doi:10.20944/preprints202201.0352.v1

Pixel Level Classification Confidence for Remote Sensing Imagery: An Evaluation of Three Interpolation Based Methods

Shiguo Jiang

Department of Geography and Planning, University at Albany, State University of New York

Arts and Sciences 227, 1400 Washington Ave, Albany, NY 12222. Email addresses:

sjiang2@albany.edu (S. Jiang).

Abstract:

Obtaining classification confidence at the pixel level is a challenging task for accuracy assessment in remote sensing image classification. Among the various methods for estimating classification confidence at pixel level, interpolation-based methods have drawn special attention in the literature. Even though they have been widely recognized in the literature, their usefulness has not been rigorously evaluated. This paper conducts a comprehensive evaluation of three interpolation-based methods: local error matrix method, bootstrap method, and geostatistical method. We applied each of the three methods to three representative datasets with different spatial resolution, spectral bands, and the number of classes. We then derive the estimated classification confidence and true classification confidence and compared the results with each other using both exploratory data analysis (bi-histogram) and statistical analysis (Willmott's *d* and Binned classification quality). The results indicate that the three interpolation methods provide some interesting insights on various aspects of estimating per-pixel classification confidence. Unfortunately, the interpolation assumes that classification confidence is smooth

across the space, which is usually not true in practice. In other words, interpolation-based methods have limited practical use.

Keywords: Per-pixel classification confidence; spatial pattern; image classification; accuracy assessment; interpolation method

1 Introduction

Reporting classification confidence at pixel level is a challenge to accuracy assessment in remote sensing image classification. Campbell (1981) is one of the earliest work recognizing the spatial variation of classification confidence. It is now widely accepted that classification confidence is neither uniformly nor randomly distributed across the classification map (Congalton, 1988; Foody, 2002; Steele et al., 1998). Factors contributing to the spatial variability of classification confidence include ground features (such as topography and elevation), land cover type and heterogeneity, patch size, and sample design (Burnicki, 2011; Congalton, 1988; Smith, 2002; van Oort et al., 2004; Yu et al., 2008). Irrespective of its source, spatial pattern of classification confidence will propagate to further applications using error-infected maps, and ignorance of the spatial variability of confidence may have negative impact on decision making (McIver and Friedl, 2001; Steele et al., 1998; van Oort et al., 2004). The traditional approach for accuracy assessment is limited in that error matrix and derived indices provide no information on the classification confidence at pixel level. Therefore, it is important to develop methods to identify the spatial distribution of classification confidence and integrate it into analysis afterwards.

Among the various studies for estimating classification confidence at pixel level, some studies use estimate per-pixel classification confidence by interpolating estimation at sample locations to the whole image. In the literature, there are three different approaches which can be named as: local error matrix method (Foody, 2005), bootstrap method (Steele et al., 1998), and geostatistical method (Kyriakidis and Dungan, 2001). In Foody (2005), local classification accuracies are derived based on local error matrices constrained in a local neighborhood. The local accuracies are then interpolated to the whole map using inverse distance squared weighted interpolation, i.e., IDW. In Steele et al. (1998), misclassification rates at sample pixels are obtained through bootstrap resampling. The misclassification rates are then interpolated to the whole map using kriging. Kyriakidis and Dungan (2001) use a more sophisticated method to derive local confusion index from global error matrix. The confusion index is then interpolated to the whole map using simple kriging. The above three methods provide insight on alternative approaches to estimate per-pixel classification confidence.

Despite the fact that these three interpolation based methods have been widely recognized and cited by various studies (see e.g., Burnicki, 2011; Comber et al., 2012; Ebrahimy et al., 2021; Foody, 2002; Löw et al., 2015; van Oort, 2007; Wickham et al., 2018), they have not been rigorously evaluated and tested. In other words, the estimated per-pixel classification confidence has not been evaluated with the true per-pixel classification confidence. An untested method has no warrantee to be correct and effective in practice. In this paper, we evaluate the effectiveness of the above three methods using carefully selected representative datasets. The organization of this paper is summarized as follows. Section 2 describes the methodology for this study which is further divided into four subsections. Section 3 presents the results. Section 4 discusses the results and concludes this paper.

2 Methodology

The methodology of this paper is illustrated in Figure 1. First, remote sensing image is classified to produce class map and estimate classification confidence map (more detail to follow). Second, class map is compared with reference map to create indicator map for error and correct pixels. Third, pixels are grouped into bins based on classification confidence, and binned classification quality is then calculated by averaging pixels in each bin. Forth, the relationship between binned classification quality and classification confidence is examined. The above procedures are applied to three interpolation-based methods with four classifiers and three datasets. In total there are $3 \times 4 \times 3$ scenarios.

The detailed procedures to estimate classification confidence map will be explained in Section 2.2.1-2.2.3 under each of the three interpolation-based methods.

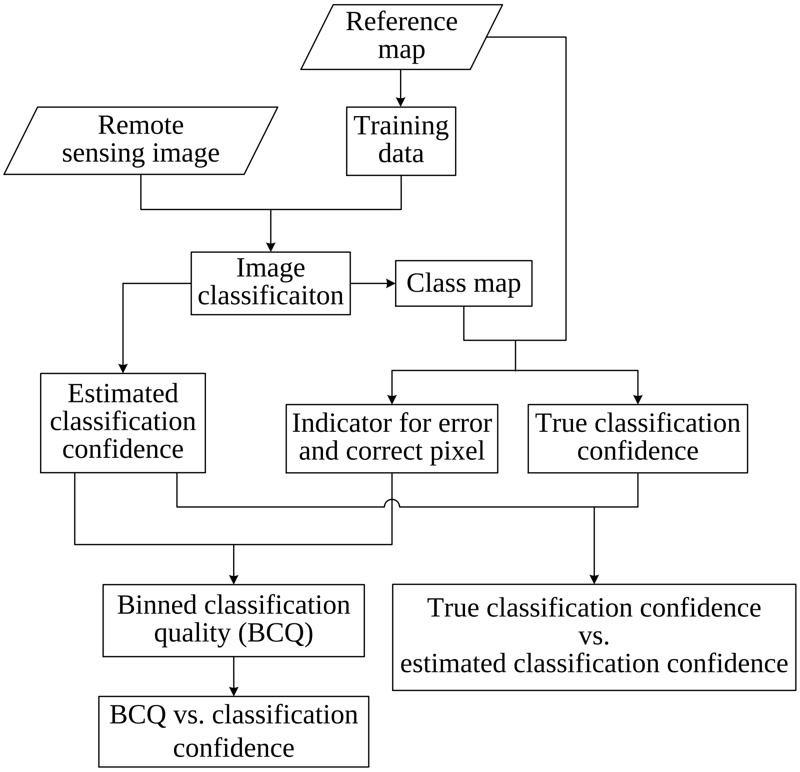

Reference map
Remote sensing image
Training data
Image classificaiton
Class map
Estimated classification confidence
Indicator for error and correct pixel
True classification confidence
Binned classification quality (BCQ)
True classification confidence vs. estimated classification confidence
BCQ vs. classification confidence

Figure 1 Flow chart of the methodology

doi:10.20944/preprints202201.0352.v1

2.1 Test datasets

Three representative datasets with different spatial resolution, spectral bands, and number of classes are used in this study. Complete coverage reference map is developed manually for each dataset, which is critical for testing and evaluating the score-based method. These datasets and corresponding reference maps are shown in Figure 2 and their characteristics are summarized in Table 1.

Table 1 Study site and data

Data Set	Study Site	Sensor	Resolution (m)	Acquisition Date	Image size (pixel)	Figure 2 subplot	
						Image	Reference
1	Kent, MD	TM	30	23/1/2010	500x500	(a)	(b)
2	Oakland, CA	QuickBird	0.6	24/3/2003	500x500	(c)	(d)
3	DC Mall	HYDICE	3	23/8/1995	1280x307	(e)	(f)

(1) Data 1: Kennedyville cropland, Landsat TM image

Data 1 is a study site in the agricultural heartland (39.30^{o} N, -76.00^{o} W) of Kent County, Maryland. The image spread across two villages: Worton and Kennedyville. The land is mainly covered with agricultural field. The north side belongs to the Sassafras River watershed while the south side belongs to the Chester River watershed. Forest covers along both rivers. Landsat image (acquisition date: 23 January 2010) is obtained from USGS Landsat Archives. A clip image of 500x500 pixels is used. Reference map is developed by experienced data analysts and is deemed accurate. Reference map consists of five classes: tree, mature crop, young crop, vacant land, and water.

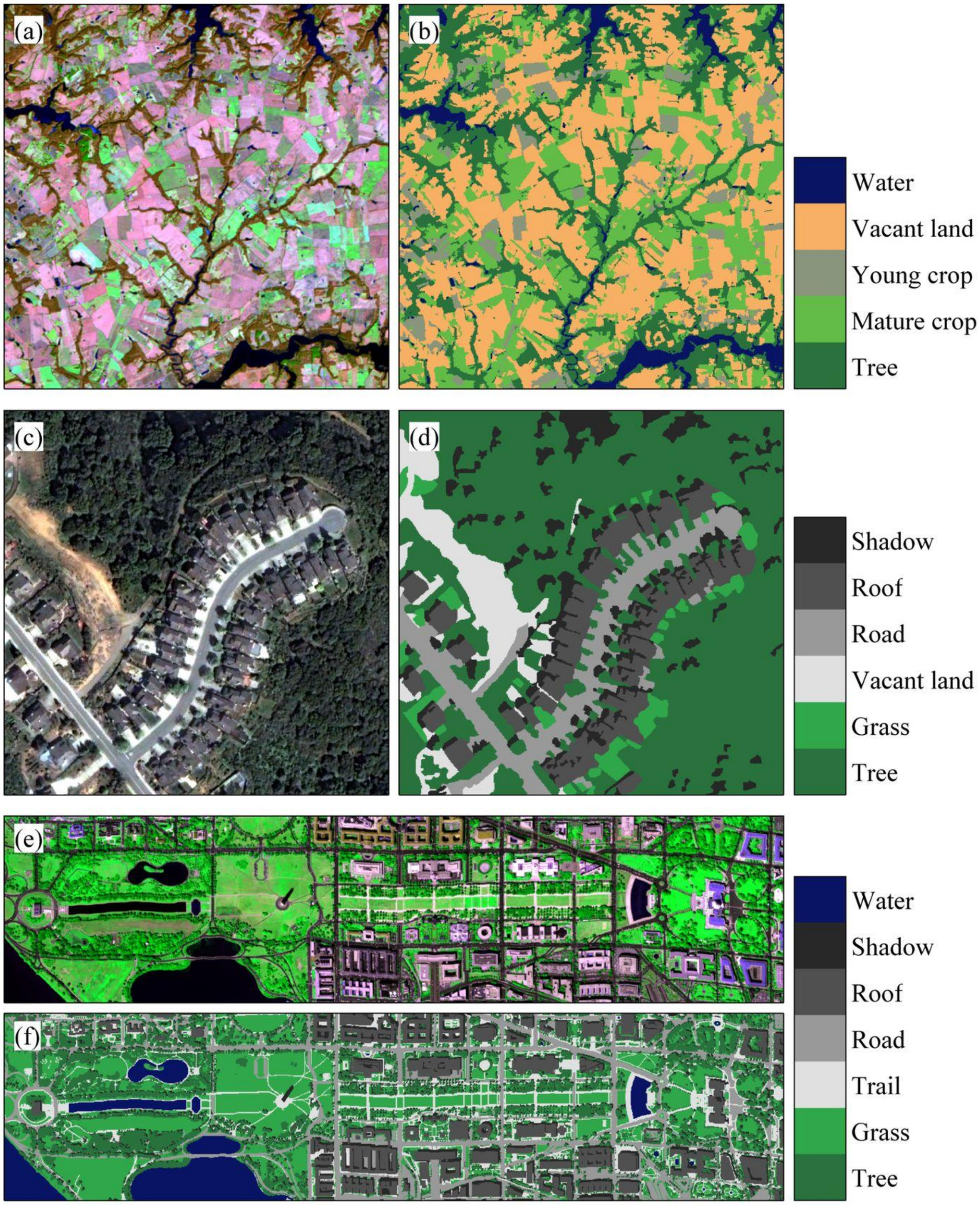


Figure 2 Datasets and corresponding reference map

(a) TM image (7-4-2 band composite), (c) QuickBird image (3-2-1 band composite), (e), HYDICE image (52-63-36 band composite); (b), (d), (f) with legend, corresponding reference maps for (a), (c), (e).

(2) Data 2: Oakland residential, QuickBird image

Data 2 is about a residential area (37.780^o N, -122.154^o W) of Oakland, California, where low density residential housing is surrounded by forest. The pan-sharpened multi-spectral QuickBird satellite image is taken with off-nadir viewing angle 11°. A clip image and reference map of 500x500 pixels are from Liu and Xia (2010). Reference map is generated by manual interpretation aided with high-resolution (0.3m) USGS orthoimage of the same area and is cross-validated by two interpreters. The reference map consists of six classes: tree, grass, vacant land, road, roof, and shadow.

(3) Data 3: DC Mall, HYDICE image

The image covers the Washington DC Mall (38.889^o N, -77.038^o W) and is taken by Hyperspectral Digital Imagery Collection Experiment (HYDICE) sensor (Landgrebe, 2003), with the size of 1280x307 pixels. The original data was collected by Spectral Information Technology Application Center of Virginia and is made available through Purdue University's MultiSpec Project (Landgrebe and Biehl, 2020). Using principal component analysis, the image dimension is reduced to 10 bands which contain 99.9% variance of the original 191 bands. Reference map is generated by manual interpretation aided with high-resolution orthoimage of the same area and is cross validated by two interpreters. The reference classification map consists of seven classes: tree, grass, trail, road, roof, shadow, and water.

For each of the three datasets, two groups of sample pixels are randomly selected, where pixels in group S_1 are used as training set, and the remaining pixels in group S_2 as test set for the classification process. S_1 and S_2 are non-overlapping and independent with each other. In this study, we set $S_1 = S_2 = N = 2000$.

2.2 Three interpolation-based methods

In this section, we briefly introduce three interpolation-based methods to estimate per-pixel classification confidence.

2.2.1 Local error matrix method

Foody (2005) proposes a method to estimate local classification accuracy by combining local error matrix and interpolation. For convenience, we name this method as local error matrix method (LEM). By partitioning test samples based on sub-regions of the area, local classification accuracy is estimated using test data in each sub-region. The local overall accuracies are then interpolated to the whole map to obtain local classification accuracy (LCA) at pixel level. Suppose we have a rectangular study area composed of pixels with a rows and b columns, i.e., $a \times b$ pixels in total. The process to estimate local classification accuracy involves the following steps.

Step 1: Image classification.

Based on S_1 training pixels, the remote sensing image is classified in the usual way to produce a class map. Global error matrix and related accuracy indices, overall accuracy, producer's accuracy, user's accuracy, and kappa, are derived based on the class map and test data. The class map will be further used to create indicator classification confidence map in Section 2.3.1.

Step 2: Definition of grid pixel.

A tessellation of $m \times n$ grid is created and overlaid with the class map. Name the pixel at the location of grid intersection as grid pixel (See Figure 3 for an example with Data 3). In total there are $m \times n$ grid pixels.

Step 3: Selection of neighbor pixels for each grid pixel.

For each grid pixel, a set of *k* neighbor pixels are selected from the test dataset. The neighbor pixels can either be selected using *k*-nearest-neighbor (*k*-NN) method, or selected from a fixed size window, e.g., fixed radius method. Figure 4 shows five example grid pixels and their *k* nearest neighbors.

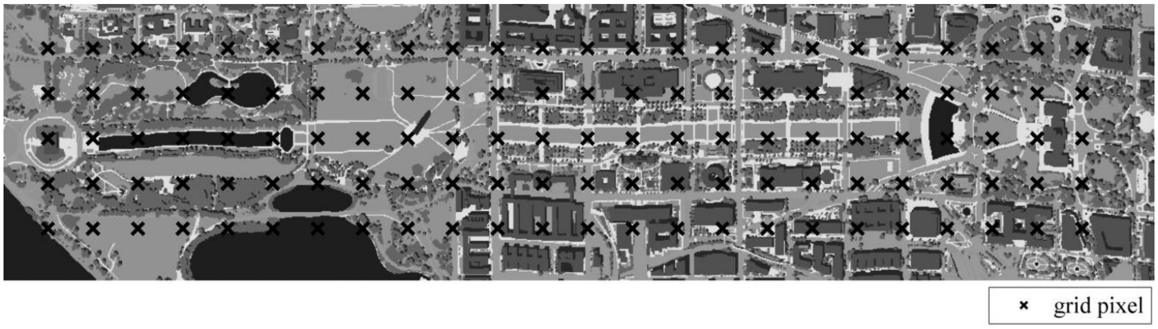


Figure 3 Grid pixels overlaying with reference map

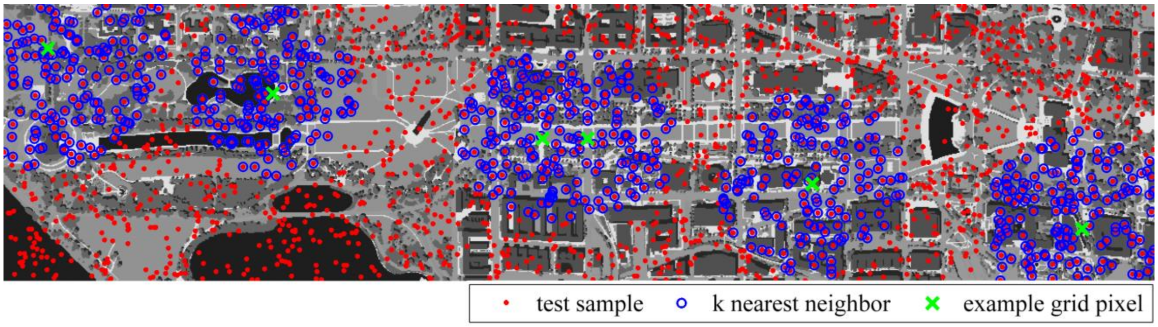


Figure 4 Example grid pixels and *k* nearest neighbors

Step 4: Calculating local classification accuracy for each grid pixel.

For each set of *k* test pixels, construct an error matrix called local error matrix. The overall accuracy derived from each local error matrix is assigned to each grid pixel as local classification accuracy (LCA).

Step 5: Interpolating LCA from grid pixel to the whole map.

The local classification accuracy at grid pixels are interpolated to the whole map using inverse distance weighting (IDW) method as in Foody (2005). The result is what we call map of *estimated* local classification accuracy.

Let u_i = LCA(x_i) be the local classification accuracy of grid pixel *i*, *i* = 1, 2, ..., $m \times n$. The local classification accuracy at pixel **x** is then estimated as,

$$u(\mathbf{x}) = \sum_{i=1}^{N} \frac{w_i(\mathbf{x}) u_i}{\sum_{j=1}^{N} w_j(\mathbf{x})}, \quad (1)$$

where $w_i(\mathbf{x})$ is the weight function,

$$w_i(\mathbf{x}) = \frac{1}{d(\mathbf{x}, \mathbf{x}_i)^p} \quad (2)$$

and,

$\mathbf{x}$ denotes target pixel, whose local classification accuracy unknown,

$\mathbf{x}_i$ are neighbor pixels, whose local classification accuracy are known (estimated using local error matrix),

d is the distance between the known pixel $\mathbf{x}_i$ and unknown pixel **x**,

N is total number of known points used in interpolation, $N \leq m \times n$.

p is the power parameter, a positive true number.

The selection of values for *m*, *n*, and *k* is subjective. For our data, square grid of size 50 pixels are used. Data 1 and 2 are of size 500×500 pixels, therefore *m* = *n* = 9. There are $9\times9=81$ grid pixels which will be used to calculate LCA. Note, pixels on the image boundary are not used as grid pixel. Data 3 is of size 1280×307 pixel, we select *m* = 24, *n* = 5, and there are 24×5 grid pixels. For all three datasets, *k* = 150. In other words, for each grid pixel, 150 closest sample pixels from the set of 2000 test samples are used to construct an error matrix. The overall accuracy is derived for each grid pixel.

2.2.2 Bootstrap method

Steele et al. (1998) combines bootstrap and kriging to estimate per-pixel misclassification rate. Misclassification rate is estimated at test pixels using bootstrap sampling and then interpolated to the whole image. The process is explained as follows.

Step 1: Generating bootstrap training set.

B sets of bootstrap training samples ware generated using the original S_1 training samples. Each bootstrap set has S_1 training samples. Since bootstrap is sampling with replacement, theoretically, the percent of original S_1 training samples that will be included in each bootstrap training set is (Hastie et al., 2009)

$$P\left[(x_i, y_i)\in D\,|\,(x_i, y_i)\in C\right]=1-\left(1-\frac{1}{n}\right)^n\approx1-e^{-1}=0.632.$$

Step 2: Image classification using bootstrap sample sets.

The classifier is trained using one bootstrap sample set to compute classification rule, which is then used to classify the image. This training and classification process is applied to each of the *B* bootstrap sample sets. Each time, we compute a new training rule and then use it to classify the image. In other words, we classify the image *B* times and get *B* versions of class maps.

Step3: Creating bootstrap class map.

B class maps are stacked together and each pixel is labeled *B* times. The mode label for each pixel is recorded and used as the bootstrap class label. This results a bootstrap class map, which will be further used to create indicator classification confidence map in Section 2.3.1.

Step 4: Calculate misclassification rate (MR) for the test data.

Compare each of the *B* class maps with the test data and record each test pixel as correctly or incorrectly classified. Suppose test pixel X_i has been incorrectly classified E_i times, misclassification rate is then estimated as MR $=E_i/B$. Note, in the original work of Steele et al. (1998), the term misclassification probability other than misclassification rate is used. We consider misclassification rate as a better term.

Step 5: Modeling semiveriogram for MR.

For kriging, MR is assumed to be an intrinsically stationary spatial process, i.e., the differences of MR between two pixels separated by a given distance have a constant mean and a constant variance. The experimental semivariogram is estimated based on MR of S_2 test pixels as,

$$\gamma(h) = \frac{1}{2N(h)} \sum_{\alpha=1}^{N(h)} \left[MR(u_\alpha) - MR(u_\alpha + h) \right]^2, \tag{3}$$

where $N(h)$ is the number of pairs of samples with distance h apart from each other. Figure 5 shows the experimental semivariogram for MR of data 1 classified using MLC. The semivariogram increase gradually and tend to be stable at certain lag distance. Mathematical models for fitting experimental semivarigram should satisfy the constraint of positive definiteness. Exponential model of the following form is widely used in the literature (Goovaerts, 1997),

doi:10.20944/preprints202201.0352.v1

$$\gamma(h)=\begin{cases} a+\left(\sigma^2-a\right)\left(1-e^{-3h/r}\right), & \text{for } h>0 \\ 0, & \end{cases}, \tag{4}$$

where a, σ^2, and r represent nugget effect, sill, and range, respectively. These parameters can be estimated using nonlinear weighted least squares method (Cressie, 1985).

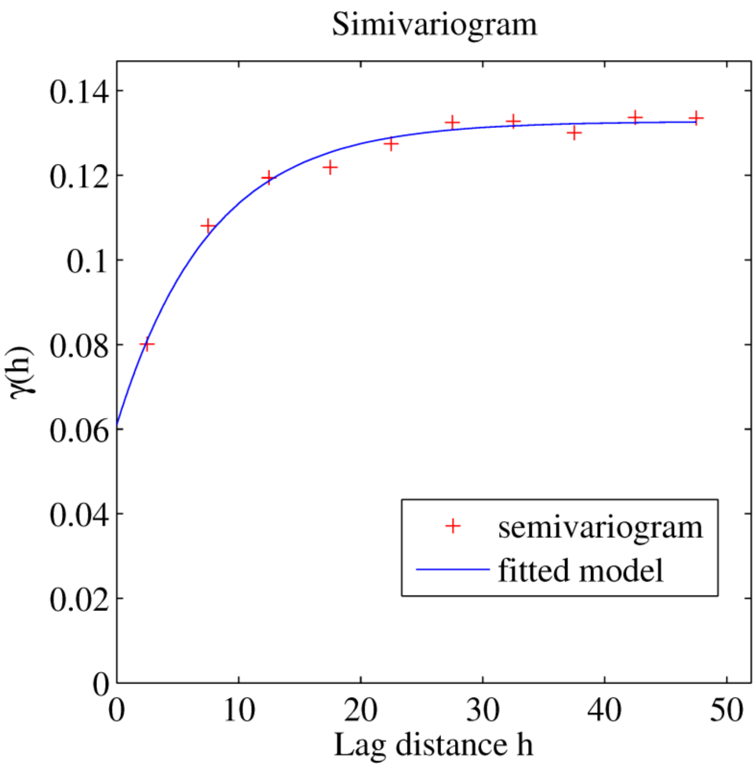


Figure 5 Experimental semivariogram and fitted model for Data 1 classified using MLC

Step 6: Predict MR for the whole map using ordinary kriging.

Under intrinsic stationary assumption for MR, we use ordinary kriging to estimate MR for all the pixels in the class map. The predicted MR for pixel t is a weighted average of the MR for n sample pixels:

$$MR_t = \sum_{i=1}^{n} w_i MR_i \tag{5}$$

where w_i is the ordinary kriging weight for the *i*th sample pixel. The weights are estimated using the semivariogram through solving of the following optimization problem:

$$\begin{gathered}\sum_{i=1}^{n} w_i(u) = 1 \\ \min\left\{\mathbf{w}^T(u)\mathbf{C}\mathbf{w}(u) + \sigma^2 - 2\mathbf{w}^T(u)\mathbf{c}(u)\right\}, \text{ subject to } \mathbf{w}^T(u)\times\mathbf{1} = 1\end{gathered} \tag{6}$$

where **C** is the covariance matrix of MR, **c** is the covariance between know pixel and target pixel, σ^2 is the variance of the MR. The ordinary kriging of weight and its variance is then estimated as

$$\begin{gathered}\mathbf{w} = \mathbf{C}^{-1}\mathbf{c} \\ \sigma_{OK}^2 = \sigma^2 - \mathbf{c}^T(u)\mathbf{C}^{-1}\mathbf{c}(u)\end{gathered} \tag{7}$$

2.2.3 Geostatistical method

Different from previous two interpolation based methods, Kyriakidis and Dungan (2001) propose a geostatistical method to map per-pixel classification confidence. Their central idea is to combine error matrix with kriging to predict per-pixel classification confidence. The geostatistical method includes the following steps.

Step 1: Image classification.

Classify image in the usual way using S_1 training data, same as Step 1 for local error matrix method. Construct the regular (global) error matrix using S_2 test data. For the convenience of discussion bellow, a typical error matrix is shown in Table 2. The class map will also be used to create indicator classification confidence map in Section 2.3.1.

Step 2: Define indicator variable of class labels for each pixel.

doi:10.20944/preprints202201.0352.v1

Let the class label of pixel *u* be a random variable *s*(*u*). Based on the indicator framework of Journel (1986), class label for each pixel can be coded as a set of *K* local probabilities, each associated with the *k*th class *k*:

$$Pr\{s(u) \ = k \,|\, \text{info}(u)\} s, k = \ 1,\ 2,\ ..., K, \tag{8}$$

which represents the probability of class *k* observed at location *u* on the ground given the classification results info(*u*).

Equation (8) can be further specified based on the different information used (Goovaerts, 1997). For image classification, the following two scenarios are used.

(1) For the reference map, equation (8) turns out as follows,

$$I(u;k) = \begin{cases} 1, \text{ if } s(u) = k \\ 0, \text{ if not} \end{cases} \quad k = 1, 2, ..., K. \tag{9}$$

(2) For classification map, equation (8) is expressed as,

$$y(u;k) \ = \ Pr\{s(u) \ = k \,|\, x(u) \ = k'\} \ = p\left(k \,|\, k'\right), k = \ 1,\ 2,\ ..., K, \tag{10}$$

where *p*(*k* | *k*') is the proportion of a pixel of class *k* on the reference map, given that it is classified as class *k*', and *x*(*u*), the class label for pixel *u* on the class map. *p*(*k* | *k*') can be estimated as:

$$p(k \,|\, k') = \frac{\sum_{\alpha=1}^{n(u)} I(u_\alpha;k) J(u_\alpha;k')}{\sum_{\alpha=1}^{n(u)} J(u_\alpha;k')}, \quad k, k' = 1, 2, ..., K, \tag{11}$$

where $I(u_\alpha;k)$ is the indicator for reference map *s*(*u*; *k*) as in equation (9), $J(u_\alpha;k')$, the indicator of classification map *x*(*u*; *k*') defined as

$$J(u;k') = \begin{cases} 1, & \text{if } x(u) = k' \\ 0, & \text{if not} \end{cases} \qquad k' = 1, 2, \ldots, K. \tag{12}$$

In practice, $p(k\,|\,k')$ is estimated using the error matrix as in Table 2, i.e., divide $x_{k'k}$ by corresponding row total $x_{k'+}$ and obtain

$$p(k\,|\,k') = \frac{x_{k'k}}{x_{k'+}} = \frac{x_{k'k}}{\sum_{k=1}^{K} x_{k'k}}, \tag{13}$$

Specifically, when $k' = k$, $p(k|k')$ is the user's accuracy.

Table 2 A typical error matrix

			Reference data					Row total
	Class	1	2	…	k	…	K	
Class map	1	x_{11}	x_{12}	…	x_{1k}	…	x_{1K}	x_{1+}
	2	x_{21}	x_{22}	…	x_{2k}	…	x_{2K}	x_{2+}
	…	…	…	…	…	…	…	…
	k'	$x_{k'1}$	$x_{k'2}$	…	$x_{k'k}$	…	$x_{k'K}$	$x_{k'+}$
	…	…	…	…	…	…	…	…
	K	x_{K1}	x_{K2}	…	x_{Kk}	…	x_{KK}	x_{K+}
Column total		x_{+1}	x_{+2}	…	x_{+k}	…	x_{+K}	N

Step 3: Define residual for class probability.

The conditional probability in equations (11) and (13) can be viewed as the mean of the indicator RV $I(u;k)$ at location u:

$$p(k\,|\,k') \;= E\left\{I(u;k)\right\}, k = \;1,\;2,\;...,K. \tag{14}$$

In other words, the conditional probability obtained from the error matrix is the "average" spatial variability of land classes on the ground. Therefore, the residual for class probabilities can be defined as follows,

$$r(u_\alpha;k) = I(u_\alpha;k) - y(u_\alpha;k). \tag{15}$$

Step 4: Model the residuals using empirical semivariogram similar to step 5 in Section 2.2.2. As usual, exponential model is used to fit the semivariogram.

Step 5: Predict probabilities of class k allocated to pixel u using simple indicator kriging:

$$p^*_{sIK}(u;k) = y(u;k) + \sum_{\alpha=1}^{n(u)} w(u_\alpha;k)\left[I(u_\alpha;k) - y(u_\alpha;k)\right], \tag{16}$$

where the weights $w(u_\alpha;k)$ are determined by solving the simple kriging system:

$$\min\left\{\mathbf{w}^T(u;k)\mathbf{C}\mathbf{w}(u;k) + \sigma(k)^2 - 2\mathbf{w}^T(u;k)\mathbf{c}(u;k)\right\} \tag{17}$$

where $\mathbf{C}$ is the covariance matrix of residual $r(u;k)$, $\mathbf{c}$ is the covariance between know pixel and target pixel, σ^2 is the variance of the $r(u;k)$. The simple kriging of weight and its variance is then estimated as

$$\begin{aligned} \mathbf{w}(u;k) &= \mathbf{C}^{-1}\mathbf{c}(u;k) \\ \sigma(k)^2_{OK} &= \sigma(k)^2 - \mathbf{c}^T(u;k)\mathbf{C}^{-1}\mathbf{c}(u;k) \end{aligned} \tag{18}$$

Step 6: Construct per pixel confusion index based on $p(u;k)$.

According to suggestions of Kyriakidis and Dungan (2001), the following local index of map quality, called per pixel confusion index (CI), is constructed,

$$c(u) = \left[1 - p^m(u)\right]\left(\frac{K}{K-1}\right), \tag{19}$$

where $p^m(u) = \max\{p(u;k), k = 1, 2, ..., K\}$, and $K/(K\text{-}1)$ is a standardization factor. When a pixel is classified with without uncertainty, i.e., $p^m(u) = 1$, $c(u) = 0$. When the probability of a pixel belonging to each class is equal, the lowest classification confidence is arrived. In such case, $p^m(u) = p(u;k) = 1/K$, $c(u) = 1$.

2.2.4 Clarification of concepts

It is necessary to clarify four concepts used in this study: local classification accuracy (LCA), misclassification rate (MR), confusion index (CI), and per-pixel classification confidence.

- Local classification accuracy (LCA) is used by Foody (2005), referring to the overall accuracy derived from local error matrix. The higher LCA is, the more accurate the classification is.
- Misclassification rate (MR) is used by Steele et al. (1998) to measure the number of misclassification divided by the total number of bootstrap classification. The lower MR is, the less uncertain the classification is.
- Confusion index (CI) is defined by Kyriakidis and Dungan (2001) to characterize the confusion of class assignment in image classification. The lower CI is, the less confusion of class assignment, i.e., the less uncertain the classification is.
- Per-pixel classification confidence is defined as the probability of a pixel being correctly classified.

2.3 Estimated classification confidence vs. true classification confidence

To evaluate the performance of three methods, we compare the map of *estimated classification confidence* with a map of *true classification confidence*. Using each of the three methods in Section 2.2.1-2.2.3, we get a map of estimated per-pixel classification confidence, which takes the form of continuous values ranging between 0-1. For convenience, we call the estimated classification confidence map as continuous map of estimated classification confidence. Local error matrix method and bootstrap method have similar procedures to estimate per-pixel classification confidence: (1) Generate classification confidence at sample pixels; (2) Interpolate estimation at sample pixels to the whole map. Different from these two methods, the geostatistical method generates a set of class probabilities for each pixel belonging to each class. These class probabilities are then used to create a local index of classification confidence.

Due to the difference in three methods, we design different approaches to obtain the map of true classification confidence.

First, for each of the three methods, a class map is obtained through the methods in 2.2.1-2.2.3. The class map is then compared with the full coverage reference map to create an indicator map where each pixel is indicated as either correctly classified or incorrectly classified.

Second, for local error matrix method and bootstrap method, we design another type of true classification map in continuous form. We create this continuous map of classification confidence by extending the original method to all pixels without interpolation. Procedures for this second type of true classification confidence map are different for local error matrix method and bootstrap method. This continuous true classification confidence map is used to examine the interpolation effect of both methods.

For local error matrix method, we obtain a continuous map of true classification confidence by extending the local error matrix from sample pixels to all the pixels. In other words, we construct local error matrix as Step 3 and 4 in Section 2.2.1 for each pixel, not just for the grid pixel. Our full coverage reference data make this possible. The overall accuracy estimated from each local error matrix is regarded as the true classification confidence for each pixel.

For bootstrap method, similar to Step 4 in Section 2.2.2, we create a map of true MR by comparing each bootstrap class map with the full coverage map and calculate MR for each pixel, not just for the test sample pixels.

It should be noted that due to the special procedures of geostatistical method, there is no way to obtain a meaningful map of true classification confidence in continuous form.

2.4 Approaches to evaluating method performance

2.4.1 Evaluation scheme

Table 3 shows the evaluation scheme for this study.

Table 3 Evaluation scheme

	Exploratory data analysis	**Statistical analysis**
Estimated classification confidence for correct and error pixels	LEM, BM, GM	LEM, BM, GM
Continuous maps of estimated vs. true classification confidence	LEM, BM	LEM, BM

Note: LEM - local error matrix method, BM - bootstrap method, GM - geostatistical method.

First, we evaluate two relationships between the classification confidence map derived from each of the tree methods. For all three methods, we divide the pixels in each dataset into

doi:10.20944/preprints202201.0352.v1

two groups: correct pixels and error pixels. We then examine the relationship between the estimated classification confidence for correct pixels and error pixels. It is assumed that correct pixels tend to have high classification confidence, while error pixels have low classification confidence. In other words, the classification confidence should be distinguishable for correct pixels and error pixels. For local error matrix method and bootstrap method, we will also examine the relationship between the estimated continuous classification confidence and true continuous classification confidence. If an interpolation-based method is effective in predicating classification confidence, the estimated classification confidence should agree well with the true classification confidence.

Second, we evaluate the three methods in two ways: exploratory data analysis (EDA) and statistical analysis. For exploratory data analysis, we use one EDA tool, bi-histogram. For statistical analysis, we construct two statistics, Willmott's *d* and Binned classification quality (BCQ). Willmott's *d* measures the similarity between the estimated continuous classification confidence map and the true continuous classification confidence map (Willmott, 1982, 1981; Willmott et al., 2012). BCQ is calculated based on binned pixels and thus examines the relationship between LAC/MR/CI and proportion of correct pixels in each bin.

2.4.2 Bi-histogram

The bi-histogram is a graphical tool for examining the distribution of two datasets by the histograms of both datasets. In this paper, we use bi-histogram to explore the two types of relationships introduced in Table 3. We will plot two bi-histograms: (1) bi-histogram of the estimated classification confidence of correct pixels vs. error pixels; (2) bi-histogram of the estimated vs. true continuous classification confidence (for local error matrix method and bootstrap method only).

2.4.3 Measure of agreement

As discussed in Section 2.3, for local error matrix method and bootstrap method, continuous maps of true and estimated classification confidence are obtained. In this section, the agreement of two maps is evaluated. If both maps agree well with each other, the method is effective in estimating per-pixel classification confidence.

There are several widely used measures of agreement/disagreement of two datasets (Ji and Gallo, 2006), including Pearson correlation coefficient (*r*), coefficient of determination (r^2), mean absolute error (MAE), root mean square error (RMSE), Willmott's index of agreement (*d*), Mielke's measure of agreement (ρ), Robinson's coefficient of agreement (*A*), and Ji and Gallo's agreement coefficient (*AC*). A detailed review of these measures is referred to Ji and Gallo (2006). In this paper, we use Willmott's *d* to measure the agreement between the continuous map of estimated classification confidence and true classification confidence. Willmott's *d* has two main advantages that serve our purpose: (1) Bounded in a range from 0 to 1, indicating degree of agreement ranging from complete disagreement to complete agreement. (2) Non-dimensional, thus easier to interpret than the widely used RMSE.

Similar to most dimensionless measures of agreement, Willmott's *d* is designed in the following form (Willmott, 1982, 1981; Willmott et al., 2012)

$$\rho = 1 - \frac{\delta}{\mu}, \tag{20}$$

where δ is a dimensioned average error-magnitude, μ the potential error, i.e., the basis of comparison. Willmott's *d* is expressed as

$$\begin{aligned} d &= 1 - \frac{SSE}{PE} \\ &= 1 - \frac{\sum_{i=1}^{N}\left[\left(X_i - \bar{X}\right) - \left(Y_i - \bar{X}\right)\right]^2}{\sum_{i=1}^{N}\left[\left|X_i - \bar{X}\right| - \left|Y_i - \bar{X}\right|\right]^2} \\ &= 1 - \frac{\sum_{i=1}^{N}\left[\left(X_i - Y_i\right)\right]^2}{\sum_{i=1}^{N}\left[\left|X_i - \bar{X}\right| - \left|Y_i - \bar{X}\right|\right]^2} \end{aligned} \tag{21}$$

where X_i denotes true classification confidence, $\bar{X}$ is the mean of X_i, Y_i is the predicated classification confidence, N is the number of pixels in the map. Willmott's d is a measure based on the sum of squares, δ is the sum of the squared errors, and μ is the overall sum of the squares for absolute values of two partial differences from the true mean, $\left|X_i - \bar{X}\right|$ and $\left|Y_i - \bar{X}\right|$. According to equation (21), the lower limit of d is zero, indicating complete disagreement, and the upper limit of d is one, indicating complete agreement.

2.4.4 Binned classification quality

Using the three methods introduced in Section 2.2.1-2.2.3, three measures of per-pixel classification confidence are obtained: LCA, MR, and CI. Binned classification quality (BCQ) is calculated using LCA, MR, and CI.

After calculation of BCQ for LCA, MR, and CI, the following hypothesises are examined:

- For LCA, BCQ$_q$ has positive relationship with b_q.
- For MR, BCQ$_q$ has negative relationship with b_q.
- For CI, BCQ$_q$ has negative relationship with b_q.

2.5 Summary of methodology

Here is a summary of the methodology for this paper.

Step 1: Image classification. The three datasets are classified using four commonly used classifiers: MLC, NN, SVM, and BDT. For local error matrix method and geostatistical method, each involves $3 \times 4 = 12$ scenarios of image classification and thus results 12 versions of class maps. For bootstrap method, there are $12 \times B$ versions of class maps.

Step 2: Creating classification confidence map. For local error matrix method and bootstrap method, two types of classification confidence map are obtained: continuous map and indicator map. For geostatistical method, only the indicator classification confidence map is generated.

Step 3: Calculating two statistics: Willmott's *d* and BCQ

- Willmott's *d* is calculated by comparing estimated LCA/MR with the true LCA/MR respectively.
- BCQ is calculated by grouping pixels based on each of the three measures, i.e., LCA, MR, and CI.

Step 4: Evaluating three methods using EDA tool (bi-histogram) and two statistics (BCQ and Willmott's *d*).

The results from three methods are also compared to examine their difference in predicting per-pixel classification confidence.

3 Results

3.1 Local error matrix method

3.1.1 Maps of classification confidence

Figure 6 shows the maps of estimated LCA based on maximum likelihood classification. Each panel is created by interpolating the overall accuracy from the local error matrix for each grid pixel. There are some spatial patterns in the estimated LCA map. For example, Data 1 has

doi:10.20944/preprints202201.0352.v1

low LCA in west and southwest side, and a bump of high LCA stretching from southwest to northeast. According to the reference map of Data 1, those areas with low LCA are highly heterogeneous patches, while areas with high LCA are homogeneous agricultural land. Data 2 has a valley of low LCA stretching from southwest to northeast, a hotspot of high LCA in southeast corner, and a high patch in the north side. Areas with low LCA are mainly composed of residential building and concrete surfaces, and these two classes are easily to be confused. Areas with high LCA are mainly forest areas and are more homogeneous. Data 3 has low LCA stretching from mid-north to mid-south and southeast, and high LCA in mid-west and northeast. Similar to Data 2, areas of low LCA in Data 3 is due to the confusion between roof and concrete surfaces. Areas of high LCA are vegetated area which are of less confusion.

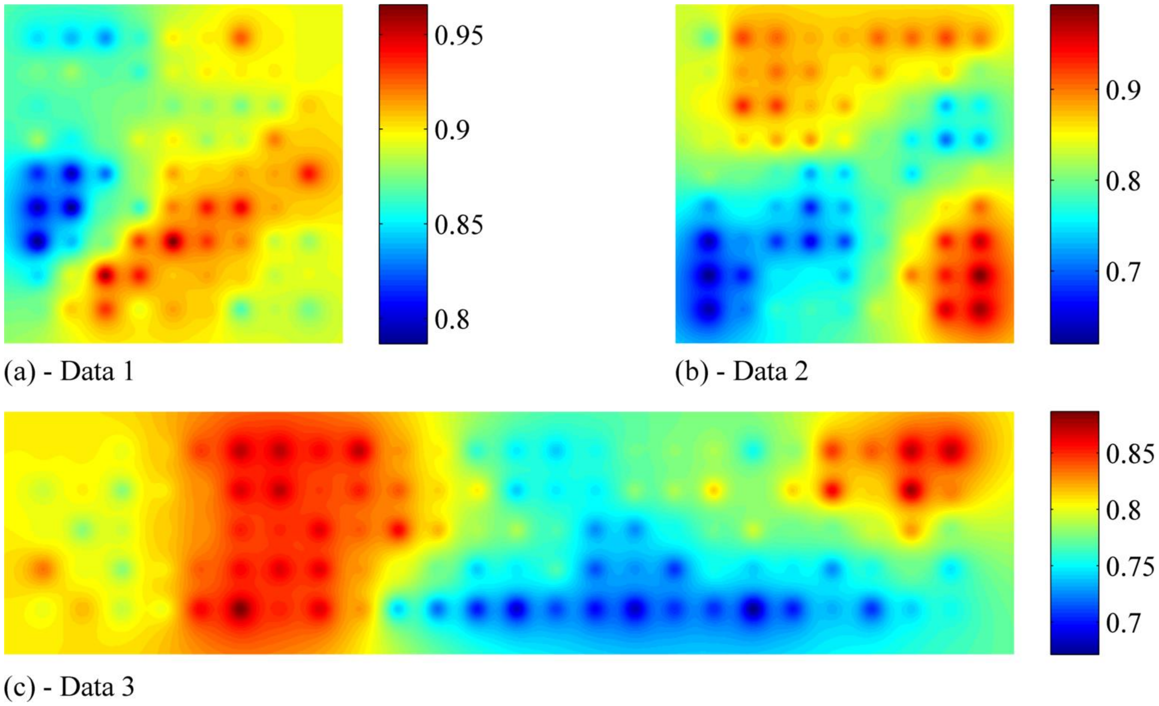
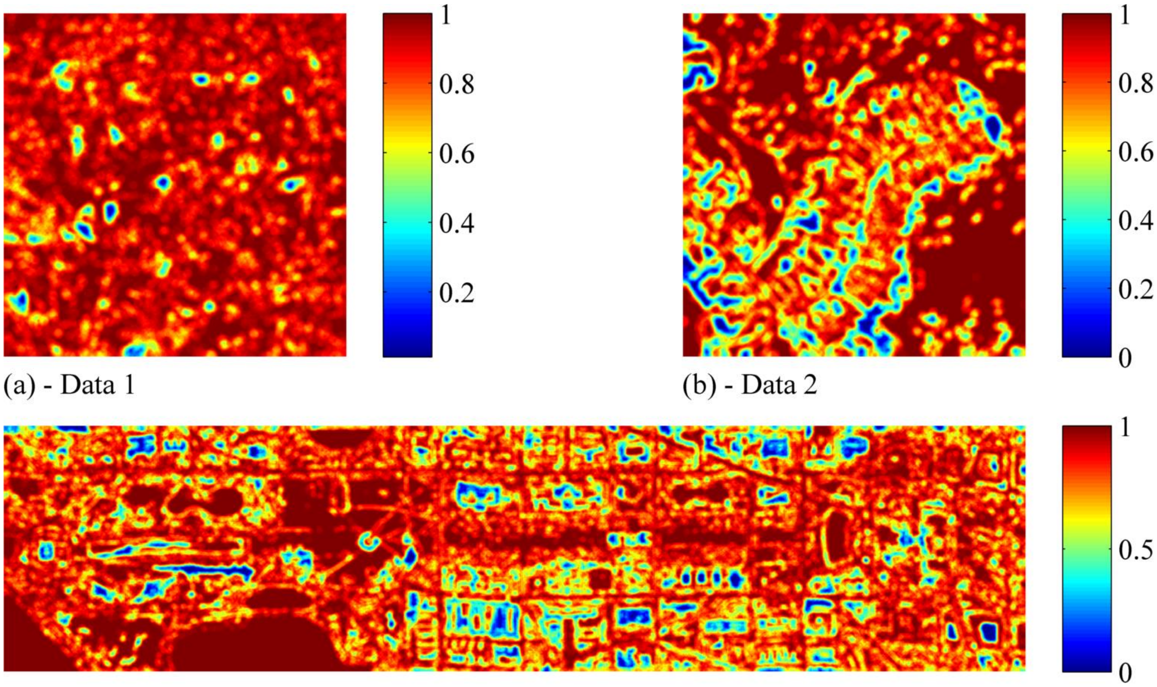


Figure 6 Map of estimated LCA from IDW interpolation (MLC)

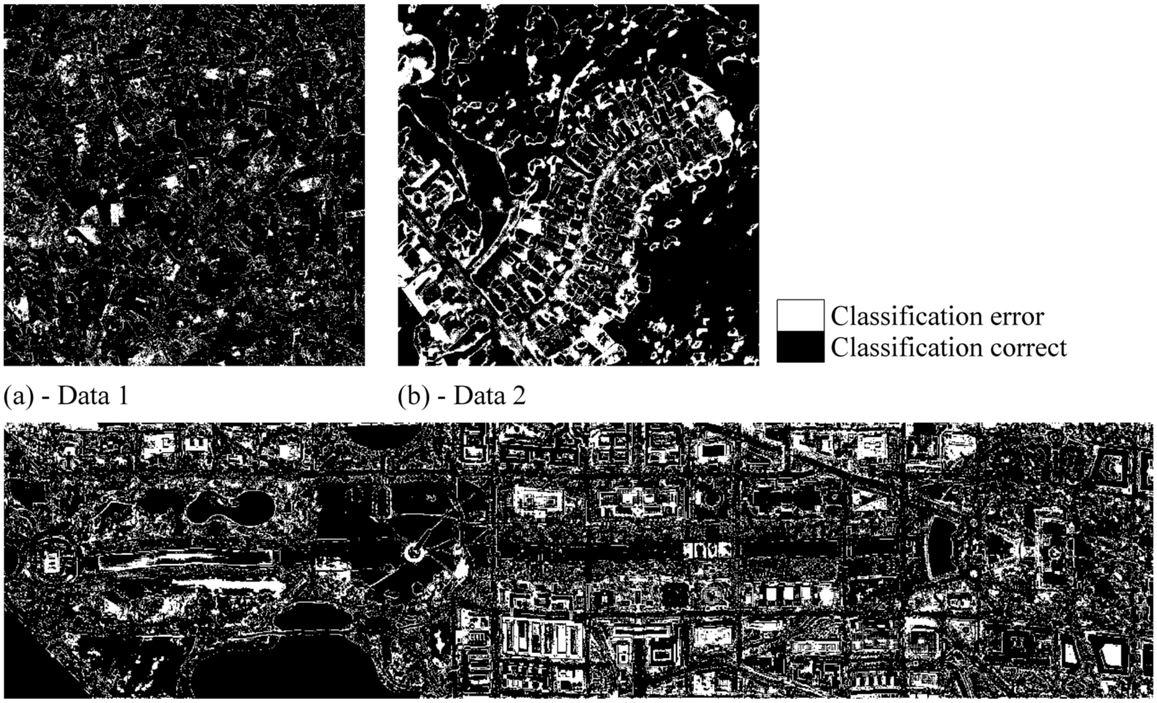


(a) - Data 1

(b) - Data 2

(c) - Data 3

Figure 7 Maps of true LCA (MLC)



(a) - Data 1

(b) - Data 2

(c) - Data 3

Figure 8 Indicator maps of classification error (MLC)

Figure 7 shows the true LCA based on maximum likelihood classification which is the overall accuracy from the local error matrix for each pixel. shows the indicator maps of classification error/correct which are generated by comparing the class map with the full coverage reference map. As expected, plots in Figure 7 are smoother than the plots in Figure 8, while plots in Figure 6 are much smoother than those in Figure 7. The visual difference of three figures is quite evident.

3.1.2 Estimated LCA of correct and error pixels

Figure 9 shows the bi-histogram of estimated LCA for correct pixels and error pixels. The shapes of corresponding histograms for both types of pixels look quite similar. Most pixels, either correctly classified or incorrectly classified, have LCA concentrated in certain range. For Data 2, correct pixels seem to be a little more concentrated in high LCA than error pixels.

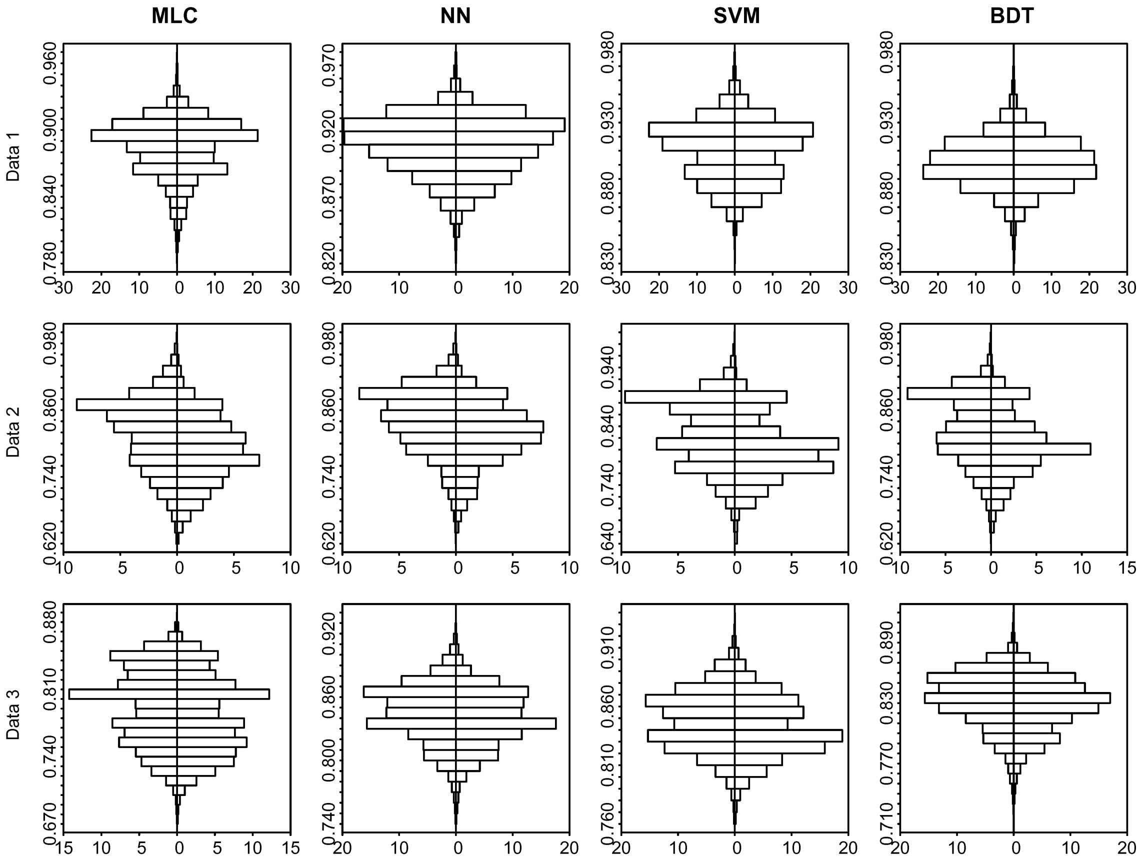


Figure 9 Bi-histogram of estimated LCA for correct and error pixels

Note: y - LCA; x: proportion of pixels. Left side - correct pixel, right side - error pixel.

3.1.3 Comparing the continuous map of estimated and true LCA

The visual difference between Figure 6, Figure 7 is evident. First, the distribution of LCA is different. Figure 10 gives the bi-histograms of true vs. estimated LCA for all three data with four classifiers. The true LCA is more spread out than the estimated LCA.

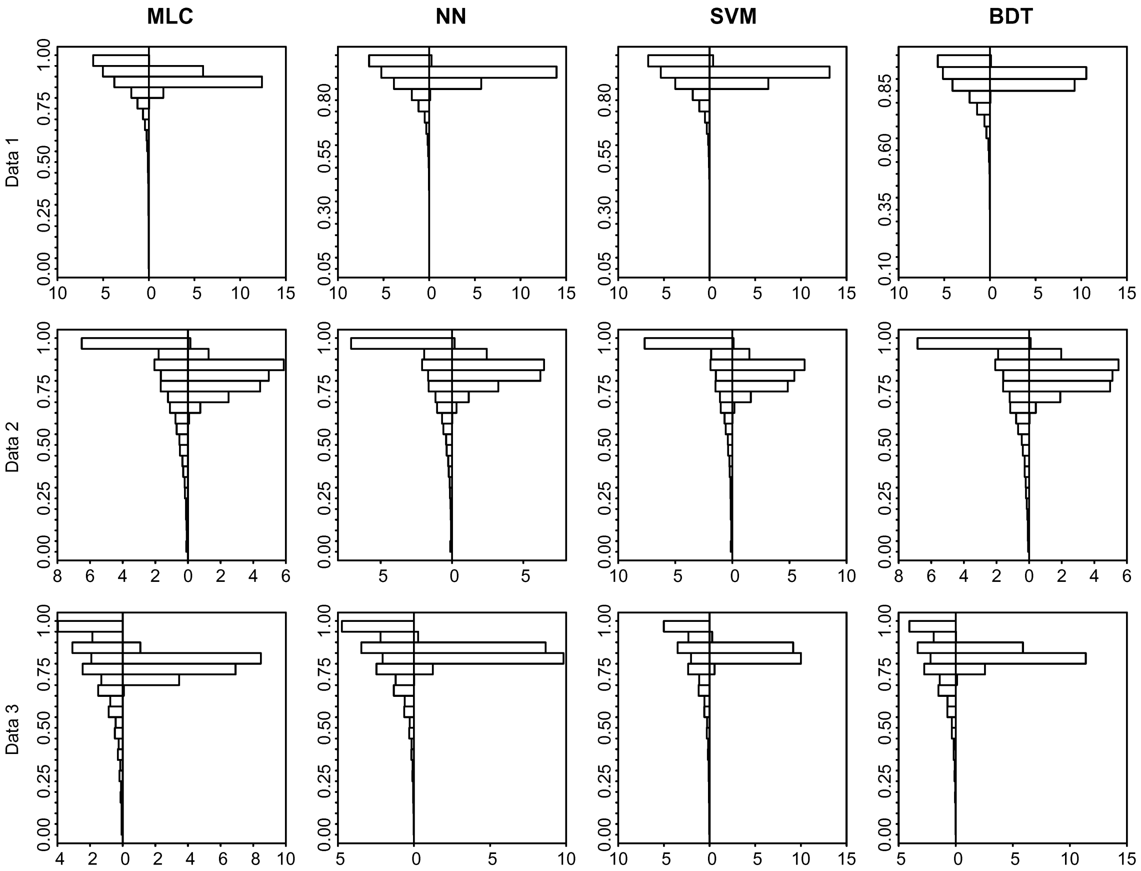


Figure 10 Bi-histogram of true and estimated LCA

Note: y: LCA, x: proportion of pixels. Left - true LCA, right - estimated LCA.

Table 4 Willmott's *d* of estimated and true LCA

	MLC	NN	SVM	BDT
Data 1	0.224	0.262	0.249	0.220
Data 2	0.449	0.401	0.393	0.430
Data 3	0.291	0.260	0.242	0.283

In summary, local error matrix method produces poor results for classification confidence. This is due to the internal weakness of local error matrix method, i.e., the use of interpolation.

3.1.4 Relationship between binned classification quality and LCA

As introduced in Section 2.4.4, the hypothesis to test is: If a pixel has high LCA, the probability of this pixel being a correct pixel is high. In other words, there should be a positive relationship between BCQ_q and b_q. Figure 11 shows scatter plot of binned classification quality (BCQ_q) against LCA (b_q) where LCA is divided into 30 bins. The patterns of the scatter plots vary with data and classifiers. For Data 1, there seems no clear relationship between BCQ_q and b_q. The scatter plots of Data 2 show linear positive relationship between BCQ_q and b_q although there are some outliers when LCA is small. The scatter plots of Data 3 show some linear positive relationship between BCQ_q and b_q with MLC. As to Data 3 with NN, SVM, BDT, the scatter plots show a "V" shape pattern. Besides dividing LCA into Q = 30 bins, we also tried other values of Q = 40, 50, ..., 100. The general pattern of the scatter plot is not sensitive to the number of bins, Q.

Table 5 shows the correlation coefficient R between BCQ_q and b_q. The low values of R indicate weak correlation between BCQ_q and b_q. It can be concluded that LCA is not a good measure for characterizing per-pixel classification confidence.

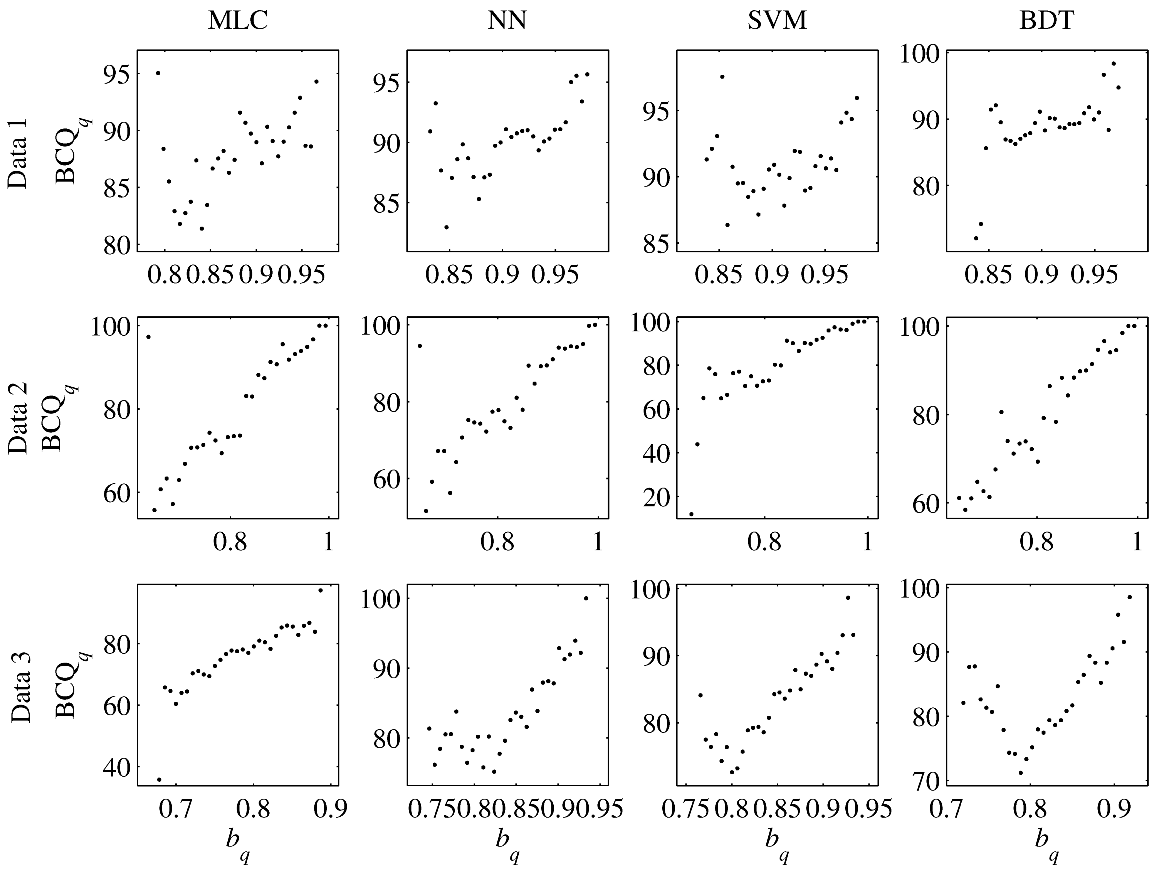


Figure 11 Scatter plot of BCQ$_q$ vs. b_q based on LCA for all the pixels

Table 5 Correlation coefficients (R) of BCQ$_q$ vs. b_q for LCA (Q=30)

	MLC	**NN**	**SVM**	**BDT**
Data 1	0.5349	0.6569	0.2932	0.6547
Data 2	0.8291	0.8309	0.8362	0.9634
Data 3	0.8821	0.8350	0.8861	0.5532

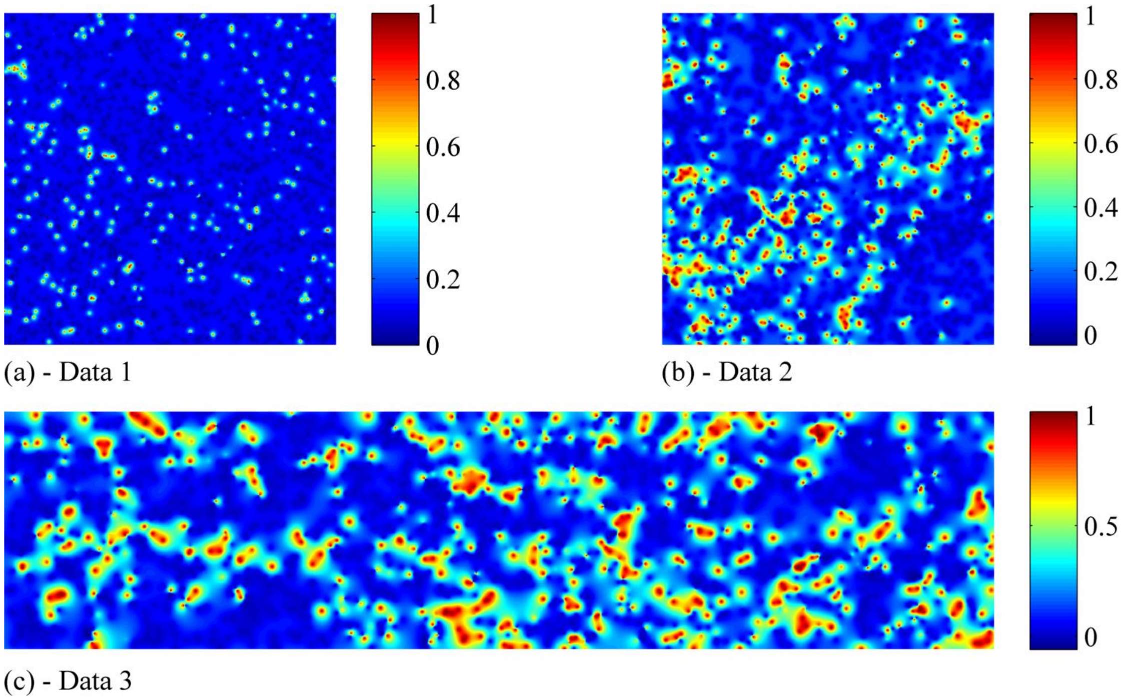
doi:10.20944/preprints202201.0352.v1

3.2 Bootstrap method

3.2.1 Maps of classification confidence

Figure 12 shows the maps of estimated MR based on maximum likelihood classification. Each panel is created by interpolating the MR from S_2 test sample pixels using ordinary kriging. There are some spatial clusters in the MR map. The clusters in Data 1 seem to spread randomly across the whole map, while the clusters for Data 2 and 3 have certain spatial pattern: high MR clusters are mainly located in the roof and concrete surface area. This confirms previous findings in Section 3.1 that classifiers have higher error in distinguishing roof and concrete surfaces.

Figure 13 shows the true MR based on bootstrap method. MR in each panel is obtained by comparing each bootstrap class map with the full coverage reference map and record each pixel as correctly or incorrectly classified. Same as MR for test pixels, MR for each pixel is the number of misclassification times divided by the number of bootstrap iterations, i.e., *B*. Figure 14 shows the indicator maps of classification error generated by comparing the bootstrap class map with the full coverage reference map. Figure 12 is visually different from both Figure 13 and Figure 14 in that the interpolation smoothed out true classification confidence. Therefore, bootstrap method may not be a good method for per-pixel classification confidence. Again, this is due to the internal weakness of using interpolation to estimate classification confidence.

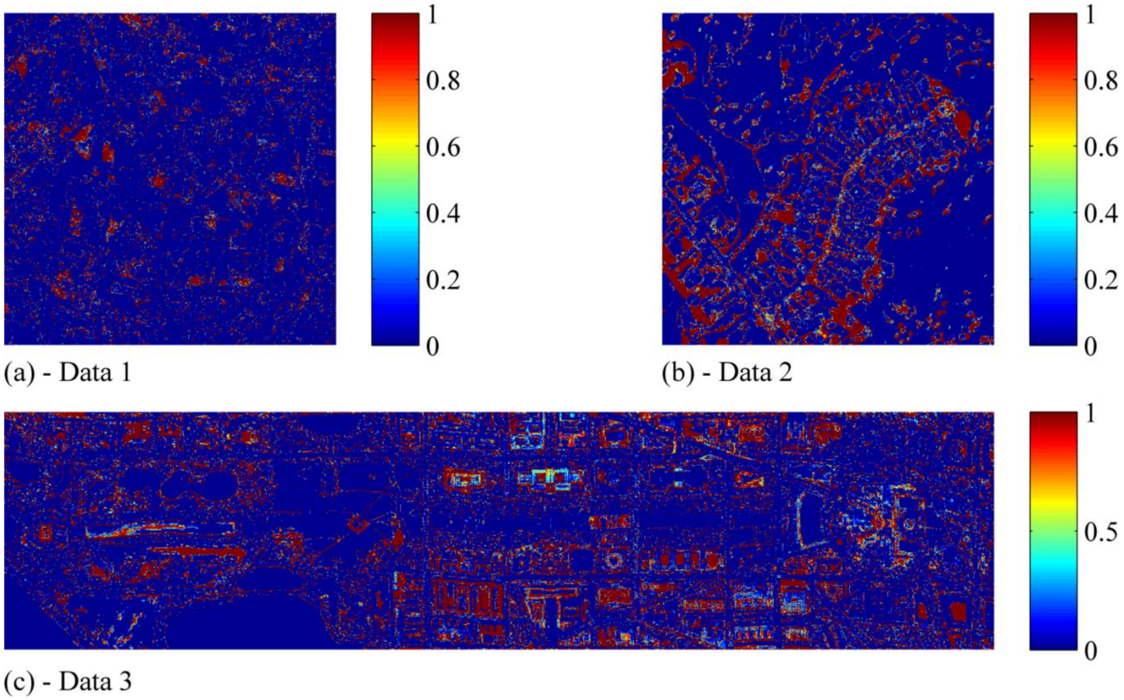


(a) - Data 1

(b) - Data 2

(c) - Data 3

Figure 12 Maps of estimated MR from bootstrap method (MLC)



(a) - Data 1

(b) - Data 2

(c) - Data 3

Figure 13 Maps of true MR from bootstrap method (MLC)

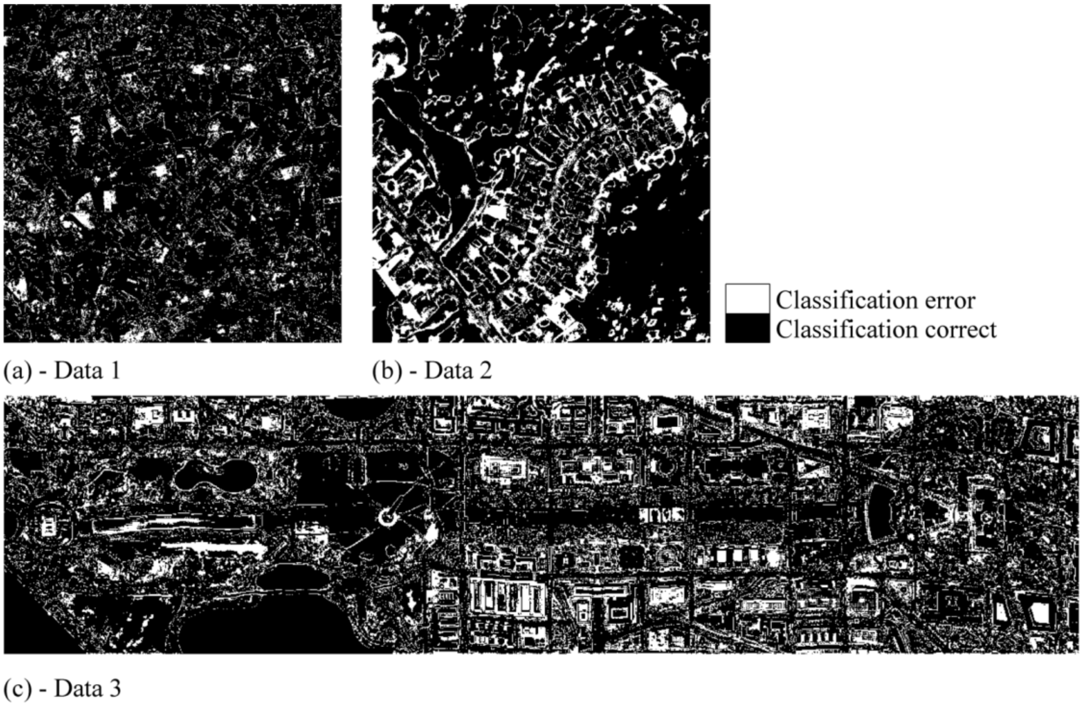


Figure 14 Indicator maps of true classification error from bootstrap method (MLC)

3.2.2 Estimated MR of correct and error pixels

Same as the evaluation of local error matrix method, the bi-histogram of MR for correct pixels and error pixels are plotted in Figure 15. MR for correct and error pixels have similar distribution. The visual shapes of corresponding histograms for correct pixels and error pixels look quite similar except for Data 2. Most pixels, either correctly classified or incorrectly classified, have MR concentrated below 0.3. The similar distribution of correct pixels and error pixels indicate that MR cannot be used to distinguish both types of pixels. It may be concluded that MR is not a good measure for per-pixel classification confidence.

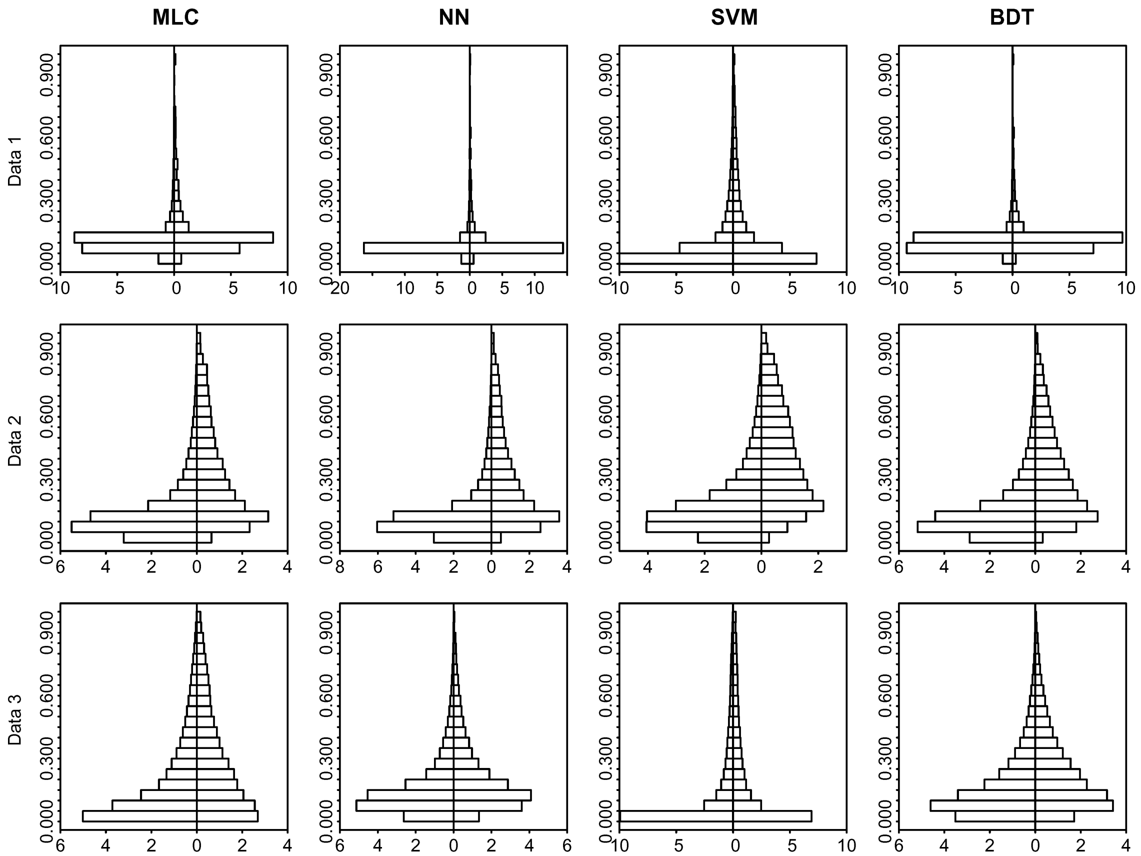


Figure 15 Bi-histogram of MR for correct and error pixels

Note: y - MR; x: proportion of pixels. Left side - correct pixel, right side - error pixel.

3.2.3 Comparing the continuous map of estimated and true MR

Figure 16 gives the bi-histogram of estimated and true MR for all three data with four classifiers. The difference of true and estimated MR is clear. True MR are mainly concentrated in low values, while the estimated MR are more spread out, especially for Data 2 and Data 3.

Table 6 shows the Willmott's *d* calculated from the true and estimated MR, indicating low agreement between estimated MR and true MR. The result from Willmott's *d* confirms the findings in Figure 16.

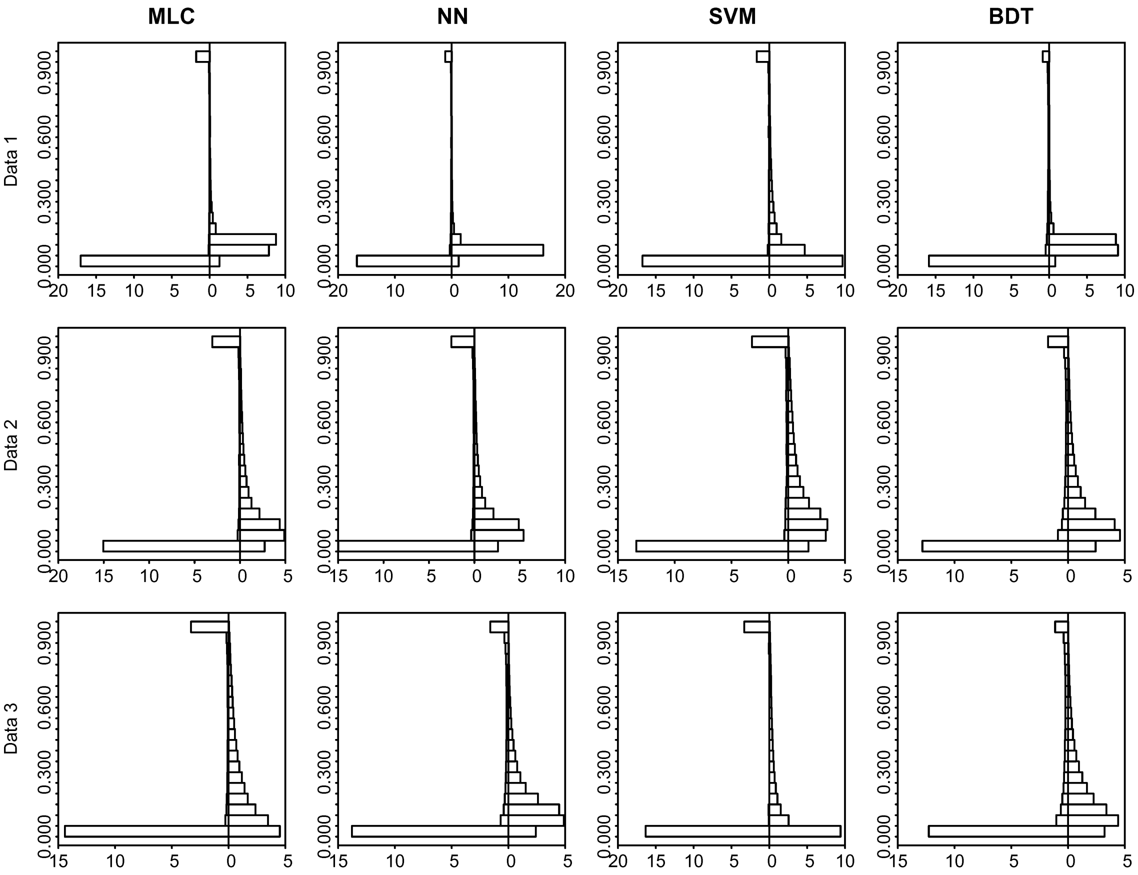


Figure 16 Bi-histogram of true and estimated MR

Note: y: MR, x: proportion of pixels. Left - true values, right - estimated values.

Table 6 Willmott's *d* for estimated and true MR

	MLC	**NN**	**SVM**	**BDT**
Data 1	0.2656	0.2023	0.3852	0.2047
Data 2	0.5517	0.5444	0.6162	0.5858
Data 3	0.4729	0.4072	0.4393	0.4480

In summary, bootstrap method is not efficient for estimating classification confidence. The key issue is due to the interpolation effect.

3.2.4 Relationship between binned classification quality and MR

Figure 17 shows the scatter plots of binned classification quality (BCQ$_q$) against MR (b_q) where MR is divided into 30 bins with equal distance. There is negative relationship between BCQ$_q$ and b_q. The higher MR is, the higher probability that a pixel is an error pixel. **Error! Reference source not found.** also shows that the exact relationship between different datasets and classifiers are different. For Data 1 with MLC, NN, and BDT, the curves are steep when MR< 0.8, 0.7, 0.6 respectively, then they level off. For Data 1 with SVM, Data 2 with all classifiers, the curves look much similar: BCQ decreases slowly when MR<0.8, then BCQ drops quickly to zero. For Data 3, the curves of the scatter plots decrease slowly when MR is smaller than 0.9, and then drop sharply when MR is greater than 0.9.

It should be noted that there are some level off points in the scatter plots for Data 1 with MLC, NN, BDT. This indicates that pixels with MR greater than certain cutoff values are all misclassified. The cutoff value for Data 1 with MLC, NN, and BDT are 0.9, 0.7667, and 0.7333.

Table 7 shows the correlation coefficient R between BCQ$_q$ and b_q. The absolute values of R are not high which indicate the correlation between BCQ$_q$ and b_q is not strong. It can be concluded that LCA is not a good measure for characterizing per-pixel classification confidence.

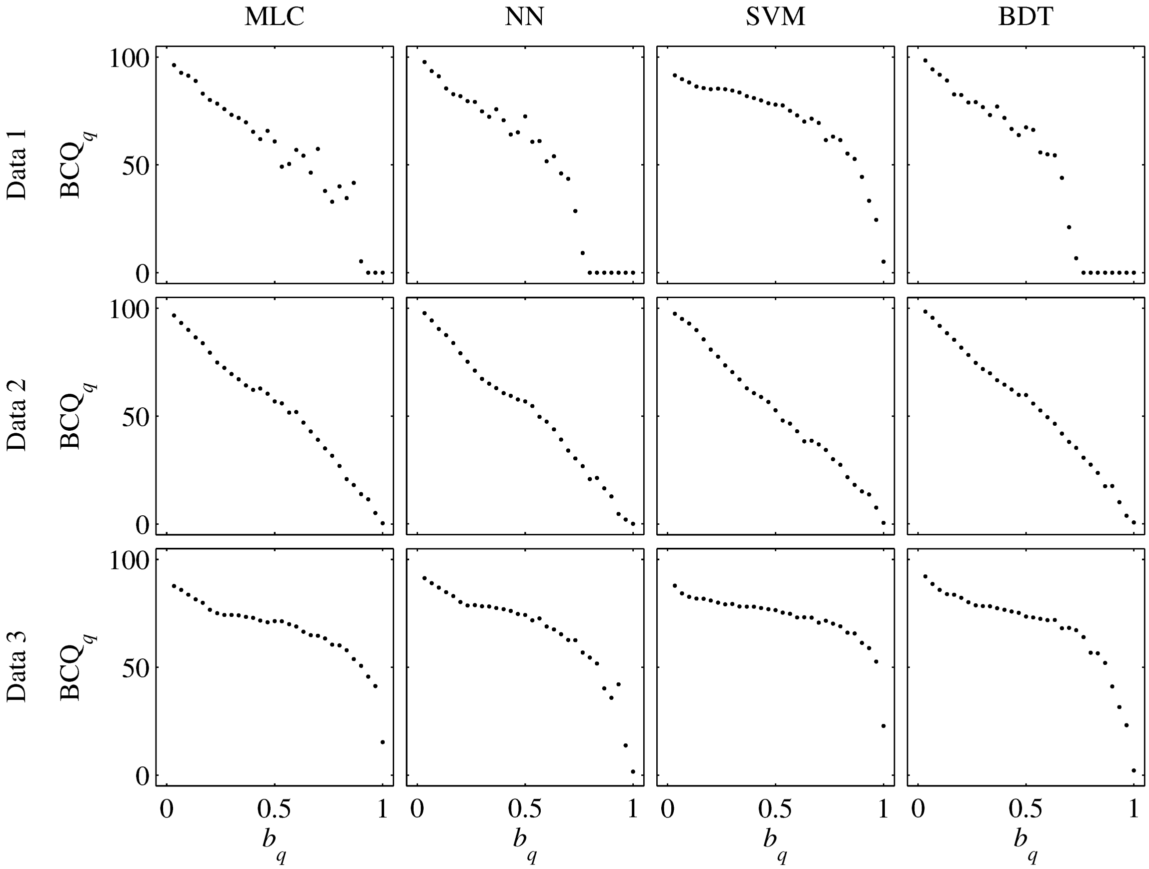


Figure 17 Scatter plot of BCQ$_q$ vs. b_q based on estimated MR

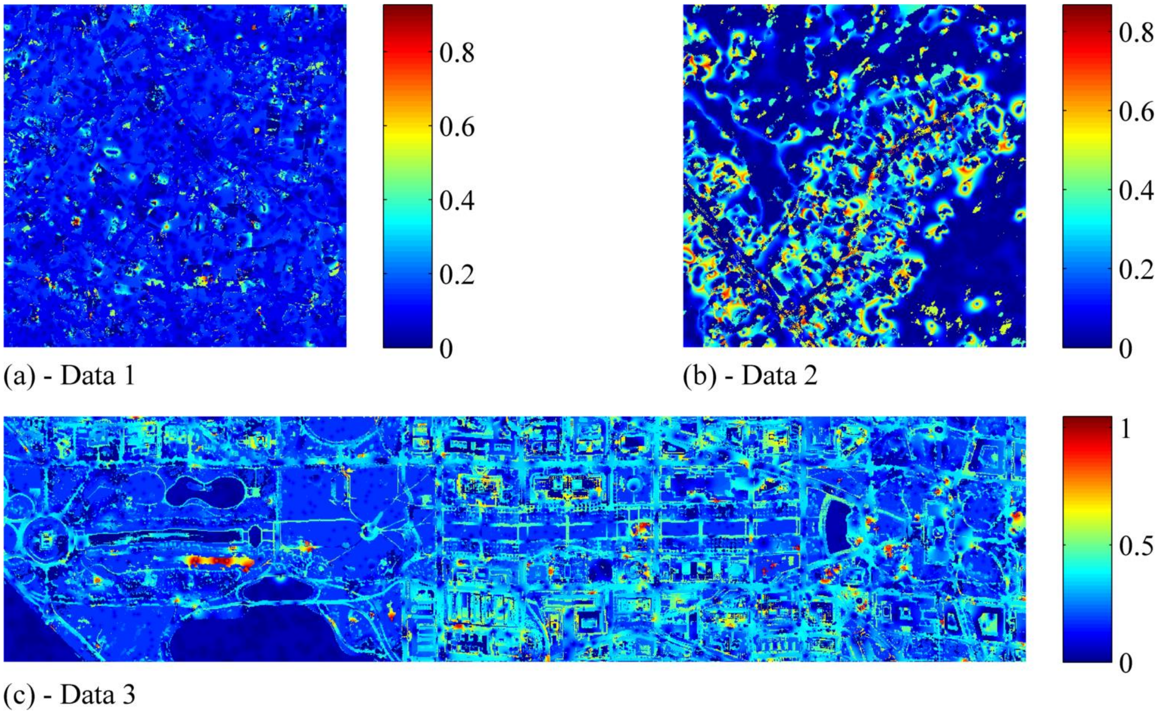
Table 7 Correlation coefficients (R) of BCQ$_q$ vs. b_q for MR (Q=30)

	MLC	**NN**	**SVM**	**BDT**
Data 1	-0.9525	-0.9551	-0.8792	-0.9539
Data 2	-0.9917	-0.9943	-0.9977	-0.9945
Data 3	-0.8886	-0.8889	-0.8087	-0.8628

3.3 Geostatistical method

3.3.1 Maps of estimated classification confidence

Figure 18 shows the maps of estimated CI based on maximum likelihood classification. Each panel is created using the geostatistical method introduced in Section 2.2.3. There are some spatial patterns in the estimated CI map. For example, data 1 has some hotspots scattered across the map. For Data 2 and 3, confusion is high in roof and concrete surface area. Deferent from the other two methods, the geostatistical method does not provide a true CI map. The indicator maps of true classification error are the same as that for the local error matrix method, which is shown in Figure 8. The comparison of Figure 18 and Figure 8 shows that CI estimated from geostatistical method have quite different spatial patterns of true classification error.



(a) - Data 1

(b) - Data 2

(c) - Data 3

Figure 18 Map of estimated CI (MLC)

3.3.2 Estimated CI of correct and error pixels

Figure 19 shows the bi-histogram of CI for correct pixels and error pixels. CI for correct pixels tends to concentrate at lower values, while CI for error pixels is more spread out across the range of [0,1] except for Data 1. For Data 1, CI for correct pixels concentrated between 0-0.2 while CI for error pixels have another small peak at the intervals of [0.3,0.5]. For Data 2, CI for correct pixels concentrated in [0-0.2] while CI for error pixels quite spread out on [0, 0.6]. CI for Data 3 have more variability with correct pixels spread out in [0, 0.5] and error pixels in [0, 0.7].

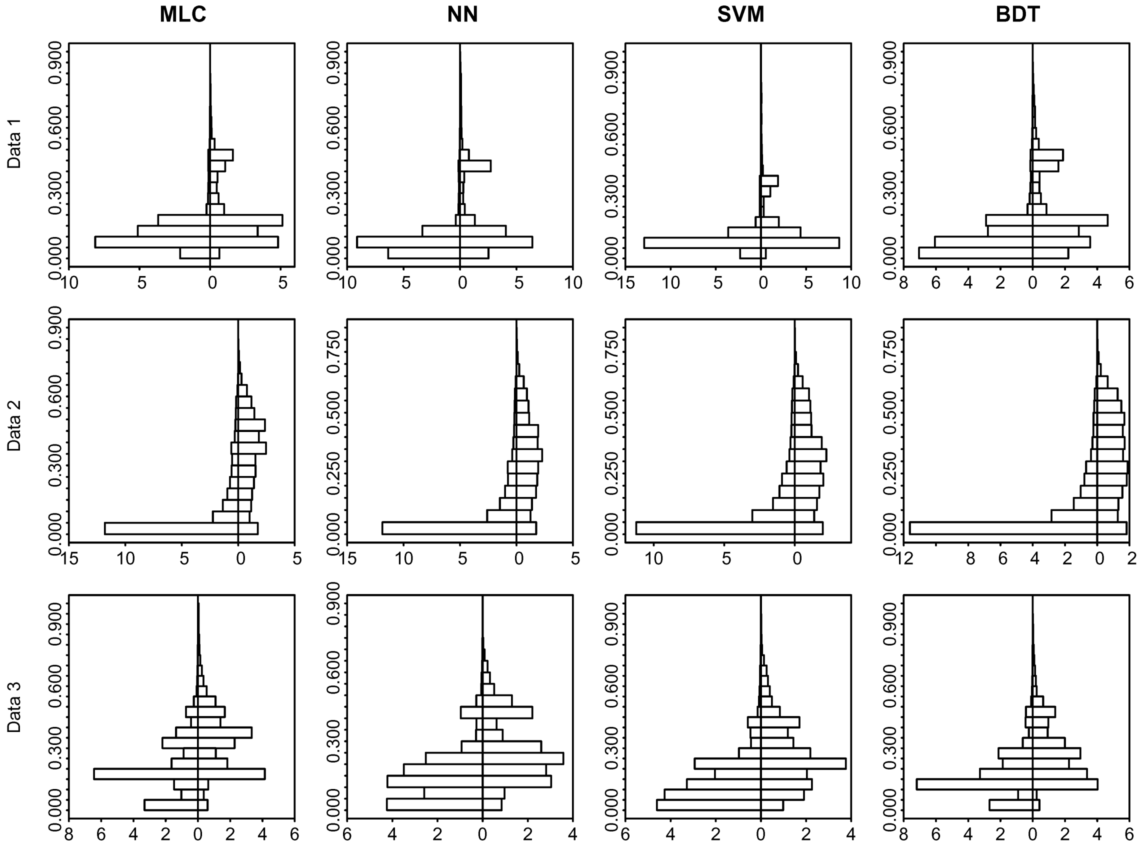


Figure 19 Bi-histogram of CI

Note: y - CI; x: proportion of pixels. Left side - correct pixel, right side - error pixel.

doi:10.20944/preprints202201.0352.v1

3.3.4 Relationship between binned classification quality and CI

Figure 20 shows the scatter plots of binned classification quality (BCQ$_q$) against CI (b_q) where CI is divided into 30 bins with equal distance. Generally, there is a negative relationship between BCQ$_q$ and b_q. However, the trend is not monotonic, especially for Data 2 with NN and SVM, Data 1 and Data 3 for all classifiers. The scatter plots for Data 2 with MLC and BDT mostly have a monotonically decreasing trend. In summary, CI is not a good measure for predicating per-pixel classification confidence.

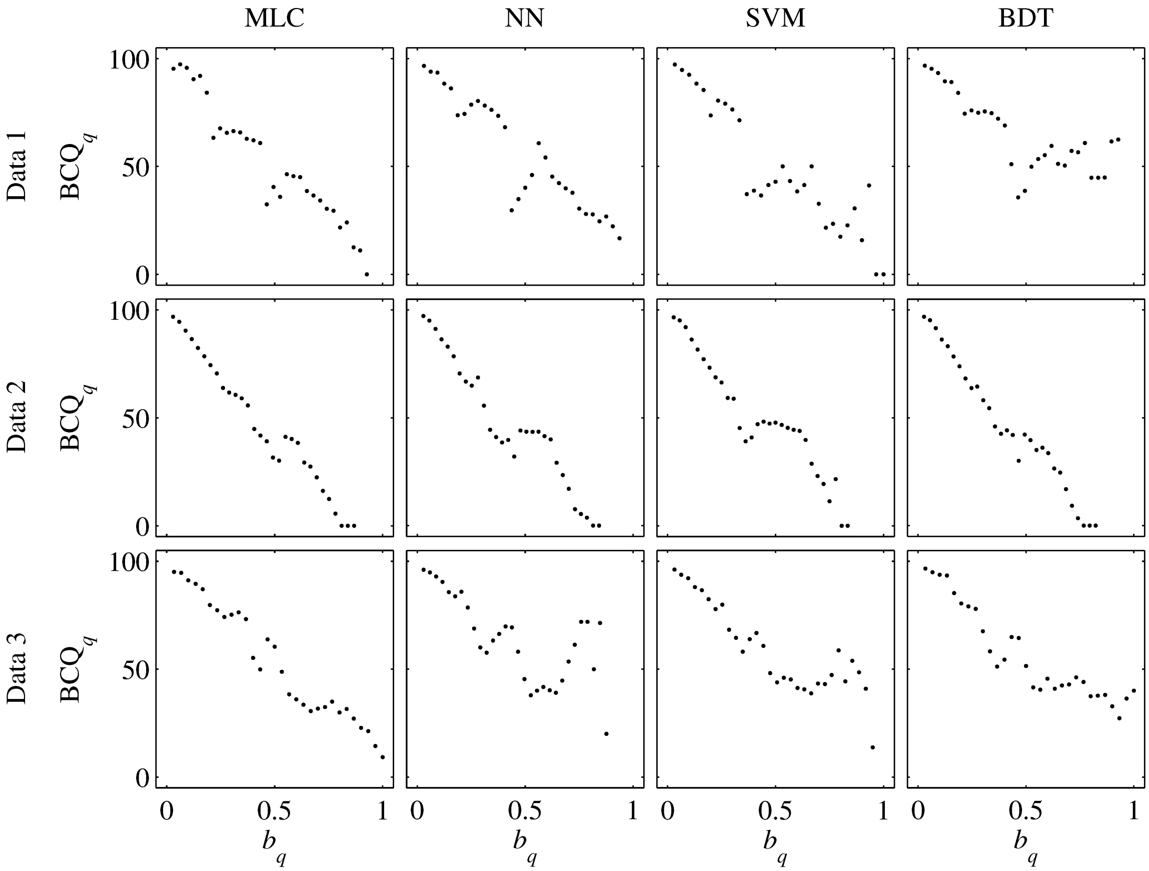


Figure 20 Scatter plot of BCQ$_q$ vs. b_q based on CI

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 24 January 2022 doi:10.20944/preprints202201.0352.v1

Table 8 shows the correlation coefficient R between BCQ$_q$ and b_q. The absolute values of R vary great with datasets and classifiers. They are high for Dat 1 with MLC, NN, and SVM, Data 2 with all classifiers, and Data 3 with MLC, and BDT, and low in other scenarios.

It can be concluded that LCA is not a good measure for characterizing per-pixel classification confidence.

Table 8 Correlation coefficients (R) of BCQ$_q$ vs. b_q for CI (Q=30)

	MLC	**NN**	**SVM**	**BDT**
Data 1	-0.9660	-0.9370	-0.9215	-0.7857
Data 2	-0.9868	-0.9656	-0.9555	-0.9856
Data 3	-0.9774	-0.7181	-0.8857	-0.9286

4 Discussion and conclusions

The performance of LCA, MR, and CI is varies across data and classifiers. Table 9 is a summary of the results for three interpolation based methods. The results can be summarized as follows.

(1) Local error matrix method is the least reliable one among the three methods. There is significant difference between estimated LCA and the true LCA. The estimated LCA are mostly concentrated around the overall accuracy of the whole map (between 0.7-0.9) while the true LCA are much more spread. The distributions of correct and error pixels are similar, which means LCA cannot distinguish error pixels from correct pixels. In other words, local error matrix is not a good method for estimating per-pixel classification confidence.

Table 9 Comparing the results of three interpolation-based methods

Items	**Local error matrix method**	**Bootstrap method**	**Geostatistical method**
Classification confidence of correct and error pixels	Bi-histogram shows that the distribution of correct and error pixels are similar.	Bi-histogram shows that the distribution of correct and error pixels are similar.	Bi-histogram shows that correct pixels tends to concentrate at lower values, while error pixels are more spread.
Estimated and true classification confidence	Estimated LCA are more spread than true LCA.	True MR's for are mostly concentrated in low values while estimated MR's are more spread out.	N.A.
Binned classification quality	There is no clear relationship between BCQ and LCA.	BCQ are negatively related to MR.	BCQ is somewhat negatively related to CI. However, the relationship is not monotonic for many scenarios.

(2) Bootstrap method may be useful in estimating per-pixel classification confidence. However, the distribution of estimated MR for correct and error pixels are very similar which means MR is not a good measure to distinguish error pixels from correct pixels. Also, the estimated MR is different from the true MR. There does exist correlation between BCQ and MR. BCQ is generally negatively related to MR but the exact relationship between binned classification quality and MR varies with datasets and classifiers.

(3) Geostatistical method is of limited use and is not reliable in estimating per-pixel classification confidence. The distributions of correct pixels and error pixels are different. CI of correct pixels tends to concentrate at lower values, while error pixels are more spread. While there is generally negative relationship between binned classification quality and CI, the

relationship is not monotonic for some datasets and classifiers. Pixels with high CI may also have high classification confidence.

The three interpolation-based method and their corresponding indices characterize the per-pixel classification confidence in different ways.

Local error matrix method is in essence a kind of smoothing window method. The LCA for each grid pixel is the unweighted average classification accuracy estimated based on its neighboring test pixels through the tool of local error matrix. The LCA for non-grid pixel is the weighted average based on neighboring grid pixels through IDW interpolation. Besides IDW, other interpolation techniques, e.g., kriging, can also be used. Due to the properties of local error matrix method, map of LCA is very smooth. My study shows that it is least useful in predicting per-pixel classification confidence.

Bootstrap method characterizes the separability of a pixel among different classes by resampling training data many times and thus results different classification rules. If a pixel cannot be classified correctly using different classification rules, then it is highly probable that this pixel will be classified wrong in practice. This study shows that bootstrap method is of limited use. Although there is negative relationship between BCQ and MR, the exact mathematical relationship varies with datasets and classifiers. As an interpolation method, the estimated MR map is smooth compared to the true MR map. Another issue is that MR cannot distinguish error pixels from correct pixels well.

Geostatistical method combines global error matrix with local variation of assigning class label to each pixel. Geostatistical method generates conditional probabilities, $p(u;k)$ of class labels for each pixel. These conditional probabilities are similar to the posterior probabilities, p, directly from classifiers. The difference between $p(u;k)$ and p is: $p(u;k)$ use the information of test data while *p* does not. The geostatistical method provides ideas for further studies. For example, it may be used to improve image classification. Geostatistical method is not reliable to predict per-pixel classification confidence either because the relationship between BCQ and CI is not monotonic.

In conclusion, the three interpolation methods provide some interesting insights on various aspects of estimating per-pixel classification confidence. Unfortunately, the interpolation assumes that classification confidence is smooth across the space. However, this is usually not true in practice. The interpolation effects hinder their practical use.

Supplementary Materials: The three datasets used in this paper remote sensing is uploaded as a zip file with this paper. Please refer to the readme.txt file after unzipping the data.

Funding: This research received no external funding.

Data Availability Statement: The data is uploaded with the paper as Supplementary Materials.

Conflicts of Interest: The author declares no conflict of interest.

References

Burnicki, A.C., 2011. Modeling the probability of misclassification in a map of land cover change. Photogrammetric Engineering & Remote Sensing 77, 39–49. https://doi.org/10.14358/PERS.77.1.39

Campbell, J.B., 1981. Spatial correlation effects upon accuracy of supervised classification of land cover. Photogrammetric Engineering and Remote Sensing 47, 355–363.

Comber, A., Fisher, P., Brunsdon, C., Khmag, A., 2012. Spatial analysis of remote sensing image classification accuracy. Remote Sensing of Environment 127, 237–246. https://doi.org/10.1016/j.rse.2012.09.005

Congalton, R.G., 1988. Using spatial autocorrelation analysis to explore the errors in maps generated from remotely sensed data. Photogrammetric Engineering & Remote Sensing 6.

Cressie, N., 1985. Fitting variogram models by weighted least squares. Journal of the International Association for Mathematical Geology 17, 563–586. https://doi.org/10.1007/BF01032109

Ebrahimy, H., Mirbagheri, B., Matkan, A.A., Azadbakht, M., 2021. Per-pixel land cover accuracy prediction: A random forest-based method with limited reference sample data. ISPRS Journal of Photogrammetry and Remote Sensing 172, 17–27. https://doi.org/10.1016/j.isprsjprs.2020.11.024

Foody, G.M., 2005. Local characterization of thematic classification accuracy through spatially constrained confusion matrices. International Journal of Remote Sensing 26, 1217–1228. https://doi.org/10.1080/01431160512331326521

Foody, G.M., 2002. Status of land cover classification accuracy assessment. Remote Sensing of Environment 80, 185–201. https://doi.org/10.1016/S0034-4257(01)00295-4

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Ji, L., Gallo, K., 2006. An Agreement Coefficient for Image Comparison. Photogrammetric Engineering & Remote Sensing 72, 823–833. https://doi.org/10.14358/PERS.72.7.823

Journel, A., 1986. Constrained interpolation and qualitative information—the soft kriging approach. Mathematical Geology 18, 269–286.

Kyriakidis, P.C., Dungan, J.L., 2001. A geostatistical approach for mapping thematic classification accuracy and evaluating the impact of inaccurate spatial data on ecological model predictions. Environmental and Ecological Statistics 8, 311–330.

Landgrebe, D., Biehl, L., 2020. MultiSpec: A Freeware Multispectral Image Data Analysis System. Purdue University, West Lafayette, IN.

Landgrebe, D.A., 2003. Signal theory methods in multispectral remote sensing. John Wiley & Sons.

Liu, D., Xia, F., 2010. Assessing object-based classification: advantages and limitations. null 1, 187–194. https://doi.org/10.1080/01431161003743173

Löw, F., Knöfel, P., Conrad, C., 2015. Analysis of uncertainty in multi-temporal object-based classification. ISPRS Journal of Photogrammetry and Remote Sensing 105, 91–106.

McIver, D.K., Friedl, M.A., 2001. Estimating pixel-scale land cover classification confidence using nonparametric machine learning methods. IEEE Transactions on Geoscience and Remote Sensing 39, 1959–1968. https://doi.org/10.1109/36.951086

Smith, J.H., 2002. Impacts of patch size and land-cover heterogeneity on thematic image classification accuracy. Photogrammetric Engineering & Remote Sensing 68, 65–70.

Steele, B.M., Winne, J.C., Redmond, R.L., 1998. Estimation and mapping of misclassification probabilities for thematic land cover maps. Remote Sensing of Environment 66, 192–202. https://doi.org/10.1016/S0034-4257(98)00061-3

van Oort, P.A.J., 2007. Interpreting the change detection error matrix. Remote Sensing of Environment 108, 1–8. https://doi.org/10.1016/j.rse.2006.10.012

van Oort, P.A.J., Bregt, A.K., de Bruin, S., de Wit, A.J.W., Stein, A., 2004. Spatial variability in classification accuracy of agricultural crops in the Dutch national land-cover database. International Journal of Geographical Information Science 18, 611–626. https://doi.org/10.1080/13658810410001701969

Wickham, J., Stehman, S.V., Homer, C.G., 2018. Spatial patterns of the United States National Land Cover Dataset (NLCD) land-cover change thematic accuracy (2001–2011). International Journal of Remote Sensing 39, 1729–1743. https://doi.org/10.1080/01431161.2017.1410298

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bulletin of the American Meteorological Society 63, 1309–1313.

Willmott, C.J., 1981. On the validation of models. Physical geography 2, 184–194.

Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. International Journal of climatology 32, 2088–2094.

Yu, Q., Gong, P., Tian, Y.Q., Pu, R., Yang, J., 2008. Factors affecting spatial variation of classification uncertainty in an image object-based vegetation mapping. Photogrammetric Engineering & Remote Sensing 74, 1007–1018. https://doi.org/10.14358/PERS.74.8.1007