2005). Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote Sensing of Environment*, *96*(3–4), 375–398.

Coburn, C. A., & Roberts, A. C. B. (2004). A multiscale texture analysis procedure for improved forest stand classification. *International Journal of Remote Sensing*, *25*(20), 4287–4308.

Colby, J. D., & Keating, P. L. (1998). Land cover classification using landsat TM imagery in the tropical highlands: The influence of anisotropic reflectance. *International Journal of Remote Sensing*, *19*(8), 1479–1500.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

Cover, T. M., & Hart, P. E. (1967). Nearest Neighbor Pattern classification. *IEEE Transactions on Information Theory*, *IT-13*(1), 21–27.

Dash, J., Mathur, A., Foody, G. M., Curran, P. J., Chipman, J. W., & Lillesand, T. M. (2007). Land cover classification using multi-temporal MERIS vegetation indices. *International Journal of Remote Sensing*, *28*(6), 1137–1159.

DeFries, R. S., Houghton, R. A., Hansen, M. C., Field, C. B., Skole, D., & Townshend, J. (2002). Carbon emissions from tropical deforestation and regrowth based on satellite observations for the 1980s and 1990s. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(22), 14256–14261.

Del Frate, F., Schiavon, G., Solimini, D., Borgeaud, M., Hoekman, D. H., & Vissers, M. A. M. (2003). Crop classification using multiconfiguration C-band SAR data. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(7 PART I), 1611–1619.

Duca, R., & Del Frate, F. (2008). Hyperspectral and multiangle CHRIS-PROBA images for the generation of land cover maps. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(10), 2857–2866.

Dymond, C. C., Mladenoff, D. J., & Radeloff, V. C. (2002). Phenological differences in tasseled cap indices improve deciduous forest classification. *Remote Sensing of Environment*, *80*(3), 460–472.

Evans, J., van Donkelaar, A., Martin, R. V., Burnett, R., Rainham, D. G., Birkett, N. J., et al. (2013). Estimates of global mortality attributable to particulate air pollution using satellite imagery. *Environmental Research*, *120*, 33–42.

Fauvel, M., Benediktsson, J. A., Chanussot, J., & Sveinsson, J. R. (2008). Spectral and spatial classification of hyperspectral data using SVMs and morphological profiles. *IEEE Transactions on Geoscience and Remote Sensing*, *46*(11), 3804–3814.

Fialko, Y., Sandwell, D., Simons, M., & Rosen, P. (2005). Three-dimensional deformation caused by the Bam, Iran, earthquake and the origin of shallow slip deficit. *Nature*, *435*(7040), 295–299.

Franklin, S. E., & Wulder, M. A. (2002). Remote sensing methods in medium spatial resolution satellite data land cover classification of large areas. *Progress in Physical Geography*, *26*(2), 173–205.

Franklin, S. E., Peddle, D. R., Dechka, J. A., & Stenhouse, G. B. (2002). Evidential reasoning with landsat TM, DEM and GIS data for landcover classification in support of grizzly bar habitat mapping. *International Journal of Remote Sensing*, *23*(21), 4633–4652.

Fukushima, H., Higurashi, A., Mitomi, Y., Nakajima, T., Noguchi, T., Tanaka, T., et al. (1998). Correction of atmospheric effect on ADEOS/OCTS ocean color data: Algorithm description and evaluation of its performance. *Journal of Oceanography*, *54*(5), 417–430.

Geerling, G. W., Labrador-Garcia, M., Clevers, J. G. P. W., Ragas, A. M. J., & Smits, A. J. M. (2007). Classification of floodplain vegetation by data fusion of spectral (CASI) and LiDAR data. *International Journal of Remote Sensing*, *28*(19), 4263–4284.

Gilbert, M., Xiao, X., Pfeiffer, D. U., Epprecht, M., Boles, S., Czarnecki, C., et al. (2008). Mapping H5N1 highly pathogenic avian influenza risk in Southeast Asia. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(12), 4769–4774.

Giustarini, L., Hostache, R., Matgen, P., Schumann, G. J. P., Bates, P. D., & Mason, D. C. (2013). A change detection approach to flood mapping in urban areas using TerraSAR-X. *IEEE Transactions on Geoscience and Remote Sensing*, *51*, 2417–2430.

Gong, P., Wang, J., Yu, L., Zhao, Y., Zhao, Y., Liang, L., et al. (2013). Finer resolution observation and monitoring of global land cover: First mapping results with Landsat TM and ETM+ data. *International Journal of Remote Sensing*, *34*, 2607–2654.

Grekousis, G., Mountrakis, G., & Kavouras, M. (2015). An overview of 21 global and 43 regional land-cover mapping products. *International Journal of Remote Sensing*, *36*, 5309–5335.

Guerschman, J. P., Paruelo, J. M., Di Bella, C., Giallorenzi, M. C., & Pacin, F. (2003). Land cover classification in the Argentine pampas using multi-temporal landsat TM data. *International Journal of Remote Sensing*, *24*(17), 3381–3402.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., et al. (2013). High-resolution global maps of 21st-century forest cover change. *Science*, *342*, 850–853.

Hansen, M. C., Stehman, S. V., Potapov, P. V., Loveland, T. R., Townshend, J. R. G., DeFries, R. S., et al. (2008). Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(27), 9439–9444.

Heikkinen, V., Korpela, I., Tokola, T., Honkavaara, E., & Parkkinen, J. (2011). An SVM classification of tree species radiometric signatures based on the Leica ADS40 sensor. *IEEE Transactions on Geoscience and Remote Sensing*, *49*(11 PART 2), 4539–4551.

Held, A., Ticehurst, C., Lymburner, L., & Williams, N. (2003). High resolution mapping of tropical mangrove ecosystems using hyperspectral and radar remote sensing. *International Journal of Remote Sensing*, *24*(13), 2739–2759.

Hong, B., Limburg, K. E., Hall, M. H., Mountrakis, G., Groffman, P. M., Hyde, K., et al. (2012). An integrated monitoring/modeling framework for assessing human–nature interactions in urbanizing watersheds: Wappinger and Onondaga Creek watersheds, New York, USA. *Environmental Modelling & Software*, *32*, 1–15.

Huang, H., Gong, P., Clinton, N., & Hui, F. (2008). Reduction of atmospheric and topographic effect on landsat TM data for forest classification. *International Journal of Remote Sensing*, *29*(19), 5623–5642.

Hussain, M., Chen, D., Cheng, A., Wei, H., & Stanley, D. (2013). Change detection from remotely sensed images: From pixel-based to object-based approaches. *ISPRS Journal of Photogrammetry and Remote Sensing*, *80*, 91–106.

Ji, C. Y. (2000). Land-use classification of remotely sensed data using Kohonen self-organizing feature map neural networks. *Photogrammetric Engineering and Remote Sensing*, *66*(12), 1451–1460.

Keegan, K. M., Albert, M. R., McConnell, J. R., & Baker, I. (2014). Climate change and forest fires synergistically drive widespread melt events of the Greenland ice sheet. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(22), 7964–7967.

Khatami, R., & Mountrakis, G. (2012). Implications of classification of methodological decisions in flooding analysis from hurricane Katrina. *Remote Sensing*, *4*(12), 3877–3891.

Knyazikhin, Y., Schull, M. A., Stenberg, P., Mõttus, M., Rautiainen, M., Yang, Y., et al. (2013). Hyperspectral remote sensing of foliar nitrogen content. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(3) (811–812 + E185–E192).

Kuplich, T. M., Freitas, C. C., & Soares, J. V. (2000). The study of ERS-1 SAR and landsat TM synergism for land use classification. *International Journal of Remote Sensing*, *21*(10), 2101–2111.

Lambin, E. F., & Meyfroidt, P. (2011). Global land use change, economic globalization, and the looming land scarcity. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(9), 3465–3472.

Lillesand, T. M., Kiefer, R. W., & Chipman, J. W. (2004). *Remote sensing and image interpretation* (5th ed.). Hoboken NJ: Wiley.

Liu, H., & Weng, Q. (2012). Enhancing temporal resolution of satellite imagery for public health studies: A case study of West Nile virus outbreak in Los Angeles in 2007. *Remote Sensing of Environment*, *117*, 57–71.

Lobitz, B., Beck, L., Huq, A., Wood, B., Fuchs, G., Faruque, A. S. G., et al. (2000). Climate and infectious disease: use of remote sensing for detection of vibrio cholerae by indirect measurement. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(4), 1438–1443.

Lu, D., & Weng, Q. (2007). A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, *28*(5), 823–870.

McMenamin, S. K., Hadly, E. A., & Wright, C. K. (2008). Climatic change and wetland desiccation cause amphibian decline in Yellowstone National Park. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(44), 16988–16993.

Mendenhall, C. D., Sekercioglu, C. H., Brenes, F. O., Ehrlich, P. R., & Daily, G. C. (2011). Predictive model for sustaining biodiversity in tropical countryside. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16313–16316.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Altman, D., Antes, G., et al. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine*, *6*(7).

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*, 247–259.

Mountrakis, G., Watts, R., Luo, L., & Wang, J. (2009). Developing collaborative classifiers using an expert-based model. *Photogrammetric Engineering and Remote Sensing*, *75*(7), 831–843.

Mundt, J. T., Glenn, N. F., Weber, K. T., Prather, T. S., Lass, L. W., & Pettingill, J. (2005). Discrimination of hoary cress and determination of its detection limits via hyperspectral image processing and accuracy assessment techniques. *Remote Sensing of Environment*, *96*(3–4), 509–517.

Myneni, R. B., Dong, J., Tucker, C. J., Kaufmann, R. K., Kauppi, P. E., Liski, J., et al. (2001). A large carbon sink in the woody biomass of northern forests. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(26), 14784–14789.

Myneni, R. B., Yang, W., Nemani, R. R., Huete, A. R., Dickinson, R. E., Knyazikhin, Y., et al. (2007). Large seasonal swings in leaf area of Amazon rainforests. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(12), 4820–4823.

Nagendra, H., & Gadgil, M. (1999). Biodiversity assessment at multiple scales: Linking remotely sensed data with field information. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(16), 9154–9158.

Ogilvie, A., Belaud, G., Delenne, C., Bailly, J., Bader, J., Oleksiak, A., et al. (2015). Decadal monitoring of the Niger Inner Delta flood dynamics using MODIS optical data. *Journal of Hydrology*, *523*, 368–383.

Peterson, D. L., Price, K. P., & Martinko, E. A. (2002). Discriminating between cool season and warm season grassland cover types in northeastern Kansas. *International Journal of Remote Sensing*, *23*(23), 5015–5030.

Potapov, P. V., Turubanova, S. A., Tyukavina, A., Krylov, A. M., McCarty, J. L., Radeloff, V. C., et al. (2015). Eastern Europe's forest cover dynamics from 1985 to 2012 quantified from the full Landsat archive. *Remote Sensing of Environment*, *159*, 28–43.

Ranson, K. J., Sun, G., Kharuk, V. I., & Kovacs, K. (2001). Characterization of forests in Western Sayani Mountains, Siberia from SIR-C SAR data. *Remote Sensing of Environment*, *75*(2), 188–200.

Ricchetti, E. (2000). Multispectral satellite image and ancillary data integration for geological classification. *Photogrammetric Engineering and Remote Sensing*, *66*(4), 429–435.

Richter, R. (1997). Correction of atmospheric and topographic effects for high spatial resolution satellite imagery. *International Journal of Remote Sensing*, *18*(5), 1099–1111.

Rindfuss, R. R., Walsh, S. J., Turner, B. L., II, Fox, J., & Mishra, V. (2004). Developing a science of land change: Challenges and methodological issues. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(39), 13976–13981.

Rio, J. N. R., & Lozano-García, D. F. (2000). Spatial filtering of radar data (RADARSAT) for wetlands (brackish marshes) classification. *Remote Sensing of Environment*, *73*(2), 143–151.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536.

Schmidt, K. S., Skidmore, A. K., Kloosterman, E. H., Van Oosten, H., Kumar, L., & Janssen, J. A. M. (2004). Mapping coastal vegetation using an expert system and hyperspectral imagery. *Photogrammetric Engineering and Remote Sensing*, *70*(6), 703–715.

Schwalm, C. R., Williams, C. A., Schaefer, K., Baldocchi, D., Black, T. A., Goldstein, A. H., et al. (2012). Reduction in carbon uptake during turn of the century drought in western North America. *Nature Geoscience*, *5*, 551–556.

Sedano, F., Gong, P., & Ferrão, M. (2005). Land cover assessment with MODIS imagery in southern African Miombo ecosystems. *Remote Sensing of Environment*, *98*(4), 429–441.

Sesnie, S. E., Gessler, P. E., Finegan, B., & Thessler, S. (2008). Integrating landsat TM and SRTM-DEM derived variables with decision trees for habitat classification and change detection in complex neotropical environments. *Remote Sensing of Environment*, *112*(5), 2145–2159.

Shackelford, A. K., & Davis, C. H. (2003). A hierarchical fuzzy classification approach for high-resolution multispectral data over urban areas. *IEEE Transactions on Geoscience and Remote Sensing*, *41*(9 PART I), 1920–1932.

Skidmore, A. K., Pettorelli, N., Coops, N. C., Geller, G. N., Hansen, M., Lucas, R., et al. (2015). Environmental science: Agree on biodiversity metrics to track from space. *Nature*, *523*, 403–405.

Sluiter, R., & Pebesma, E. J. (2010). Comparing techniques for vegetation classification using multi- and hyperspectral images and ancillary environmental data. *International Journal of Remote Sensing*, *31*(23), 6143–6161.

Smits, P. C., Dellepiane, S. G., & Schowengerdt, R. A. (1999). Quality assessment of image classification algorithms for land-cover mapping: A review and a proposal for a cost-based approach. *International Journal of Remote Sensing*, *20*(8), 1461–1486.

Song, C., Woodcock, C. E., Seto, K. C., Lenney, M. P., & Macomber, S. A. (2001). Classification and change detection using landsat TM data: When and how to correct atmospheric effects? *Remote Sensing of Environment*, *75*(2), 230–244.

Stehman, S. V. (2006). Design, analysis, and inference for studies comparing thematic accuracy of classified remotely sensed data: A special case of map comparison. *Journal of Geographical Systems*, *8*(2), 209–226.

Strahler, A. H. (1980). The use of prior probabilities in maximum likelihood classification of remotely sensed data. *Remote Sensing of Environment*, *10*(2), 135–163.

Su, L., Chopping, M. J., Rango, A., Martonchik, J. V., & Peters, D. P. C. (2007a). Differentiation of semi-arid vegetation types based on multi-angular observations from MISR and MODIS. *International Journal of Remote Sensing*, *28*(6), 1419–1424.

Su, L., Chopping, M. J., Rango, A., Martonchik, J. V., & Peters, D. P. C. (2007b). Support vector machines for recognition of semi-arid vegetation types using MISR multi-angle imagery. *Remote Sensing of Environment*, *107*(1–2), 299–311.

Syed, T. H., Famiglietti, J. S., Chambers, D. P., Willis, J. K., & Hilburn, K. (2010). Satellite-based global-ocean mass balance estimates of interannual variability and emerging trends in continental freshwater discharge. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(42), 17916–17921.

Taylor, S., Kumar, L., & Reid, N. (2010). Mapping *Lantana camara*: Accuracy comparison of various fusion techniques. *Photogrammetric Engineering and Remote Sensing*, *76*(6), 691–700.

Thenkabail, P. S., Enclona, E. A., Ashton, M. S., & Van Der Meer, B. (2004). Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications. *Remote Sensing of Environment*, *91*(3–4), 354–376.

Tottrup, C. (2004). Improving tropical forest mapping using multi-date landsat TM data and pre-classification image smoothing. *International Journal of Remote Sensing*, *25*(4), 717–730.

Townshend, J. R., Masek, J. G., Huang, C., Vermote, E. F., Gao, F., Channan, S., et al. (2012). Global characterization and monitoring of forest cover using Landsat data: Opportunities and challenges. *International Journal of Digital Earth*, *5*, 373–397.

Tsai, F., & Philpot, W. D. (2002). A derivative-aided hyperspectral image analysis system for land-cover classification. *IEEE Transactions on Geoscience and Remote Sensing*, *40*(2), 416–425.

Watts, J. D., Powell, S. L., Lawrence, R. L., & Hilker, T. (2011). Improved classification of conservation tillage adoption using high temporal and synthetic satellite imagery. *Remote Sensing of Environment*, *115*(1), 66–75.

Weng, Q. (2012). Remote sensing of impervious surfaces in the urban areas: Requirements, methods, and trends. *Remote Sensing of Environment*, *117*, 34–49.

Wilkinson, G. G. (2005). Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Transactions on Geoscience and Remote Sensing*, *43*(3), 433–440.

Yang, C., Everitt, J. H., & Johnson, H. B. (2009). Applying image transformation and classification techniques to airborne hyperspectral imagery for mapping Ashe juniper infestations. *International Journal of Remote Sensing*, *30*(11), 2741–2758.

Youngentob, K. N., Roberts, D. A., Held, A. A., Dennison, P. E., Jia, X., & Lindenmayer, D. B. (2011). Mapping two eucalyptus subgenera using multiple endmember spectral mixture analysis and continuum-removed imaging spectrometry data. *Remote Sensing of Environment*, *115*(5), 1115–1128.

Yu, L., Liang, L., Wang, J., Zhao, Y., Cheng, Q., Hu, L., et al. (2014). Meta-discoveries from a synthesis of satellite-based land-cover mapping research. *International Journal of Remote Sensing*, *35*(13), 4573–4588.