Contents lists available at SciVerse ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Assessing reference dataset representativeness through confidence metrics based on information density

Giorgos Mountrakis *, Bo Xi

Department of Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry,1 Forestry Dr., Syracuse, NY 13210, USA

ARTICLE INFO

Article history:
Received 18 January 2012
Received in revised form 10 October 2012
Accepted 31 January 2013
Available online 6 March 2013

Keywords:
Representativeness
Classification
Quality
Reference data
Reliability
Training

ABSTRACT

Land cover maps obtained from classification of remotely sensed imagery provide valuable information in numerous environmental monitoring and modeling tasks. However, many uncertainties and errors can directly or indirectly affect the quality of derived maps. This work focuses on one key aspect of the supervised classification process of remotely sensed imagery: the quality of the reference dataset used to develop a classifier. More specifically, the representative power of the reference dataset is assessed by contrasting it with the full dataset (e.g. entire image) needing classification. Our method is applicable in several ways: training or testing datasets (extracted from the reference dataset) can be compared with the full dataset. The proposed method moves beyond spatial sampling schemes (e.g. grid, cluster) and operates in the multidimensional feature space (e.g. spectral bands) and uses spatial statistics to compare information density of data to be classified with data used in the reference process. The working hypothesis is that higher information density, not in general but with respect to the entire classified image, expresses higher confidence in obtained results. Presented experiments establish a close link between confidence metrics and classification accuracy for a variety of image classifiers namely maximum likelihood, decision tree, Backpropagation Neural Network and Support Vector Machine. A sensitivity analysis demonstrates that spatially-continuous reference datasets (e.g. a square window) have the potential to provide similar classification confidence as typically-used spatially-random datasets. This is an important finding considering the higher acquisition costs for randomly distributed datasets. Furthermore, the method produces confidence maps that allow spatially-explicit comparison of confidence metrics within a given image for identification of over- and under-represented image portions. The current method is presented for individual image classification but, with sufficient evaluation from the remote sensing community it has the potential to become a standard for reference dataset reporting and thus allowing users to assess representativeness of reference datasets in a consistent manner across different classification tasks.

1. Introduction

Remote sensing offers significant contributions to environmental monitoring. As an important product of classification of remotely sensed data, thematic maps are crucial to various applications, and are the basis to understand associated land cover changes (Foody and Mathur, 2004). However, many uncertainties and errors can directly or indirectly affect the quality of thematic maps; for example, uncertainties resulting from the image acquisition system, variable illumination, atmospheric conditions, and complex spatial patterns of heterogeneous landscapes. Thus, the accurate allocation of each pixel to the correct class is a challenging task. Numerous methods have been developed to assess the quality of the thematic map products, as these products have a wide range of environmental applications. In this paper, we focus on one key aspect in the classification process from remotely sensed imagery: the quality of the reference dataset used to perform a classification, and more specifically, how representative the reference dataset is to the overall image.

For the remainder of the paper it is assumed that the classification is a result of a supervised process, since training data are necessary for the proposed methodology. As agreed by Rebbapragada et al. (2008), the quality of the supervised land cover mapping products depends on two main factors, the accuracy of the classifier used to produce the thematic maps and the quality of the labeled data used to train the classifier. A considerable amount of

* Corresponding author. Address: Department of Environmental Resources Engineering, State University of New York College of Environmental Science and Forestry, 419 Baker Hall, 1 Forestry Dr., Syracuse, NY 13210, USA. Tel.: +1 315 470 4824; fax: +1 315 470 6958.

E-mail address: gmountrakis@esf.edu (G. Mountrakis).

0924-2716/$ - see front matter © 2013 International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS) Published by Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.isprsjprs.2013.01.011

research is dedicated to algorithmic development for classification accuracy improvements (e.g. Swain and Davis, 1978; Foody and Mathur, 2004). A variety of means to assess the classification accuracy has also been presented. The traditional and most widely used methods to assess classification accuracy are classification confusion or error matrices (Congalton, 1991; Foody, 2002, 2009). In addition, several measures (e.g., entropy and fuzzy set-based accuracy assessment) assessing the classification accuracy at local or pixel level have been proposed to improve the classification accuracy assessment (Binaghi et al., 1999; Foody, 1996; Liu et al., 2004; Mitchell et al., 2008; Steele et al., 1998; Van der Wel et al., 1998). Lately, the validity of several established approaches has been questioned (Stehman, 2004; Pontius and Millones, 2011).

For clarity, a few definitions are provided here. Assuming an image in need of classification, all pixels of that image would fall either in the reference dataset or cover the rest of the image. In practice for supervised classification a reference dataset is often separated into training and testing datasets, with the former used to develop the classification rules and the latter to assess the classification accuracy. The effectiveness and accuracy of the classification algorithms depend upon the training dataset selection. It is well known that different classifiers (e.g. a decision tree vs. a neural network) can generate different classification results and therefore accuracies for the same training data, as each algorithm uses different characteristics or features of the training data to define the decision boundaries for class definition (Foody and Mathur, 2006). Some studies (e.g., Lu et al., 2004; Kandrika and Roy, 2008) have supported the above statement and identified three main reasons behind variable accuracy of different classifiers: the use of coarse resolution satellite images (e.g., Landsat TM), the landscape heterogeneity, and the limited representativeness of the training dataset.

Thus, the quality of the training dataset is a significant concern before the class allocation process, and can have a larger impact on classification accuracy than the implemented classification technique (Campbell, 2007). Several training data selection methods have been proposed and compared in the literature (Kavzoglu, 2009). Richards and Jia (1999) stated that the traditional training data sampling strategies are generally polygon-based (e.g. on screen selection of polygonal training data). Within a selected polygon or training region, the sample of a respective class is usually homogenous and the sample may be auto-correlated and under-representative of an informational class. In addition, by seeking the extremes in the feature space, some methods (e.g. the Pixel Purity Index (ENVI, 1999)) try to find the 'purest' pixels in an image, while Lesparre and Gorte (2006) stated that in some cases (for example images of natural vegetation) it can be difficult to obtain sufficient pure training samples to accurately estimate the spectra of the classes. Several semi-automated and automated sampling strategies and training methods have been studied to minimize bias, generate pure spectra, and achieve high classification performance (McCaffrey and Franklin, 1993; Cano et al., 2007). For instance, Sebban et al. (2000) presented the prototype selection algorithms as training data pre-processing techniques for decision tree simplification. Reeves and Bush (2001) used a genetic algorithm for training dataset selection in radial-based function networks and obtained improved generalization.

Furthermore, researchers have investigated general requirements for training set size and location within different data dimensionalities in order to mitigate the data dimensionality problem (a.k.a. the Hughes phenomenon) and obtain high classification accuracy (Barandela et al., 2004; Jain et al., 2000; Raudys and Jain, 1991; Shahshahani and Landgrebe, 1994; Van Niel et al., 2005). A proposed rule about the relationship between training sample size n and data dimensionality p is that n lies between $10p$ and $100p$ (Lillesand et al., 2004). Washer and Landgrebe (1984) stated that larger numbers of training samples may be required to adequately estimate class statistics when high data dimensionality is involved (e.g. in hyperspectral data). Foody and Mathur (2004) also indicated that small training sets may result in low classification accuracy, which is especially apparent for analyses using hyperspectral sensor data. Van Niel et al. (2005) demonstrated that the needed number of training samples should also be determined by considering the complexity of the discrimination problem (e.g. a lower number of training samples may be necessary for a simple discrimination problem). In addition, a limited yet salient feature set simplifies both the training pattern and the classifiers that are built on the selected pattern (Jain et al., 2000).

Several means of evaluating the representativeness of a given training dataset have been proposed in the literature. Representativeness is defined as the degree of expressiveness of the training dataset with respect to the rest of the image and/or the ground spectral properties. The graphical representations of the spectral response patterns are often used to evaluate the quality of training samples in different bands. Researchers investigate whether training sets are normally distributed and whether different training classes are spectrally separable (Lillesand et al., 2004). Richards et al. (1999) stated that the weaknesses in the selection of the training sets can also be identified by using some form of threshold or limit on the classifiers (e.g., the minimum distance classifier and Maximum Likelihood Classifier). However, these conventional evaluation methods only examine the training dataset without taking into account the entire image, and/or are restricted to a specific classification method that they adopt.

This work investigates reference data quality through a novel methodology that evaluates the representativeness of the training (or testing) data when compared to the full scene to be classified. The method offers two important characteristics: (i) it is independent of the classification methodology, and (ii) it does not require any labeling of data. The proposed method compares the multidimensional footprint (in the spectral or feature space) of a new unlabeled point with the footprint of the data used in the training (or testing) process. The underlying theory is based on spatial statistics applied to this multidimensional feature space. Unlabeled pixels that match or surpass the spatial density (i.e. spatial clustering) of the training (or testing) sample would exhibit higher classification confidence. Classification confidence should not be confused with classification accuracy. The link between confidence and accuracy is further investigated in later sections.

2. Methodology

This section presents the theory behind the proposed method followed by an example to further clarify implementation. For simplicity Sections 2.1 and 2.2 examine only one of possible implementations, how representative is a training dataset with respect to the remaining image that is not part of the reference data. In order to do so, each point that is not part of the reference dataset is compared with the entire training dataset.

2.1. Theory

The proposed methodology operates in the multidimensional spectral space, where each dimension corresponds to one spectral band. If further processing has taken place and additional features are introduced (e.g. normalized band differences, texture metrics), each feature is assumed to correspond to one dimension. For consistency, the rest of the paper will use the term multidimensional feature space. Because the proposed method relies on Euclidean distance calculations on that multidimensional feature space, it is recommended to scale the training dataset so that all dimensions occupy the same range. In that case all the unlabelled points would also need to be scaled based on the scaling equations extracted from the training dataset. Furthermore, to avoid dimensionality correlations a principal component analysis may take place as an independent pre-processing step.

Our working assumption is that the classification confidence of an unlabelled point P, where P is not part of the reference dataset, increases when it is surrounded in close proximity by a high number of training points in the multidimensional feature space. Spatial statistical methods are employed to assess confidence as follows. Given a training dataset with n points all Euclidean distances are calculated between every possible pair in that training dataset. A modified K function (Ripley, 1976) is then used to count the number of points falling within a given neighborhood of each training point. The neighborhood is defined as a multidimensional sphere with a specified multidimensional distance. The resulting K function for a training dataset K_{TS} is defined over multiple distances h as:

$$K_{TS}(h) = \sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}) \tag{1}$$

where h is a predefined Euclidean distance in the multidimensional feature space, n is the number of the training points, d_{ij} is the Euclidean distance between two points in that multidimensional feature space , and $I_h(d_{ij})$ is an indicator function defined as:

$$I_h(d_{ij}) = \begin{cases} 1 & \text{for } d_{ij} \leq h, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

In other words, the K function counts the average number of neighboring points that fall within a given multidimensional sphere of radius h. This K_{TS} function acts as the basis for comparison and it is solely created from all points within the training dataset.

The K_{TS} benchmark function is compared to the corresponding K function for an unlabelled point P, defined as K_P. Euclidean distances are calculated from the unlabeled point P and each point in the training dataset, producing a $1 \times n$ distance vector. This distance vector is used to obtain $K_P(h)$ as follows:

$$K_p(h) = n \times \frac{n-1}{n} \times \sum_{i=1}^{n} I_h(d_i) = (n-1) \times \sum_{i=1}^{n} I_h(d_i) \tag{3}$$

where n is the number of training points, and d_i is the Euclidean distance between the unlabelled point P and the ith point in the training dataset. $I_h(d_i)$ is the same indicator function from Eq. (2). Because the $K_{TS}(h)$ function contains a double summation, $K_P(h)$ has to be amplified by n times. However, when calculating $K_P(h)$ there is one additional point that can be counted, since $K_{TS}(h)$ does not count the point that is the center of the multidimensional sphere. Therefore, $K_P(h)$ has to be scaled down slightly by a factor of $\frac{n-1}{n}$, resulting in the final form of Eq. (3).

As mentioned, the calculation of the expected number of neighbors in the training dataset is expressed by the $K_{TS}(h)$ function and the observed neighbors for an unlabelled point are manifested through $K_P(h)$. A normalized comparison between the two takes place using a new $Z(h)$ statistic:

$$Z(h) = \frac{K_p(h) - K_{TS}(h)}{K_P(h) + K_{TS}(h)} \tag{4}$$

Positive $Z(h)$ values indicate that at a given h distance (i.e. a multidimensional sphere of radius h) the unlabelled point P is surrounded by a larger than the average number of training points. This would translate in higher than average confidence in the classification of that unlabelled pixel.

A further adjustment can be made to introduce larger weights on nearby points expressed by a scaling factor defined as $W(h)$ in the following equation:

$$Z(h) = W(h) \times \frac{K_P(h) - K_{TS}(h)}{K_P(h) + K_{TS}(h)} \tag{5}$$

Two different methods are proposed, Linear and Gaussian, to assign the weights, although others can be easily incorporated. The Linear weights, $W_{LIN}(h)$, are computed according to distances h in the following equation:

$$W_{LIN}(h) = 1 - \frac{h}{h\_\mathrm{max}} \tag{6}$$

where *h_max* is the maximum distance found in the training dataset. The Gaussian weights, $W_G(h)$, are calculated using:

$$W_G(h) = e^{-\frac{h^2}{2\times c^2}} \tag{7}$$

where parameter c is the standard deviation of a Gaussian curve, and it controls the falling off speed. This standard deviation can be defined by users using different percentiles of the distance values found in every pair within the training dataset (i.e. calculate all possible distance pairs, order them and assign as the standard deviation the 50% percentile of those distances). Example profiles of the different weight functions are shown in Fig. 1.

The number following the "G" notation of the graph corresponds to the distance percentile assigned as the standard deviation (e.g. G10 uses as standard deviation the 10th percentile distance). It is recommended that multiple weight schemes are examined to accommodate for different spatial behavior (e.g. the Linear in combination with the G10).

All statistics presented to this point are a function of scale as expressed through different Euclidean distances h. From the practical perspective, it is necessary to calculate a single confidence value for each unlabelled point; therefore the $Z(h)$ statistic should be aggregated over all distances. Initially, the positive $Z(h)$ values are aggregated using Eq. (8) to a single value defined as Z^+ and the negative $Z(h)$ values to a single Z^- value using Eq. (9), respectively, as follows:

$$Z^+ = \sum_{i=1}^{n^+} Z(h), \quad \text{for } Z(h) > 0 \tag{8}$$

$$Z^- = \sum_{j=1}^{n^-} Z(h), \quad \text{for } Z(h) < 0 \tag{9}$$

In Eq. (8) n^+ is the total number of the positive $Z(h)$, and in Eq. (9) n^- is the total number of the negative $Z(h)$. Fig. 2 shows the profile of the $Z(h)$ statistic and how Eqs. (8) and (9) are applied.

At the last step for individual pixel processing, the final confidence value, denoted as C, is calculated for each unlabelled point (pixel) as follows:

$$C = \frac{Z^+ + Z^-}{Z^+ + |Z^-|} \tag{10}$$

Values of the C statistic range from -1 to 1. A positive C suggests higher positive than negative $Z(h)$ values; thus, more training points are surrounding the unlabelled point and therefore the classifier can make more educated decisions. The greater the C value is, the higher the confidence. In contrast, if negative C values are obtained this would translate to a limited number of training points around the unlabelled point, which indicates that the training dataset might not be representative enough to classify that pixel.

Confidence values from individual unlabelled pixels can lead to a confidence map (e.g. see Fig. 17), which provides a visual guide for under and oversampled pixels with respect to the training dataset. It is also useful to quantify the overall representativeness qual-

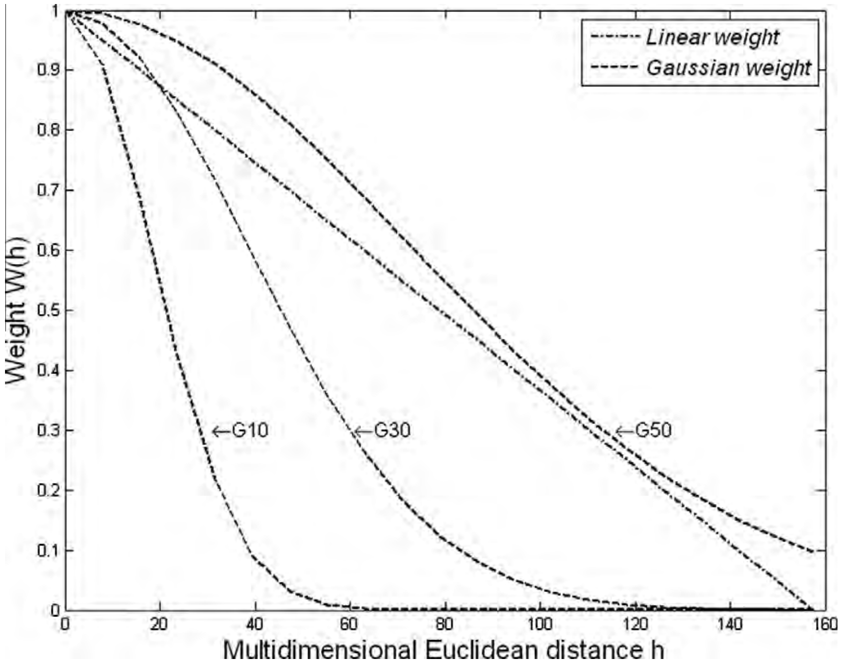

Fig. 1. Linear and Gaussian weight schemes for the $Z(h)$ statistic.

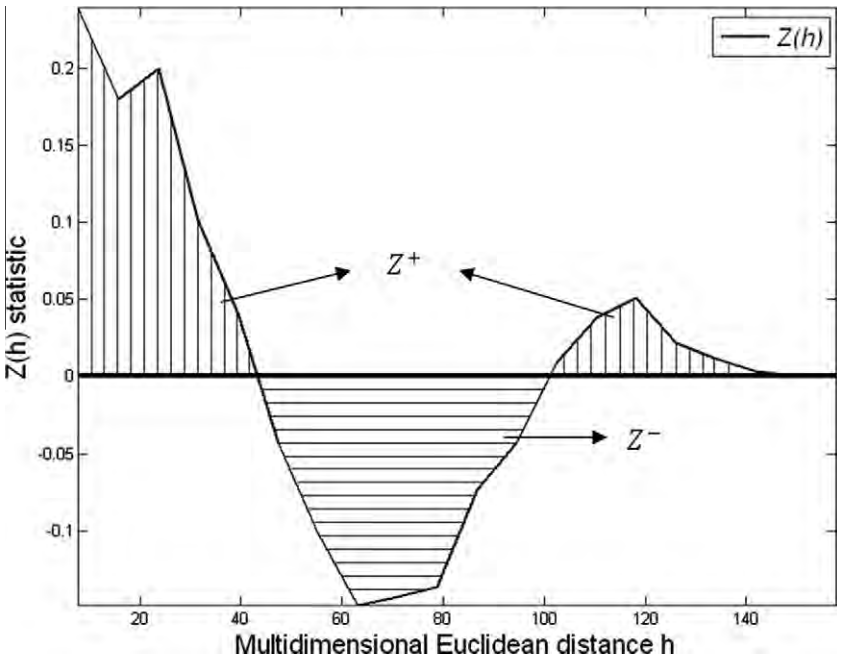

Fig. 2. The profile of the $Z(h)$ statistic showing positive $Z(h)$ and negative $Z(h)$.

ity of the training dataset for a given simulation site (C_{global}). This is expressed through a simple weighted average of the confidence of all associated pixels in the simulation site, as follows:

$$C_{global} = \frac{\sum_i^m C_i \times Q_i}{m} \tag{11}$$

where C_i is the confidence value C for each pixel and m is the total number of pixels. The Q_i value is a user-defined weight for each pixel and its default value is 1. These weights could be used to increase the contribution to C_{global} from specific regions and/or used to highlight classes with rare representation. The above statistics are implemented in Matlab software and the code is freely available.

2.2. Implementation example

To illustrate the proposed method, three example point cases are selected as unlabelled pixels (shown in Fig. 3 as points $P1$, $P2$ and $P3$). The quality of the training dataset for these cases is evaluated. The three unlabeled points are intentionally selected to represent different confidence levels. Point $P1$ is surrounded by a high number of training points and therefore classification would exhi-

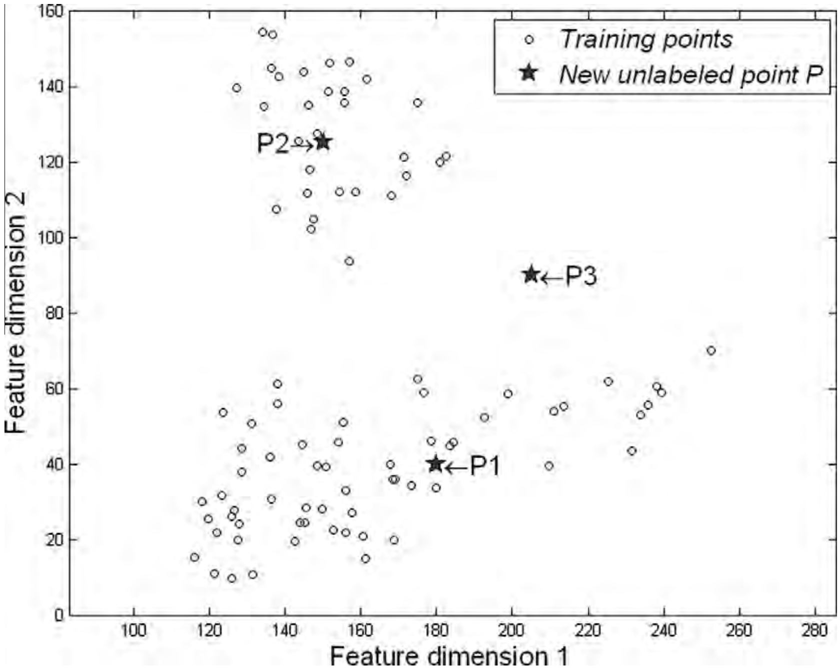

Fig. 3. A two-dimensional feature space showing a training dataset and three new unlabeled points $P1$, $P2$, $P3$.

bit high confidence. Point $P2$ is in the center of another cluster of the training dataset but this cluster has a lower number of training points compared with the cluster associated with $P1$. The last case of point $P3$ is far away from training points, which should lead to lower classification confidence.

The $K(h)$ and $Z(h)$ statistics in Fig. 4 correspond to the three unlabelled points $P1$, $P2$, and $P3$ in Fig. 3 and are calculated using Eqs. (1)–(4). Two versions of the $K(h)$ function are presented, the $K_{TS}(h)$ and $K_P(h)$, in dashed line and solid line respectively. Using these two functions comparisons can be made between the observed number of training points around the unlabelled point P at certain distances h (using $K_P(h)$) and the expected number of points from studying exclusively the training dataset (expressed through $K_{TS}(h)$). If $K_P(h)$ is above $K_{TS}(h)$, which generates a positive $Z(h)$ value, the number of training points around the unlabelled point P is greater than the average number of training points within the specific feature distance h. Thus, more information is available to train a classifier, leading to higher confidence. For the $Z(h)$ statistic figures (Fig. 4b, d and f), the zero baseline is highlighted to allow an easier visual comparison between positive and negative Z values. There are large differences in the $Z(h)$ values among the three points. For point $P1$, the positive $Z(h)$ values dominate the entire range of the feature distances h. For point $P2$, the $Z(h)$ values are positive in the beginning, then they decrease, stay negative for a certain range and finally go back to positive for large h distances, suggesting an average confidence level. For point $P3$, the negative Z values are dominant, indicating low confidence.

Table 1 lists the confidence C values for each unlabelled point using Eqs. (4)–(10). The C_{EQ} confidence metric is produced with Equal weight to all distances (i.e. ignoring the $W(h)$ value in Eq. (5)). C_{LIN} and C_G are calculated using the Linear weights and the Gaussian weights according to feature distances in Eqs. (6) and (7), respectively. Points $P1$ and $P3$ are the two extreme cases, with $P1$ exhibiting high confidence through values close to one, and $P3$ low confidence, reaching the other limit value of minus one. Both points $P1$ and $P3$ show strong responses independently of the implemented weighting schemes. Looking at the confidence of point $P2$, although consistently positive, it is dependent to some extent on the weighting scheme. Confidence progressively drops as larger distances are incorporated in the overall calculation; see G10, G30 and G50 in Fig. 1 and their associated confidence values in Table 1. If the C_{G10} value is examined for point $P2$ the value is high (0.99). This suggests that within smaller distances around point $P2$, the amount of the surrounding training points is greater than average. However, once the feature distance increases beyond that range, no more additional training points are found which results in the negative $Z(h)$ values and lower C_{G30} and C_{G50} values. It is worthwhile to test different weighting schemes to reflect the confidence changes, an issue examined in the experiments section with associated discussion.

3. Linking confidence metrics and classification accuracy

The previous sections discussed how a dataset's representativeness is captured in the proposed confidence metric. As an example, the representativeness of a training dataset was examined with respect to the entire image. From the spatial statistics employed it is expected that our method does capture representativeness. The next logical question is whether representativeness goes beyond the level of providing confidence on the obtained results, can it also improve classification accuracy? From the practical perspective a confidence metric has substantial value if it exhibits a strong positive link to classification accuracy. The inverse does not necessarily hold true, namely that if confidence increases but accuracy does not then the dataset in not more representative. It could be an indication that the classification task may be easy so even non-representative datasets could offer high accuracy, or that the classifier does not have the modeling capability to capture additional more representative information.

Two study sites were used and multiple classification algorithms were implemented to further assess the confidence–accuracy relationship. Section 3.1 presents a multi-class problem using high spatial resolution QuickBird imagery and Section 3.2 offers another example of binary classification for a Landsat scene. It should be noted that in these two sections confidence is measured as the representativeness of a training dataset with respect to the testing dataset. Pixels that do not belong in either the training or testing dataset are ignored for this analysis.

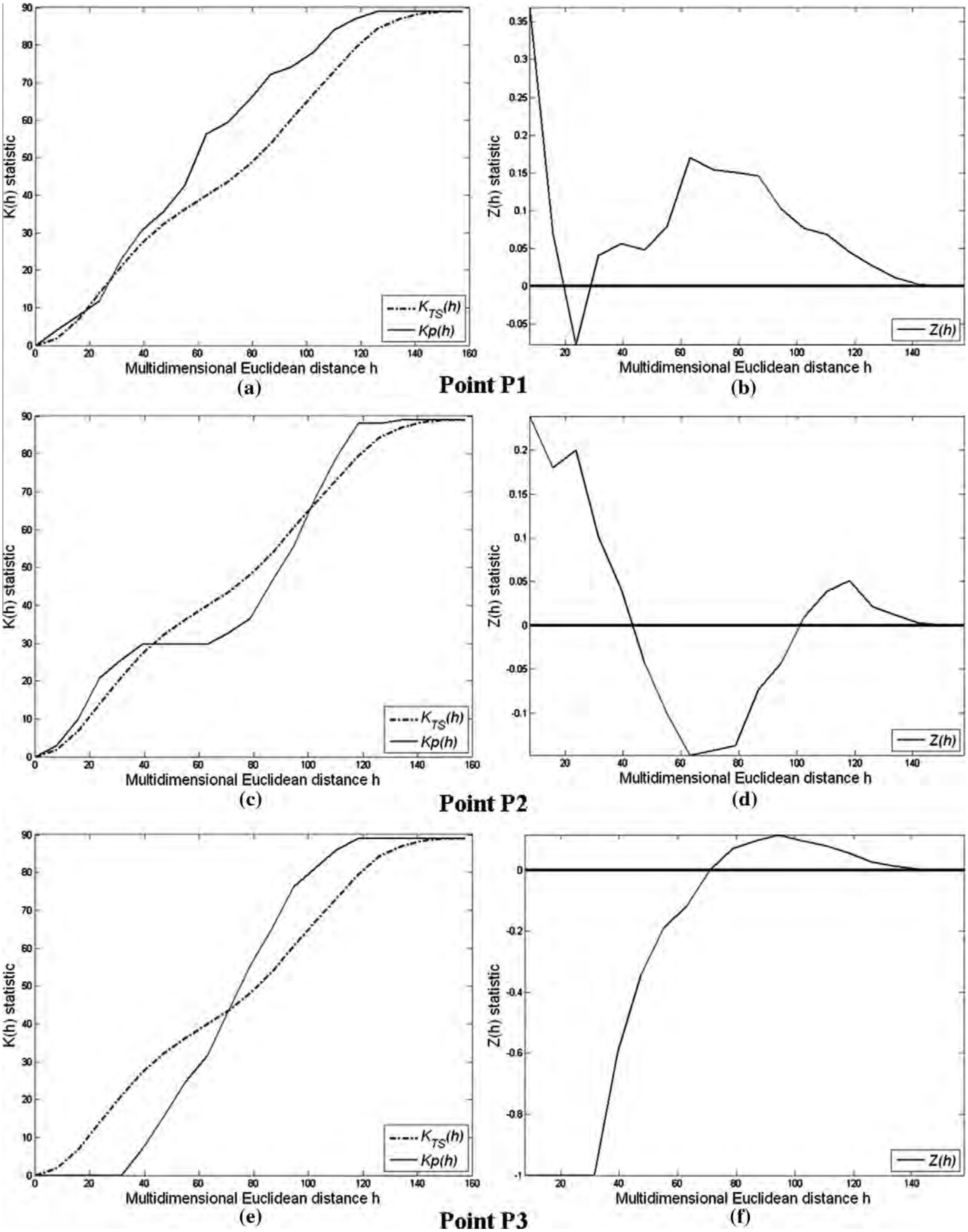

Fig. 4. $K(h)$ and $Z(h)$ statistics, with (a and b) corresponding to point $P1$, (c and d) to point $P2$, and (e and f) to point $P3$ of Fig. 3.

3.1. Confidence metrics in a multi-class classification using QuickBird imagery

3.1.1. Experimental setup

For the first study area, QuickBird imagery was used to offer an assessment on higher spatial but lower spectral resolution. The study area is in Las Vegas, Nevada. The study area and the QuickBird imagery are shown in Fig. 5. It is a subset of the QuickBird imagery acquired in May 2003. The main land cover types in this study area include water, buildings, asphalt, grass, and bare soil. All four bands (i.e., blue, green, red and near-IR bands) were used after resampling at the 0.6 m pixel size. The subset consists of 720 × 720 pixels, and covers approximately 0.4 × 0.4 km.

This detailed experiment tested different training dataset size selection, different classification methods and different confidence metric estimation methods.

Table 1
Results of the *C* statistics for the three different cases.

	C_{EQ}	C_{LIN}	C_{G10}	C_{G30}	C_{G50}
*P*1	0.91	0.88	0.85	0.82	0.88
*P*2	0.16	0.30	0.99	0.64	0.28
*P*3	−0.81	−0.91	−1.00	−0.99	−0.91

Training and testing dataset selection. In order to assess the representativeness of various training datasets, each training dataset was created through a combination of representative and non-representative points (pixels). The representative points were expressed as 200 randomly selected points for each of the five classes. From within the 1000 (200 × 5) points, 50 points were randomly extracted for each class leading to a 250 point (50 × 5) testing dataset. This testing dataset was used for classification accuracy assessment (classifier validation) and it was not introduced to any classifiers during the training process. All accuracy results in Figs. 8 and 13 refer to this 250 point dataset. The remaining 750 (150 × 5) points comprise the Random dataset, which is expected to be representative of the entire study area.

In addition to the representative points an intentional effort took place to create non-representative points for each class. These artificially inserted non-representative points were necessary to support confidence variability in the various training datasets and therefore allow examination of the confidence–accuracy relationship under a wide range of conditions. The non-representative points were extracted from a polygon-based training dataset selection process. For each class, 150 reference points were randomly selected from within a single polygon corresponding to that land cover class. The five polygons corresponding to the five classes are overlaid in blue color in Fig. 5. For example, a rectangle was drawn on the water area and randomly sampled 150 points within this rectangle as the reference data for the water class. For simplicity, this dataset is called Limited, which includes 750 (150 × 5) points and is not representative for the entire study area. Hence, two extreme datasets were created of reference points, Limited and Random.

In order to artificially vary the representativeness of training datasets, sample points were randomly extracted from both the Random and Limited datasets. The higher the proportion of sample points from the Random dataset, the more representative is the resulting training dataset (according to Edwards et al., 2006). Each dataset setup is denoted as L%R%, where L corresponds to the Limited dataset, R corresponds to the Random dataset, and the % represents the percentage number of points that were randomly selected from the corresponding dataset. First, the two extreme cases were created, namely L100R0 and L0R100. For L100R0, 100% points were randomly selected from the Limited dataset without any points (0%) from the Random dataset. For L0R100, the situation was reversed, meaning 100% points were randomly extracted from the Random dataset with no points from the Limited dataset. In order to balance the L100R0 and L0R100 cases so that the association between the confidence and the classification accuracy can be fully examined, four intermediate cases were considered. For example, L75R25, suggests that 75% points of this dataset were randomly selected from the Limited dataset and 25% points were randomly selected from the Random dataset. After following this procedure six total dataset types were created with different representativeness, namely L100R0, L90R10, L75R25, L50R50, L25R75, and L0R100 moving from lower to higher representativeness.

Three training dataset sizes were tested with 100, 250, and 400 points using a stratified sampling per cover type as discussed above (equal class presence for all datasets). This setup translated into 20, 50 and 80 points for each of the five land cover classes, which were vegetation, water, road, bare soil, and building, respectively. For each of the three training dataset sizes, the aforementioned six types of training datasets with different representativeness were evaluated.

Classification methods. Multiple classification methods were implemented, namely Decision Tree (DT) (Quinlan, 1986), Maximum Likelihood Classifier (MLC), Backpropagation Neural Network (BNN) (Rumelhart et al., 1986) and Support Vector Machine (SVM, see Mountrakis et al. (2011) for a review). For consistency, none of the methods used cross-validation techniques during training. The DT implemented used full trees without any pruning. For the MLC a uni-modal Gaussian method was applied (i.e. applying one Gaussian model per land cover class). Two versions of the BNN classifiers were implemented, one with fixed, simple architecture and one with variable complex architecture. The simple BNN version was comprised of a fixed network architecture comprised of four node input layer, one hidden layer with three nodes, and an output layer with five nodes, each corresponding to a single

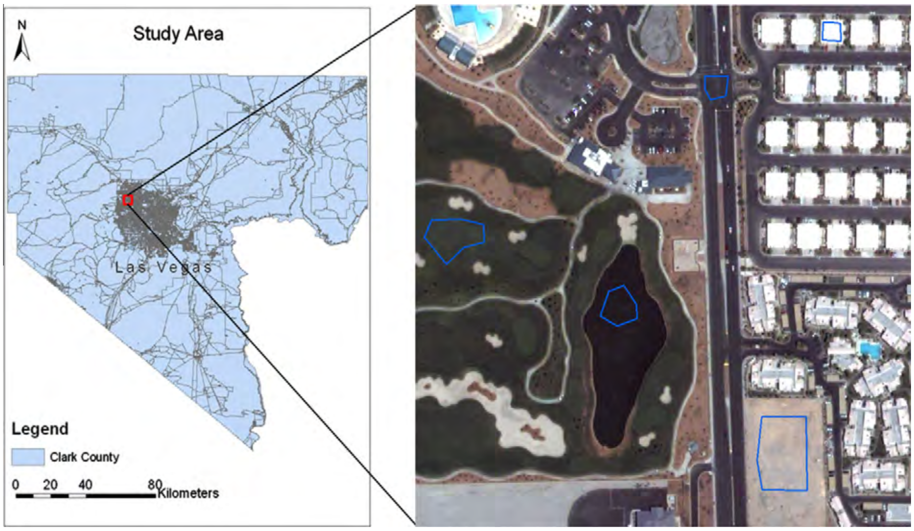

Fig. 5. Left: The first study site in Las Vegas, Nevada. Right: QuickBird imagery, acquired in May 2003, bands 3 (red), 2 (green), and 1 (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

class. The complex BNN version was comprised of the same input layer, but a first hidden layer with varying nodes from 3 to 15, a second hidden layer with nodes ranging from 0 to 6, and the same output layer with five nodes. The motivation behind using a simple and a complex BNN version was to assess performance under algorithms of varying complexity. The internal node functions for both BNN versions were hyperbolic tangent sigmoid transfer functions for the hidden layer(s) while the output layer was comprised of logistic sigmoidal functions. The winning class was assigned as the highest response node on the output layer. To avoid convergence errors and also to incorporate a procedure similar to other BNN practical implementations, for each training dataset, 100 networks were produced, and the one with the highest training accuracy was selected. Finally, the SVM algorithm used radial basis functions as the underlying kernel functions and C-SVC as the basic classification model. A grid search trying different pairs of the cost and gamma parameters took place for a subset of each dataset size. The identified parameters used for all simulations were found for the 100, 250 and 400 point training sizes with cost parameter set at 2^3, 2^5, 2^{11} respectively while the gamma value was set to 2^{-15} for all datasets.

Confidence calculation methods. Three different weight schemes (Equal, Linear, Gaussian "G10") were examined to produce the confidence metrics (see Section 2.1). The C_{global} was calculated using Eq. (11) from Section 2.1 with Q_i the default value 1, which assigned equal weight for all pixels.

Overall setup. With datasets and methods configured, the confidence metrics and classification methods were applied on each training dataset to obtain the C_{global} values and overall testing accuracy. For example, when extracting the training dataset with L90R10 type and 100 points size, 18 points (90% of 20) were randomly selected from the Limited dataset and two points (10% of 20) from the Random dataset for each land cover class. This process was repeated for 100 times and created 100 different training datasets with the same type and size to avoid bias in the randomized dataset selection. The confidence metric C_{global} value was calculated using each of these 100 training datasets. Also, a classifier was trained; the overall testing accuracy was estimated on the same testing dataset. The relationship between confidence and accuracy metrics is presented in the next section.

3.1.2. Results

For each training dataset size (100, 250, or 400 points) and each weighting scheme (Equal, Linear, Gaussian "G10") a collection of point clouds showing the relationship between confidence metrics and the classification accuracy was obtained. In Fig. 6 the relationship between confidence and classification accuracy is presented for a training size of 250 points using a decision tree as the underlying classifier and the Equal weighting scheme for the confidence metric calculation. The *X* axis represents the C_{global} values and the *Y* axis shows the overall testing accuracy of each classifier in the same testing dataset. Points resulting from the same training dataset type (e.g. L90R10) are grouped together using the same color and shape. Starting with the lower left point cloud (displayed as black diamonds) that corresponds to the L100R0 training dataset type, it is evident that low confidence metrics are associated with poor classification accuracy for the classifier (average overall testing accuracy is close to 70%). As other training dataset types are examined, increases in confidence metric values result in classification accuracy improvements, suggesting that the main hypothesis of this work may be valid.

In order to see whether this relationship holds true for different classification methods and weighting schemes, 45 graphs were produced (5 classifier types × 3 weighting schemes × 3 training dataset sizes). For visualization purposes the point cloud graphs are aggregated in 0.05 confidence range bins on the *X* axis (C_{global} values). For all points falling within a given confidence bin, the mean and standard deviation of the overall testing accuracy are displayed. In addition, at the top of each graph the mean (triangle) and three standard deviations (solid circles) are added representing confidence values exclusively from the L0R100 training datasets (e.g. only the red triangles of Fig. 6). The latter mean and standard deviation values relate to the *X* confidence axis with no reference to the *Y* accuracy axis. The resulting graphs are organized in Figs. 7–11 corresponding to the five classification methods.

A general association between confidence values and classification accuracy exists across all graphs. Following the mean values of every bin confidence values typically increase from the L100R0 datasets to the L0R100 dataset. More importantly, increases in global confidence values typically result in classification accuracy gains. This is especially evident as the classification method progressively gets more complex. For the MLC, accuracy gains are observed in increases of small confidence values, however there also seems to be a saturation point where accuracy values stop increasing or slightly decrease. Since this behavior is only present in the MLC and not in all the other more complex methods tested, one possible explanation is the limited modeling capabilities of MLC, however further investigation would be necessary for a concrete conclusion. On the other hand, complex classifiers demonstrate accuracy gains for improvements across the entire range of confidence values, which indicates a strong and consistent relationship between confidence and accuracy. This is desirable because confidence metrics can be used to select the most appropriate reference dataset before any labeling or classification takes place. The advanced modeling capabilities of the DT, the two BNN and the SVM classifiers suggest that they took further advantage of the discriminatory power of the training dataset, especially the complex BNN and the SVM classifiers (i.e. they "learned" more than the MLC). Another general observation is that the standard deviation on the overall testing accuracy tends to decrease as confidence values increase, providing further evidence on the validity of the proposed confidence metrics. On occasion, the standard deviation is small on the smallest confidence bin (leftmost bin) due to limited number of sample points falling in that confidence bin.

Looking into the specifics, the produced graphs can be grouped based on the training dataset size, the weighting scheme and the classification method. By comparing graphs vertically within each figure, we allow variable training dataset sizes while keeping the classification method (e.g. DT) and the confidence weighting scheme (e.g. Linear) constant. A first observation is that as the training dataset size increases, the standard deviation on the overall testing accuracy decreases for all confidence bins. This indicates the overfitting problem is reduced; which, in turn, implies information quality within the training datasets is increased.

Another observation is that accuracy values (*Y* axis) increase in small confidence values at a faster pace for larger training dataset sizes. Furthermore, the values and range of confidence values vary significantly for each weight scheme, with smaller ranges and higher values for Equal weights, and to larger range and smaller values moving to Linear and Gaussian weights. The smaller values of the Linear and Gaussian weights are expected as higher weight is placed on finding neighboring training points in close proximity. The range is a reflection of the specific content of the training image indicating possible clusters in the feature space of the training dataset (i.e. distinct foot print of each land cover class). This variability of confidence values suggests that confidence values cannot be examined as absolute metrics but a standardized benchmark should be developed.

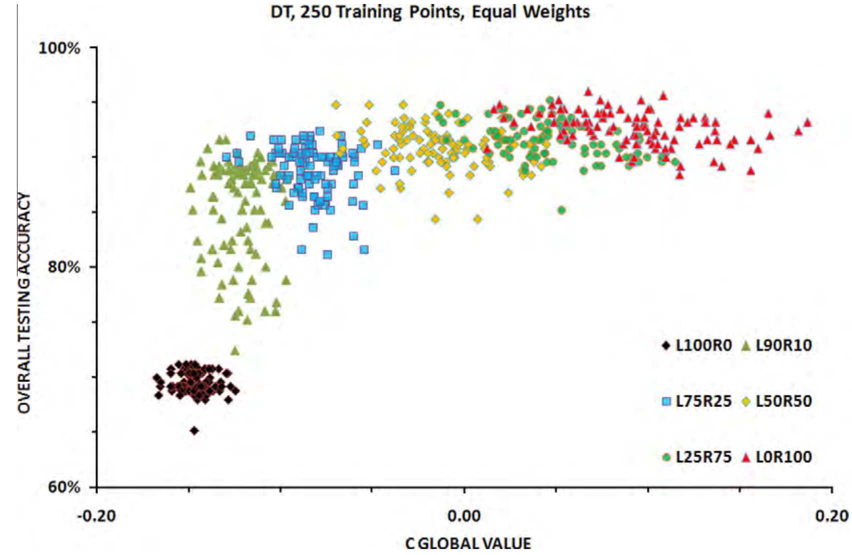

Fig. 6. Point cloud graph showing an example of the relationship between confidence metrics and the classification accuracy.

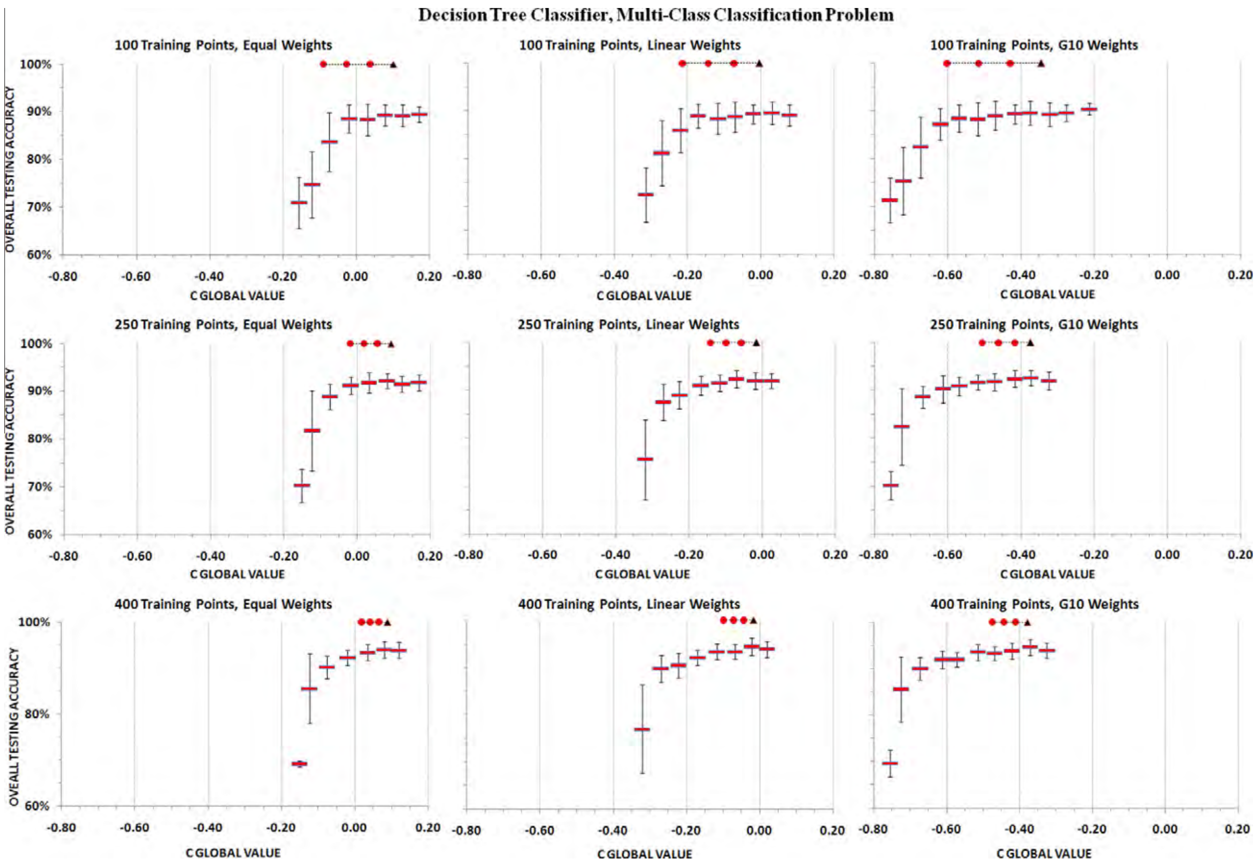

Fig. 7. The confidence–accuracy relationship generalized plots for a Decision Tree classifier using multiple training dataset sizes and confidence weight schemes.

This confidence benchmark is expressed as the confidence variability of the L0R100 dataset, provided on the upper portion of each graph as a triangle for the mean value along with three solid circles expressing the three negative standard deviations. These statistics capture the variability of a dataset exclusively using spatially random training points for the specific combination of training size, method and confidence weight scheme. The mean value can be used as the origin and the standard deviation as a scaling factor. Values of the mean are consistent if we examine vertically each figure (same weight scheme and classification method). As the training size increases mean values are similar but the standard deviation decreases. This is consistent with expectations as larger training sizes would lead to more representative datasets and therefore higher confidence values.

The confidence benchmark created from the L0R100 datasets can be contrasted with the confidence–accuracy relationship in

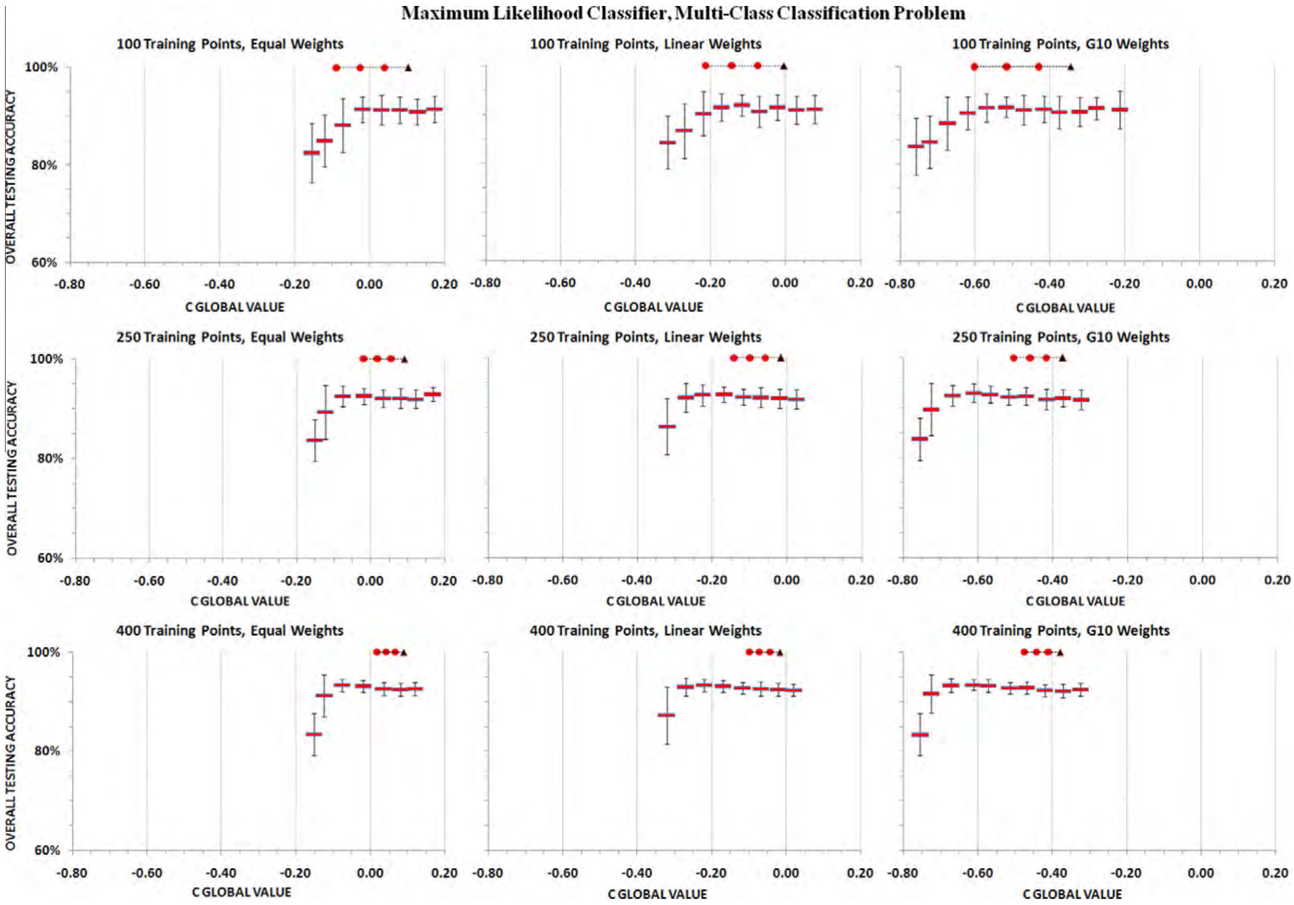

Fig. 8. The confidence–accuracy relationship generalized plots for a Maximum Likelihood Classifier using multiple training dataset sizes and confidence weight schemes.

Figs. 7–11. As classification complexity increases moving from DT to simple BNN to complex BNN and SVM, the critical confidence value (where accuracy gains become limited) progressively moves from mean minus two standard deviations to the mean minus one standard deviation to the mean value, respectively. This observation is generally observed across weight schemes and training sizes suggesting that the benchmark is an effective method to normalize confidence value variability. Even though this benchmark is consistent, its interpretation should vary depending on the classifier's complexity. Higher classifier complexity increases classification ability and therefore classifiers can identify small changes in the spatial distribution of a training dataset allowing accuracy improvements even in large confidence values. Therefore when the benchmark is compared to an obtained confidence value classification complexity should play a role in the interpretation (i.e. expect sustained accuracy improvements closer to the benchmark mean as classifier complexity increases). It is our suggestion that the stricter mean benchmark values are used across all classifiers.

3.2. Confidence metrics in a binary classification using Landsat 7 ETM+ imagery

To further investigate whether the confidence metrics depend on classification problem and remotely sensed data source, a binary classification problem was studied. Landsat ETM+ imagery was used with six bands and 30 m Ground Sample Distance (GSD). Landsat imagery is popular due to easy data access and large geographic coverage; it is also supports a good balance between spatial and spectral resolutions that allows distinction of multiple land cover types (Civco and Hurd, 1997). The image was from Onondaga County, New York with low spatial while high spectral resolution when contrasted with the previously examined Quickbird image.

Fig. 12 shows the second study area, it is a small portion of a Landsat scene and consists of 240 × 180 pixels covering approximately 7.2 × 5.4 km. This study area was mainly used for testing a binary classification of impervious surfaces, see Luo and Mountrakis (2010) for further information on this dataset. The six bands were used directly after a linear stretching was applied independently in every band.

3.2.1. Experimental setup

For this binary (Impervious vs. Non-impervious) classification problem, we tested one scenario, a DT classifier with the training dataset size of 400 points (translated into 200 points per class). The training dataset setup was similar to the previous experiment. From the training area we randomly selected 800 points (400 points per class) to compose the Random dataset. For the Limited dataset, 400 points from each of two rectangles, one within water area (Non-impervious class) and one within urban area (Impervious class), were extracted from the training area. In addition to the L100R0 and L0R100 types, the balancing intermediate cases of L75R25, L50R50 and L25R75 were implemented and one hundred datasets from each type were assessed. The testing dataset was comprised by 400 random points (200 points per class) from the testing area and there was no overlap between the training and testing dataset.

3.2.2. Results

The confidence–accuracy relationship graphs are shown in Fig. 13. The close link between the confidence metrics and classification accuracy still exists. The overlaid confidence benchmark has

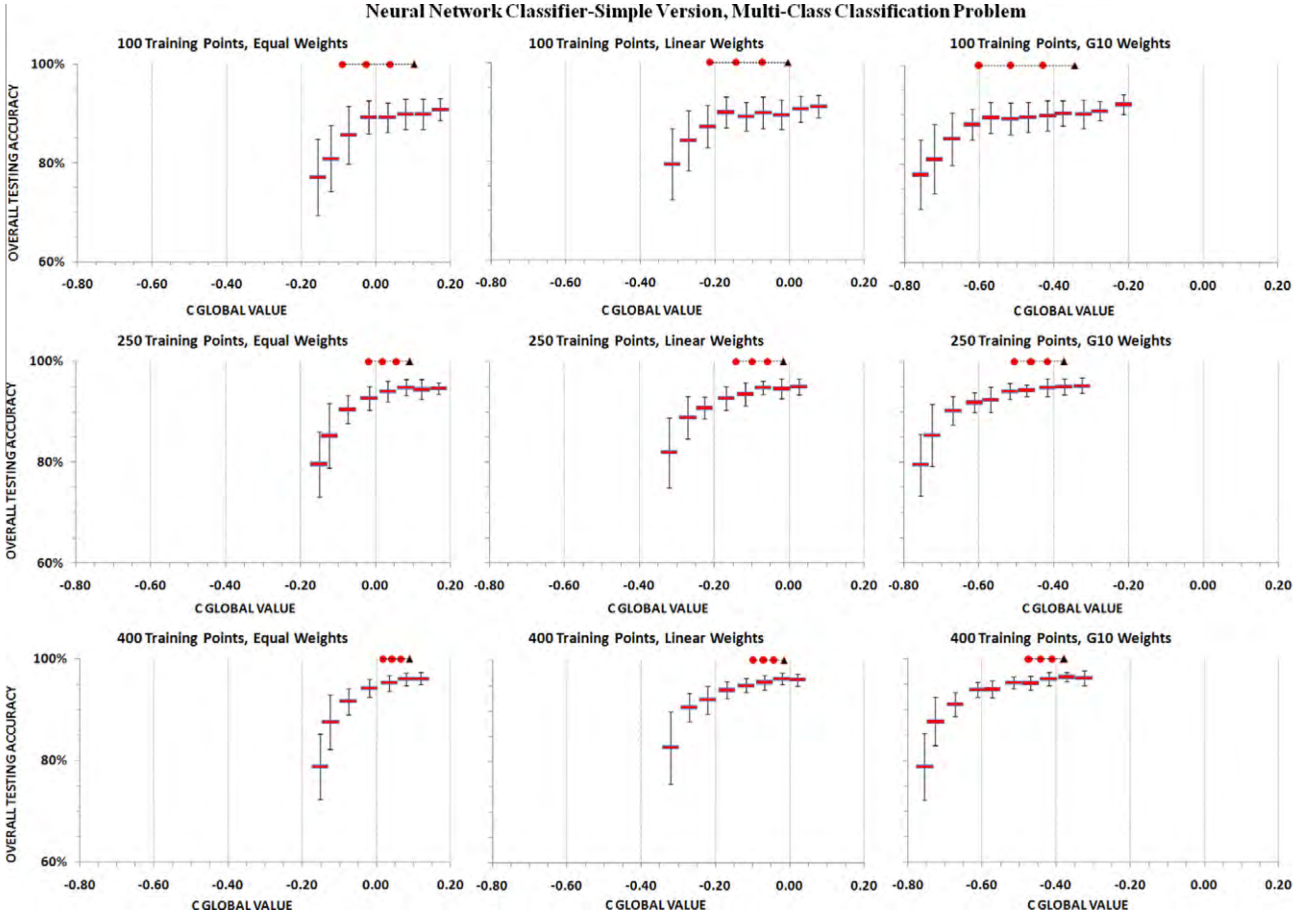

Fig. 9. The confidence–accuracy relationship generalized plots for a simple Neural Network Classifier using multiple training dataset sizes and confidence weight schemes.

the ability to identify confidence values with high accuracy performance, indicating the applicability of the proposed confidence metric on different image data sources and classification problems.

4. Sensitivity analysis example for training dataset selection

The previous section established a close link between the confidence metrics and the classification accuracy. This link is important because confidence metrics offer the significant advantage that they are calculated solely on the input features (e.g. spectral information), in other words, they do not require pixel labels. Thus, the confidence metrics can be calculated for any candidate image portion, which allows us to assess its suitability as a training (or testing) dataset based on how representative it is of the overall image. The higher the confidence value, the more representative the dataset is.

The next natural task is to investigate how to implement the proposed confidence metrics for dataset selection. Fig. 14 shows the third study area, which is in Ross County, Ohio. The Landsat scene was acquired on April 05, 2000; a 900 × 600 pixel portion of the image covering the study area was extracted. This study area covers a region of approximately 27 × 18 km and is characterized by typical land cover types, including water bodies, trees, grass, agriculture, bare soil, and urban areas. This area was partitioned into two parts, with the left half (900 × 300 pixels) as the training confidence area and the right half (900 × 300 pixels) as the testing confidence area. It should be clarified that the training area in this section refers to the portion of the image where training samples can be extracted from. Different training datasets are extracted from the training area and each is contrasted with the entire testing area to assess the training dataset's confidence. In this section there is no classification taking place and no labeled pixels were used. Instead, we use this process as a guide to find the most representative samples so later labels are assigned to them and classification training could take place.

Utilizing as an example case the above Landsat study site, confidence metrics were extracted by varying the sampling scheme and the sampling size. For all experiments in this section the Linear weighting scheme was used to calculate the confidence metric and evaluate training dataset representativeness.

4.1. Sensitivity of different training data sizes using single block training

In this experiment confidence metrics were evaluated for training datasets of variable size. Using a moving block of fixed size the entire training area in Fig. 15a (left half) was scanned without any overlapping allowed between blocks. At each location all pixels within that block were selected as the training samples. Based on the selected training dataset, the C metric was calculated for each pixel contained in the testing area (Fig. 15a, right half) and the associated unweighted ($Q_i = 1$ in Eq. (11)) average as expressed by the C_{global} statistic. Hence, every block acting as a potential training dataset was associated with a single C_{global} statistic. Three block sizes were tested, 10 × 10, 20 × 20, and 30 × 30 using 100, 400, and 900 training pixels, respectively. This sampling scheme will be referred to as BLOCK for the remainder of the paper, associated with the training sample size, for example BLOCK100. Sample

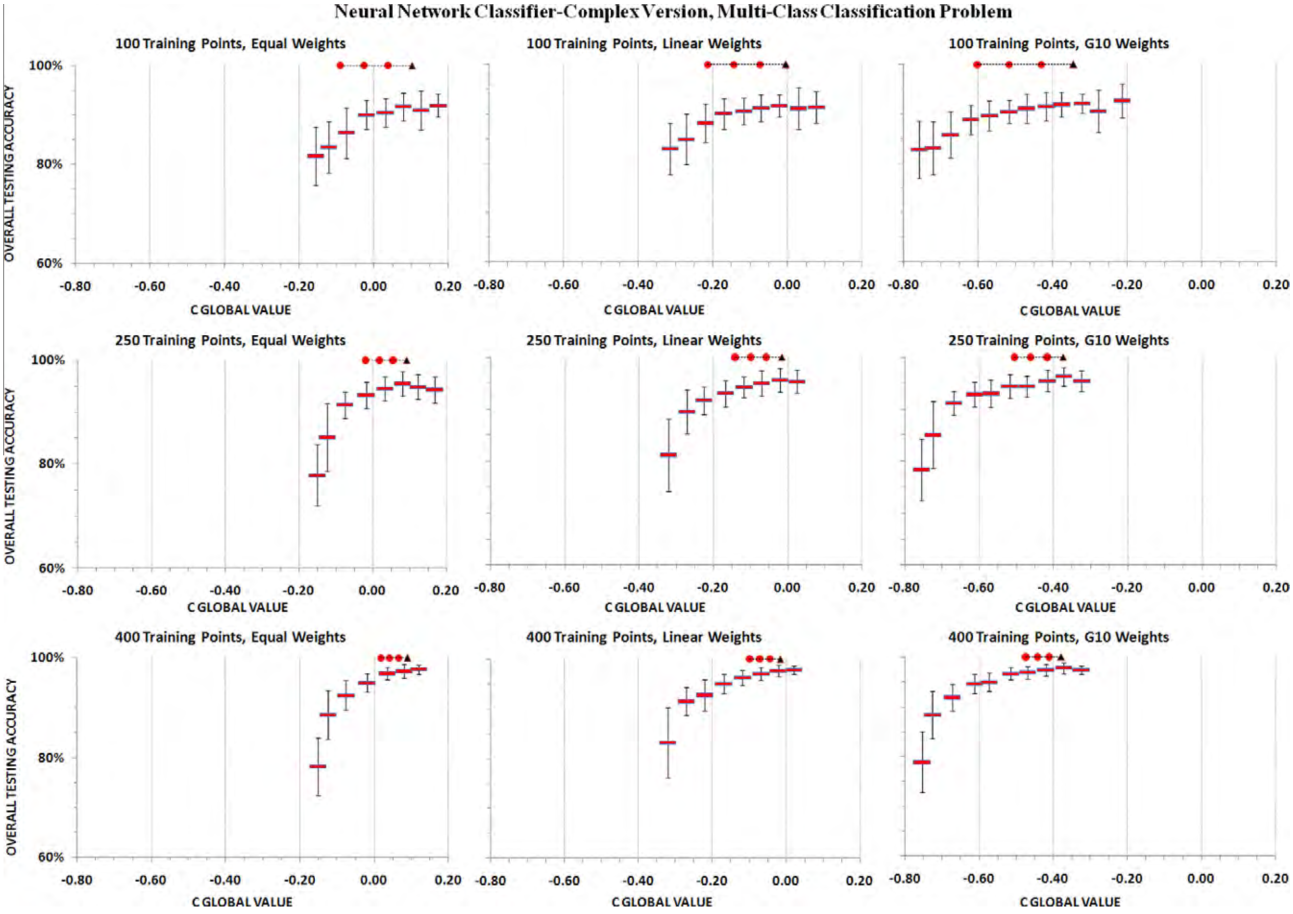

Fig. 10. The confidence–accuracy relationship generalized plots for a complex Neural Network classifier using multiple training dataset sizes and confidence weight schemes.

blocks are typically used for classification accuracy assessment of the National Land Cover Data (Homer et al., 2007).

As each training block produced a C_{global} value to quantify its total representativeness for the testing area, a visualization map was produced showing each window's C_{global} value. Fig. 15b–d shows the C_{global} spatial pattern maps for all the three sampling scenarios of 100, 400 and 900 points. These spatial pattern maps correspond to extracting training blocks from the training area in Fig. 15a. Each training block is treated as a pixel of the spatial pattern map and stores the C_{global} value it generated. Most of the training blocks in Fig. 15b, where the BLOCK100 training selection scheme is implemented, have low confidence values. In Fig. 15c and d, as the block size becomes larger, more training blocks change to pink, which represents higher confidence value (note the negative sign). The spectral image of the training block that generated the maximum C_{global} value for each block size scenario is shown in Fig. 15e–g, respectively. These training blocks include several land cover types and they are identified as highly representative blocks. In addition, using these visualization maps, a better understanding of the spatial heterogeneity of the landscape can be obtained assisting further in training dataset selection.

Since the training area was 900 × 300 pixel and no overlapping was allowed between windows, BLOCK100 scheme had 2700 unique blocks (90 × 30 possible unique locations for a 10 × 10 block in a 900 × 300 image), BLOCK400 had 675 unique blocks and BLOCK900 had 300 unique blocks. In order to compare results from each BLOCK size, the mean, standard deviation and maximum C_{global} value were calculated for all blocks, representing the BLOCK100, BLOCK400 and BLOCK900 sampling schemes. Fig. 16 presents statistically the resulting confidence variability for each training sample size. In essence, each of the three plots in Fig. 16 summarizes the corresponding values obtained in Fig. 15b–d, respectively. In addition, a two-sample Student's *t*-test assuming unequal variances was performed pairwise on the three BLOCK sampling scenarios to test the null hypothesis there is no difference between two sampling scenarios. The largest *P* value was 5.94E−26 indicating a significant difference in the confidence results of the three sampling scenarios.

During training it is expected that a user will be interested in the best performing block rather than the distribution itself. Therefore of particular interest for the training dataset selection process is to identify the single best performing block in terms of confidence value. In this particular experiment the maximum C_{global} values offered no significant difference. This implies that an equally representative training dataset can be obtained for all three sizes for the given image, if the windows are carefully selected. This is highly desirable due to the large acquisition, field visits and processing cost associated with larger datasets.

4.2. Sensitivity of different spatial sampling schemes for training data collection

The previous section examined the effect of training size using a single spatial sampling scheme, a single square block. In this section we expand on the above experiment by adding two additional spatial sampling schemes, the RANDOM and SYSTEMATIC schemes. The BLOCK scheme was identical to the previous section. Results from the three BLOCK sizes are identical in Figs. 16 and 17. Building on that, we randomly selected 100, 400 and 900 training points from the training area and repeated the training selection for 100

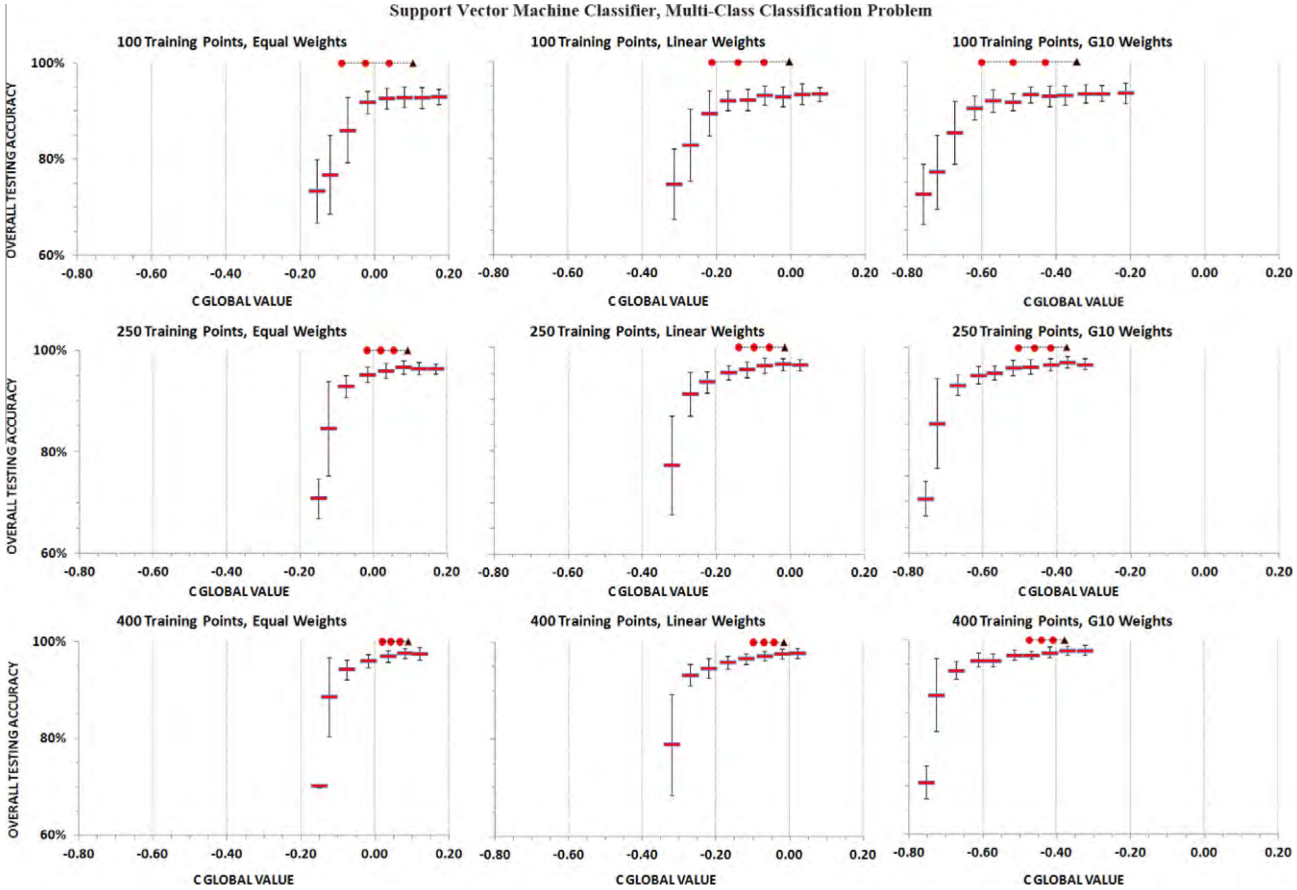

Fig. 11. The confidence–accuracy relationship generalized plots for a Support Vector Machine classifier using multiple training dataset sizes and confidence weight schemes.

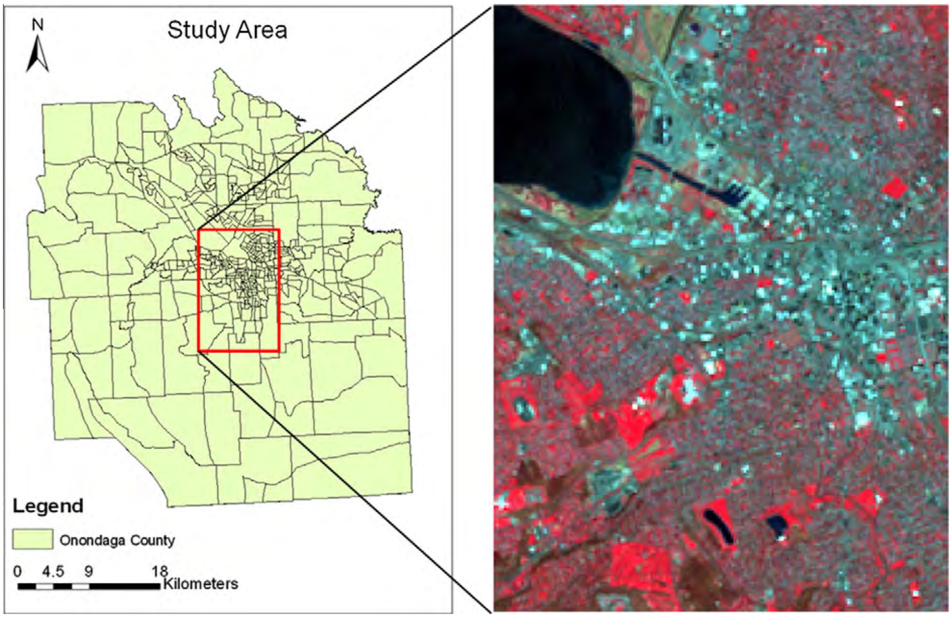

Fig. 12. Left: The second study site in Onondaga County, New York. Right: Landsat-7 ETM + imagery (path 15, row 30), acquired in April 2001, bands 4 (red), 3 (green) and 2 (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

times for each of the three sizes. These three cases were named RAND100, RAND400, and RAND900, respectively. For the SYSTEMATIC sampling scheme, systematic sampling was used to select the same numbers of training points as the BLOCK and RANDOM sampling schemes. The training area was split symmetrically into four sub-areas (each one comprised of 450 × 150 pixels), and scanned each sub-area using a moving block from the corresponding upper left corner without any overlapping allowed. Thus, these four training blocks always had the same sampling interval and each of them had 1/4 number of a training dataset points. For example, for the

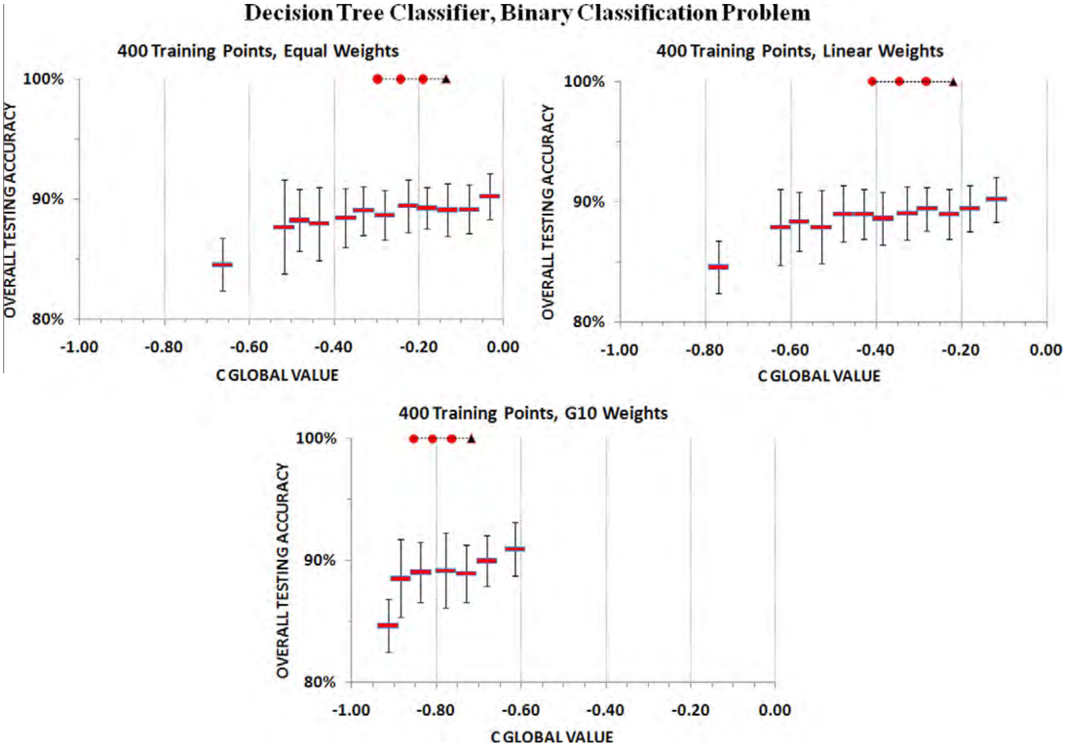

Fig. 13. The confidence–accuracy relationship generalized plots for a Decision Tree classifier on a binary classification problem using multiple confidence weight schemes.

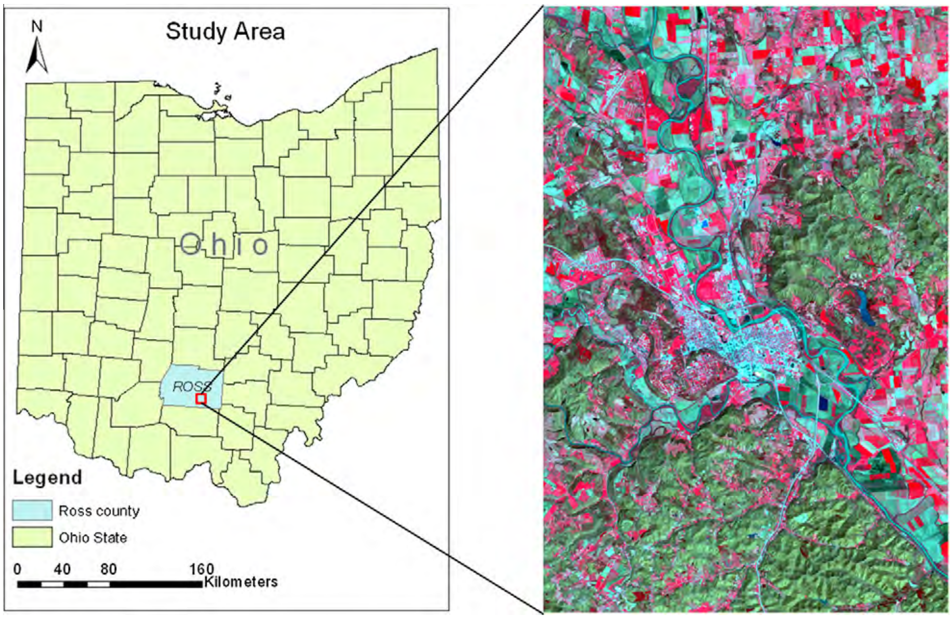

Fig. 14. Left: The first study site in Ross County, Ohio. Right: Landsat-7 ETM + imagery (path 19, row 33), acquired in April, 2000, bands 4 (red), 3 (green) and 2 (blue). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

case of 100 training points four 5 × 5 training blocks were selected from the four sub-areas to constitute a training dataset as one sample. For comparison purposes the three systematic sampling schemes were named SYST100, SYST400, and SYST900. The rationale behind this experiment is to present a potential user with a range of confidence metrics of the different sampling schemes over a given training area.

Similarly to the BLOCK100 scheme, the SYST100 had 2700 unique samples, the BLOCK400 and SYST400 both had 675 unique samples, and the BLOCK900 and SYST900 had 300 unique samples, respectively. For the three RANDOM schemes, 1000 unique samples were obtained for each size (i.e. 1000 samples of the RAND100, RAND400, and RAND900, respectively). As presented in the previous section, the mean, standard deviation and maximum C_{global} value was calculated for different schemes. Fig. 17 presents the resulting confidence variability for each of the nine schemes. Additionally, the first, second and third standard deviations were computed to the lower side of the mean for all the RANDOM sampling scenarios. This corresponds to the same process applied on the Quickbird image in Figs. 9–13.

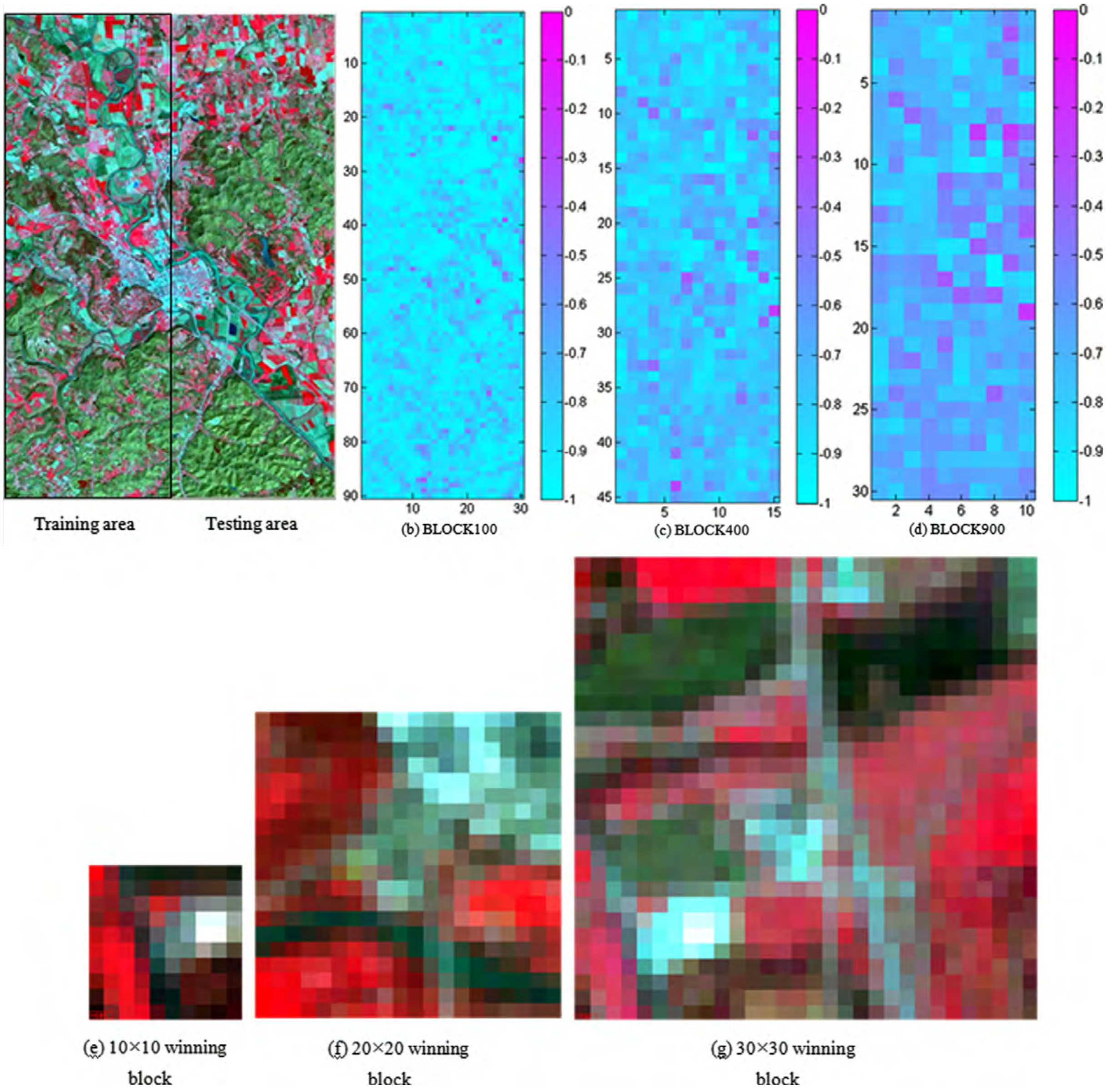

Fig. 15. C_{global} spatial pattern maps associated with the training area and generated using different training block sizes. (a) Study area. (b) BLOCK100 spatial pattern map. (c) BLOCK400 spatial pattern map. (d) BLOCK900 spatial pattern map. (e) Spectral representation of the 10 × 10 winning block. (f) Spectral representation of the 20 × 20 winning block. (g) Spectral representation of the 30 × 30 winning block.

The two-sample Student's *t*-test assuming unequal variances was performed pairwise on the nine sampling scenarios. Within the *t*-test results of all the possible pairs the greatest *P* value was 7.8E−04, indicating that all these pairs were statistically different. In addition, it is clear that the SYSTEMATIC and RANDOM sampling schemes produce, on average, more representative training datasets for every given size, especially the RANDOM sampling scheme. This is no surprise as RANDOM sampling is generally considered a superior sampling method (Edwards et al., 2006).

From the user's perspective, the interest lies in obtaining the single best training dataset. Assuming a training size of 100 or 400 points the BLOCK with the maximum C_{global} confidence (stars on Fig. 17) would on exhibit similar confidence as confidence calculated on average by the corresponding RANDOM method (RAND 100 and RAND 400, triangles on Fig. 17). Combining these results with our previous recommendation from the Quickbird image analysis (that suggested selecting datasets with confidence values close to mean confidence of a corresponding random training dataset) we would suggest these two blocks as replacements for random datasets. In the case of 900 points, where the highest block confidence is close to the mean minus two standard deviations of the random case, we would recommend against the block scheme unless the classifier employed is a fairly simple one (e.g. MLC). For all cases we would support the choice of SYSTEMATIC sampling over the RANDOM sampling since confidence values are sufficiently high. It should be clarified that the suggestions on this paragraph apply on this particular image for the distribution shown in Fig. 17; different distributions from different images may lead to different recommendations.

5. Incorporating training dataset confidence in remote sensing products

The proposed methodology may assist in the classifier training dataset selection, as demonstrated in Section 4. Similarly, it may assist in the classifier testing dataset selection. Another important

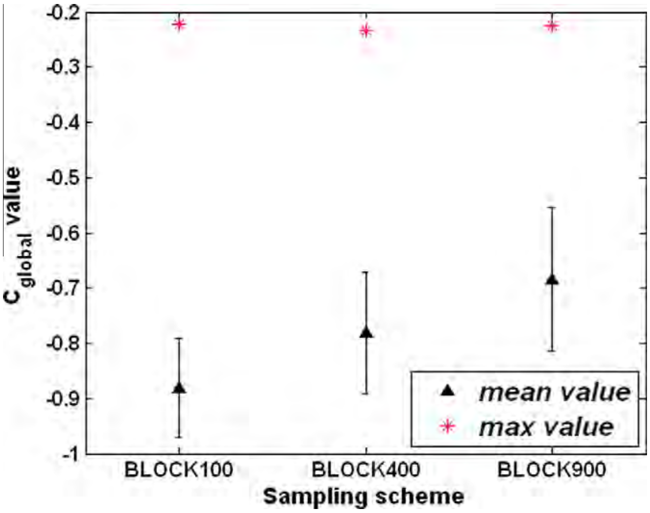

Fig. 16. Mean, standard deviation and maximum values of C_{global} for the training datasets generated by the BLOCK sampling scheme on three different dataset sizes.

function is the ability to express confidence for a classified product. At the general level the C_{global} value can be used as a summary statistic expressing the average confidence values of all pixels. Moving a step further, the spatial distribution of confidence values can be examined visually through confidence maps to assess the representativeness of the training dataset in different image regions. Fig. 18 shows the confidence maps with the maximum C_{global} values for the three sampling schemes with 400 training points (i.e. the three datasets leading to the star confidence values in Fig. 17 for 400 points). Fig. 18a is the study area in which the right part is used to test the representativeness of the selected training datasets. For the confidence maps, as discussed in Section 2.1, the confidence values range from −1 to 1 and the color scale changes from blue to pink according to the confidence values. Fig. 18b shows the best result from all the training datasets selected by the BLOCK400 sampling scheme. The urban area and forests in the middle part of the testing area have higher confidence values, whereas the water bodies, agriculture areas, and forests located in the southern portion are not well represented. Fig. 18c is the best SYST400 confidence map. Forest areas in the middle and lower portions have greater confidence values, while the confidence values for the water bodies and most agriculture areas are still low. Fig. 18d presents the result from training dataset of the RAND400 scheme with the highest confidence value. This confidence map looks similar to the one produced by SYST400 in Fig. 18c. However, in Fig. 18c confidence values are more uniformly distributed, while in Fig. 18d there is strong bias towards forested areas. Confidence maps, such as Fig. 18, should accompany mapping products as they contain important, spatially-explicit information that currently is not communicated to users.

6. Discussion and conclusions

There has been substantial work in the remote sensing field on accuracy assessment. From the initial introduction of error matrices (Card, 1982; Czaplewski and Catts, 1992) to more advanced methods (e.g. Pontius, 2000; Stehman, 1999, 2004; Congalton and Green, 2009. Pontius and Millones, 2011) it has been an active research topic. A typical accuracy metric relates a subset of an image, the reference dataset, with classification results to obtain a level of correctness in the classifier's performance. Different sampling methodologies have been proposed ranging in randomness and stratification techniques (Shine and Wakefiled, 1999; Chen and Stow, 2002; Steele, 1998; Stehman, 2009a,b). However none of these sampling techniques takes into account the actual spectral distribution of all pixels, instead it solely operates in the spatial domain (e.g. systematic or clustered sampling).

There is a clear need to quantify how representative the reference dataset is with respect to the entire image. This effort should not compete with accuracy metrics, instead it should complement them (Fig. 19).

Furthermore, a close link between the proposed confidence metrics and the classification accuracy of a classifier was established in the experiments. This suggests that the proposed dataset selection process can lead to actual gains in classification performance. The method offers a significant advantage: no pixel labels are required when evaluating reference datasets for a given image. Direct incorporation of the proposed metrics is feasible in the supervised classification process to guide and evaluate reference dataset selection, and to supply ancillary information to support remote sensing products.

From the implementation perspective, several recommendations can be provided following the presented experiments. From the confidence calculation viewpoint two recommendations are made, use of Linear weights and consideration of rare class representation. It is clear that the obtained confidence results are dependent on the weight scheme selection. The C_{global} values with Linear weights showed a strong accuracy–confidence relationship and a consistent behavior when compared to values from the random

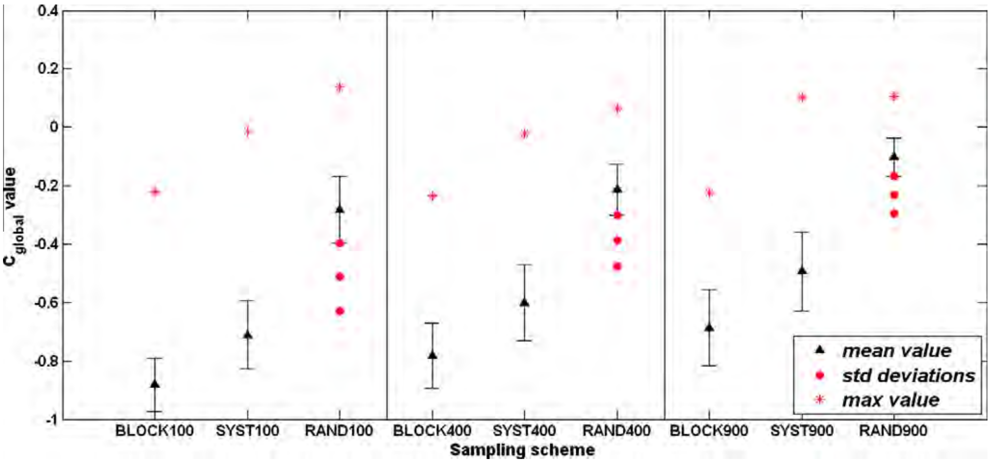

Fig. 17. Mean, standard deviations and maximum values of C_{global} for the training datasets generated by three different sampling schemes and three different dataset sizes.

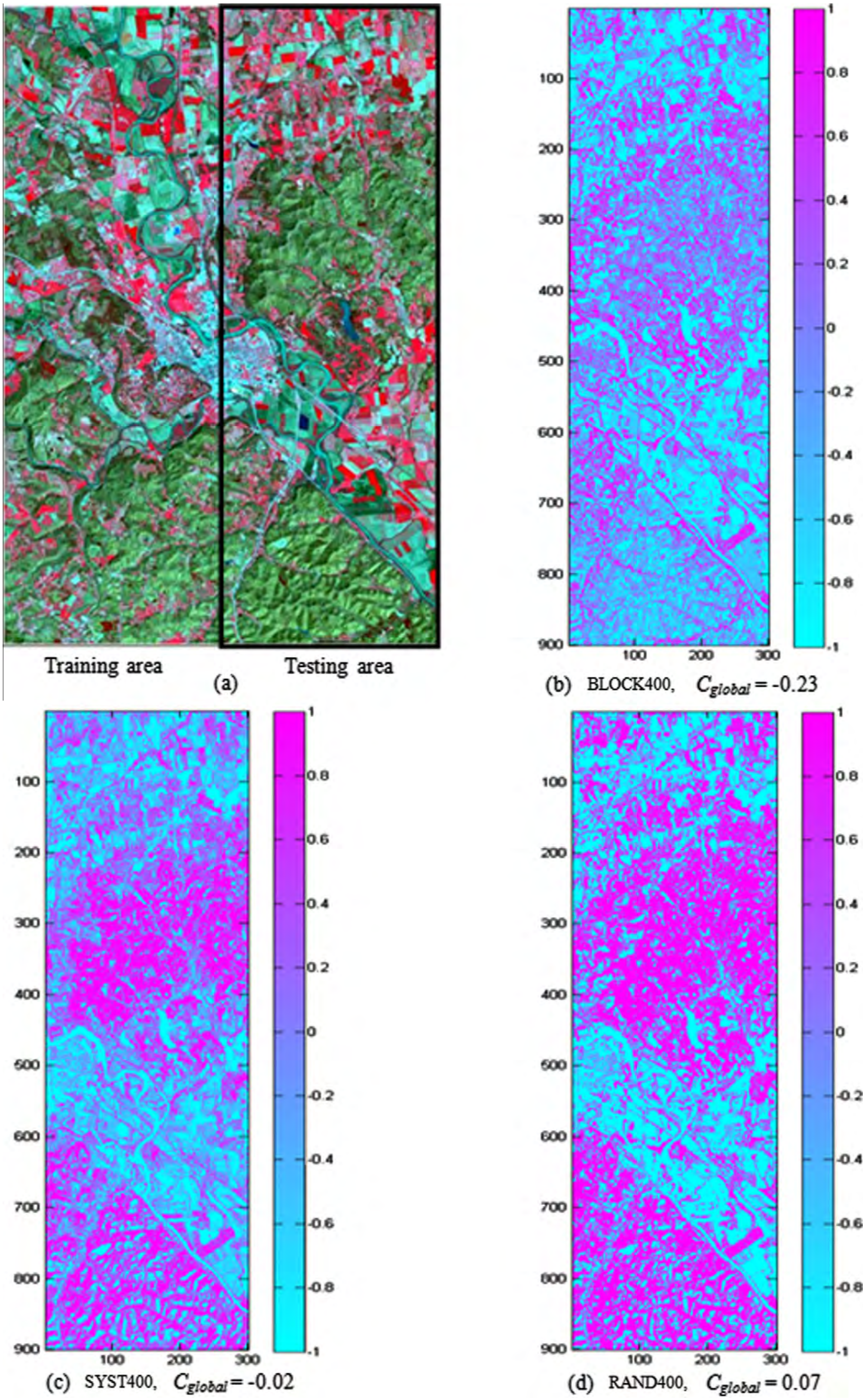
Training area
Testing area
(a)
(b) BLOCK400, C_{global} = -0.23
(c) SYST400, C_{global} = -0.02
(d) RAND400, C_{global} = 0.07
100
200
300
400
500
600
700
800
900
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

Fig. 18. Confidence maps of highest C_{global} 400-point dataset for each sampling scheme. (a) Study area. (b) BLOCK400, $C_{global} = -0.23$. (c) SYST400, $C_{global} = -0.02$. (d) RAND400, $C_{global} = 0.07$.

case benchmark. Furthermore, by design the Linear weights consider information across all scales in the feature space, while providing higher weight in smaller scales. In our experiments we did not make use of the user-defined weight Q_i, to emphasize higher contributions of specific pixels to the overall C_{global} calculation. By doing so, we intentionally chose to examine the image as a

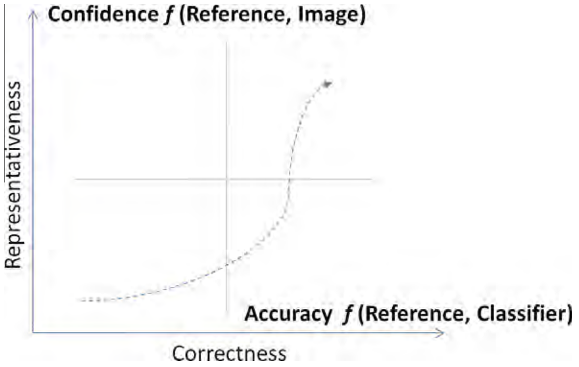
Confidence f (Reference, Image)
Representativeness
Accuracy f (Reference, Classifier)
Correctness

Fig. 19. Communication of classification results using accuracy and confidence metrics.

complete representation independently of class representation. In other words if the image contains a high quantity from a given class that should also be reflected in the training dataset. There may be a case where users are more interested in a rare land cover class. To achieve a strong contribution from that particular class the Q_i weights should be adjusted accordingly. This could take place through an unsupervised clustering process, where rare classes could be identified.

From the classification algorithm perspective, several methods were tested (Figs. 7 and 11). Methods such as the Maximum Likelihood classify pixels from the center of the cluster going outwards, while the opposite holds true for the Support Vector Machine, where the analysis concentrates on the transitional space, the so called edge pixels. The methodology delivered consistent results across all methods independently of the underlying classification mechanism. As expected, the accuracy gains from a more representative dataset are captured better with more sophisticated methods, such as the SVM or BP neural networks (Figs. 10 and 11).

From the sampling scheme evaluation perspective, Section 4 discussed under which circumstances spatially random points can be replaced with spatially neighboring points (e.g. within a block). This finding is significant because the method provides users with a potential alternative, a continuous window-based reference dataset, which may be preferable in cases where reference data are constrained to a small segment of the image (e.g. due to field work or high resolution data acquisition costs).

From the confidence assessment perspective, the reference dataset selection should not be exclusively dependent on the C_{global} statistics, for example the summary metrics provided in Fig. 17. Visualization maps of the spatial distribution of the obtained confidence (such as Fig. 18) should play a complementary and important dual role. First, it may be preferable that a reference site is selected with slightly smaller confidence than the site with the highest average value. This may be a case where the spatial distribution of confidence from one site is more uniform across the entire image. Visualization maps of standard deviations from the mean could also indicate further that spatial variability. Second, visualization maps could be used for further reference dataset refinement. Pixels with lower confidence are easily identified, and targeted additional sampling could be guided to address gaps in the feature space.

In this paper, the method is presented as a guide for optimal reference dataset selection within a single image. An interesting question arises as to whether this metric can be used to compare reference dataset selection from independent classifications on different images. It is well-known that algorithmic accuracy can be easily manipulated with misrepresented information in the reference dataset. Having a consistent confidence metric that is reported along with each classification task would add further validity to the obtained results. The initial signs are positive: through the use of the random case benchmark a standardized approach could be developed. Our tests suggest that when the confidence value reaches the confidence of the average random case the accuracy improvements are limited beyond that point. Further investigation is necessary to develop a standard as a limited number of study sites and sensor types were tested. In addition, we chose to use a subset of popular classification methods, varying from statistical to parameterized. Despite the successful relationships established for algorithms of high modeling capabilities (SVM, DT, BPNN) caution should be exercised until further results become available. This is particularly important for algorithms such as SVM that tend to operate on the ''edges'' between classes, and not necessarily require a complete class distribution for successful application. We view this paper as an initial proof of concept with additional testing necessary for wide-spread adoption. To encourage further development by the remote sensing community the source code is freely available.

Acknowledgements

The authors would like to thank Dr. Steve Stehman for multiple helpful discussions and the anonymous reviewers for improving the focus of our work.

References

Barandela, R., Valdovinos, R.M., Sánchez, J.S., Ferri, F.J., 2004. The imbalanced training sample problem: under or over sampling? Lecture Notes in Computer Science 3138, 806–814.

Binaghi, E., Brivio, P.A., Ghezzi, P., Rampini, A., 1999. A fuzzy set-based accuracy assessment of soft classification. Pattern Recognition Letters 20, 935–948.

Campbell, J.B., 2007. Introduction to Remote Sensing. Guilford Press, New York, p. 626.

Cano, J.R., Herrera, F., Lozano, M., 2007. Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability. Data & Knowledge Engineering 60, 90–108.

Card, D.H., 1982. Using known map category marginal frequencies to improve estimates of thematic map accuracy. Photogrammetric Engineering & Remote Sensing 48, 431–439.

Chen, D., Stow, D., 2002. The effect of training strategies on supervised classification at different spatial resolutions. Photogrammetric Engineering & Remote Sensing 68 (11), 1155–1162.

Civco, D.L., Hurd, J.D., 1997. Impervious surface mapping for the state of Connecticut. In: Proceedings of 1997 ASPRS/ACSM Annual Convention, April 7–10, 1997, Seattle, WA, vol. 3, pp. 124–135.

Congalton, R.G., 1991. A review of assessing the accuracy of classifications of remotely sensed data. Remote Sensing of Environment 37, 35–46.

Congalton, R., Green, K., 2009. Assessing the Accuracy of Remotely Sensed Data: Principles and Practices, second ed. CRC/Taylor & Francis, Boca Raton, FL, 183p.

Czaplewski, R.L., Catts, G.P., 1992. Calibration of remotely sensed proportion or area estimates for misclassification error. Remote Sensing of Environment 39, 29–43.

Edwards Jr., T.C., Cutler, D.R., Zimmermann, N.E., Geiser, L., Moisen, G.G., 2006. Effects of sample survey design on the accuracy of classification tree models in species distribution models. Ecological Modelling 199, 132–141.

ENVI, 1999. ENVI Tutorials; version 3.2. Better Solutions Consulting, Lafayette, Colorado, USA.

Foody, G.M., 1996. Approaches for the production and evaluation of fuzzy land cover classifications from remotely-sensed data. International Journal of Remote Sensing 17, 1317–1340.

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment 80, 185–201.

Foody, G.M., 2009. Sample size determination for image classification accuracy assessment and comparison. International Journal of Remote Sensing 30, 5273–5291.

Foody, G.M., Mathur, A., 2004. Toward intelligent training of supervised image classifications: directing training data acquisition for SVM classification. Remote Sensing of Environment 93, 107–117.

Foody, G.M., Mathur, A., 2006. The use of small training sets containing mixed pixels for accurate hard image classification: training on mixed spectral responses for classification by a SVM. Remote Sensing of Environment 103, 179–189.

Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., Van Driel, J.N., Wickham, J.D., 2007. Completion of the 2001 national land cover database for the conterminous United States. Photogrammetric Engineering & Remote Sensing 73, 337–341.

Jain, A.K., Duin, R.P.W., Mao, Jianchang., 2000. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 4–37.

Kandrika, S., Roy, P.S., 2008. Land use land cover classification of Orissa using multi-temporal IRS-P6 awifs data: a decision tree approach. International Journal of Applied Earth Observation and Geoinformation 10, 186–193.

Kavzoglu, T., 2009. Increasing the accuracy of neural network classification using refined training data. Environmental Modelling & Software 24, 850–858.

Lesparre, J., Gorte, B.G.H., 2006. Using mixed pixels for the training of a maximum likelihood classification. In: Proceedings of ISPRS Commission VII Mid-term Symposium, pp. 632–637.

Lillesand, T.M., Kiefer, R.W., Chipman, J.W., 2004. Remote Sensing and Image Interpretation. Wiley, New York, p. 763.

Liu, W., Gopal, S., Woodcock, C., 2004. Uncertainty and confidence in land cover classification using a hybrid classifier approach. Photogrammetric Engineering and Remote Sensing 70, 963–971.

Lu, D., Mausel, P., Batistella, M., Moran, E., 2004. Comparison of land-cover classification methods in the Brazilian Amazon basin. Photogrammetric Engineering and Remote Sensing 70, 723–731.

Luo, L., Mountrakis, G., 2010. Integrating intermediate inputs from partially classified images within a hybrid classification framework: an impervious surface estimation example. Remote Sensing of Environment 114 (6), 1220–1229.

McCaffrey, T., Franklin, S., 1993. Automated training site selection for large-area remote-sensing image analysis. Computers & Geosciences 19, 1413–1428.

Mitchell, S., Remmel, T., Csillag, F., Wulder, M., 2008. Distance to second cluster as a measure of classification confidence. Remote Sensing of Environment 112, 2615–2626.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: a review. ISPRS Journal of Photogrammetry and Remote Sensing 66 (3), 247–259.

Pontius, R.G., 2000. Quantification error versus location error in comparison of categorical maps. Photogrammetric Engineering and Remote Sensing 66, 1011–1016.

Pontius, R.B., Millones, M., 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. International Journal of Remote Sensing 32 (15), 4407–4429.

Quinlan, J.R., 1986. Induction of decision trees. Machine Learning 1, 81–106.

Raudys, S.J., Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Transactions on Pattern Analysis and Machine Intelligence 13, 252–264.

Rebbapragada, U., Lomasky, R., Brodley, C.E., & Friedl, M.A., 2008. Generating high-quality training data for automated land-cover mapping. In: IEEE International Geoscience & Remote Sensing Symposium, vol. 4, pp. IV-546–IV-548.

Reeves, C.R., Bush, D.R., 2001. Instance Selection and Construction for Data Mining. Kluwer Academic Publishers, pp. 339–356.

Richards, J.A., Jia, X., 1999. Remote Sensing Digital Image Analysis: An Introduction. Springer, Berlin; New York, p. 363.

Ripley, B.D., 1976. The second-order analysis of stationary point processes. Journal of Applied Probability 13, 255–266.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Rumelhart, D.E., McClelland, J.L. (Eds.), Parallel Distributed Processing. MIT Press, Cambridge, MA, pp. 318–362.

Sebban, M., Nock, R., Chauchat, J., Rakotomalala, R., 2000. Impact of learning set quality and size on decision tree performances. International Journal of Computer System and Signal 1, 85–105.

Shahshahani, B.M., Landgrebe, D.A., 1994. The effect of unlabeled samples in reducing the small sample size problem and mitigating the Hughes phenomenon. IEEE Transactions on Geoscience and Remote Sensing 32, 1087–1095.

Shine, J.A., Wakefiled, G.I., 1999. A Comparison of Supervised Imagery Classification Using Analyst-chosen and Geostatistically-chosen Training Sets. GeoComputation' 99. <http://www.geovista.psu.edu/sites/geocomp99/Gc99/044/gc_044.htm>.

Steele, B., Winne, J.C., Redmond, R., 1998. Estimation and mapping of misclassification probabilities for thematic land cover maps. Remote Sensing of Environment 66, 192–202.

Stehman, S.V., 1999. Basic probability sampling designs for thematic map accuracy assessment. International Journal of Remote Sensing 20, 2423–2441.

Stehman, S.V., 2004. A critical evaluation of the normalized error matrix in map accuracy assessment. Photogrammetric Engineering and Remote Sensing 70 (6), 743–751.

Stehman, S.V., 2009a. Sampling designs for accuracy assessment of land cover. International Journal of Remote Sensing 30 (20), 5243–5272.

Stehman, S.V., Foody, G.M., 2009b. Accuracy assessment. In: Warner, T.A., Nellis, M.D., Foody, G.M. (Eds.), The SAGE Handbook of Remote Sensing. Sage Publications, New York, pp. 297–309.

Swain, P.H., Davis, S.M., 1978. Remote Sensing: The Quantitative Approach. McGraw-Hill International Book Co., London; New York.

Van der Wel, F.J.M., Van der Gaag, L.C., Gorte, B.G.H., 1998. Visual exploration of uncertainty in remote-sensing classification. Computers & Geosciences 24, 335–343.

Van Niel, T.G., McVicar, T.R., Datt, B., 2005. On the relationship between training sample size and data dimensionality: Monte Carlo analysis of broadband multi-temporal classification. Remote Sensing of Environment 98, 468–480.

Washer, M.J., Landgrebe, D.A., 1984. A binary tree feature selection technique for limited training sample size. Remote Sensing of Environment 16, 183–194.