The process of a supervised LULC classification is composed of a training stage and a performance evaluation (or validation) stage, and a resultant confusion matrix is used for accuracy assessment. For a better definition of the classification accuracy under different situations, we devise the following general expression of LULC classification accuracy.

Let Ω represent the set of all pixels in the study area, i.e. the global dataset, and S_T and S_R, two independent datasets of known ground-truth LULC classes, represent the training sample and reference sample, respectively. We shall adopt the convention of $P(S_1, S_2)$ for a general expression of various measures of LULC classification accuracy. In this expression, S_1 and S_2 represents the training dataset and the validation dataset, respectively. Thus, the conventional class-specific accuracies in the training-sample confusion matrix and the reference-sample confusion matrix and two other global accuracy measures can be defined as shown in Table 3.

Table 3. Different measures of LULC classification accuracy and their data dependency.

Accuracy measure	Expression	Sample data dependency
Training-sample accuracy	$P_i(S_T, S_T), i = 1,2,\cdots,k$	S_T
Reference-sample accuracy	$P_i(S_T, S_R)$	S_T and S_R
Training-sample-based global accuracy	$P_i(S_T, \Omega)$	S_T
Classifier global accuracy	$P_i(\Omega, \Omega)$	None

The training-sample-based global accuracy represents the classification accuracy of the thematic map. The training-sample accuracy, reference-sample accuracy and training-sample-based global accuracy all are dependent on the training dataset, and thus, conclusions of accuracy assessment using these measures are subject to the training data uncertainty. By contrast, the classifier global accuracy represents the accuracy achieved by using the population, i.e. the global dataset, to establish the classification rules, and thus only depends on the classifier adopted for LULC classification. Given a specific classifier, the classifier global accuracy (be the producer's, user's or the overall accuracy) has a unique and theoretical value. The global accuracy, either the training-sample-based global accuracy or the classifier

global accuracy, is unknown and can only be estimated by using the confusion matrix of LULC classification results. For convenience of discussion, we consider only the producer's accuracy, although the same discussions and simulations also apply to the user's accuracy and the overall accuracy. Definition of the classifier global accuracy provides an objective measure for comparing the performance of different classifiers in LULC classification.

Most studies assessed the accuracy and uncertainty of LULC classification results by using the reference-sample confusion matrix. Such practices aim to estimate the training-sample-based global accuracy by using the reference-sample classification accuracy. However, the target accuracy, i.e. $P_i(S_T, \Omega)$, itself is dependent on the training data S_T and thus conclusions drawn from such practices are inherently influenced by the selection of training samples. Even for a given training sample S_T, the reference-sample accuracy $P_i(S_T, S_R)$ is still subject to reference-sample uncertainty. Therefore, we propose using the classifier global accuracy $P_i(\Omega, \Omega)$ as the target accuracy since it is not subject to the training and reference data uncertainty and allows the users to compare the LULC classification performance of different classifiers.

To demonstrate the usefulness and advantage of using the classifier global accuracy for assessment of LULC classification results, we conducted rigorous stochastic simulations of multi-class multivariate Gaussian distributions to mimic the LULC classification and then compare three evaluation approaches. Details of our stochastic simulations and the three evaluation approaches are described below.

Consider an example that k land-cover types ($C_i, i = 1,2,\cdots,k$) are present in a study area. Suppose that m sets of sample data, say $S = \{S_1, S_2, \cdots, S_m\}$, are available. Each sample dataset is composed of pixels of *known* class-identities from the k land-cover types. We assume that all sample datasets have been sampled by simple random sampling or stratified random sampling. In an LULC classification, one of the m sample datasets, for example S_ℓ, is chosen

as the *training* sample and the rest of m-1 datasets can be considered as *reference* samples. If all sample datasets were obtained through the same sampling criteria or procedures, the so called 'training sample' and 'reference sample' are not statistically different. The sample dataset used for determining the discriminant functions or classification rules in LULC classification is considered as the 'training sample' and any of the m sample datasets can be utilized as the training sample. Readers are reminded that evaluation of LULC classification accuracies can be conceived as a work of parameter estimation. For every evaluation approach, there exist a target accuracy, i.e. the parameter to be estimated, and an estimate of the target accuracy which is often derived from the LULC confusion matrix.

Approach I – the reference-sample-based evaluation approach

This is the commonly adopted approach for assessment of LULC classification accuracy by using the reference sample. Upon completion of an LULC classification using a particular sample dataset, say S_ℓ, as the training sample, performance of the LULC classification can be evaluated by using any of the remaining m-1 sets of reference sample $(S_j, j = 1,2,\cdots,m; j \neq \ell)$. Let $p_i(S_\ell, S_j)$ represent the producer's accuracy of the i-th land-cover class using S_ℓ as the training sample and the j-th sample dataset S_j as the reference sample. We refer to $p_i(S_\ell, S_j)$ as the *reference-sample* producer's accuracies. This evaluation approach aims to estimate the training-sample-based global accuracy, i.e. $p_i(S_\ell, \Omega)$, by using the reference-sample classification accuracy, i.e. $p_i(S_\ell, S_j)$, as the estimator. Apparently, for a given set of training sample S_ℓ, the value of $p_i(S_\ell, S_j)$ varies with land-cover classes and reference samples, and the estimation can be expressed by

$$\hat{p}_i(S_\ell, \Omega) = p_i(S_\ell, S_j);\ \ell \neq j. \tag{8}$$

Using a large number of reference samples $(S_j, j = 1,2,\cdots,m; j \neq \ell)$, the uncertainty of the estimator can be evaluated. As the number of reference samples increases, we can expect

the mean value of the reference-sample producer's accuracy approaches to the true producer's global accuracy achieved by using S_ℓ as the training sample, i.e.,

$$\frac{1}{(m-1)}\sum_{\substack{j=1,\\ j\neq\ell}}^{m} p_i(S_\ell, S_j) \xrightarrow[m\to+\infty]{} p_i(S_\ell, \Omega). \tag{9}$$

In most practices of remote sensing LULC classification, we usually have limited number of reference samples. Therefore, using only one or a few sets of reference samples, it is difficult to conduct a meaningful evaluation of the classification results. Equation (9) shows that, at its best, this approach can only provide a good estimate of the producer's *global* accuracy $p_i(S_\ell, \Omega)$ achieved by a specific training sample S_ℓ.

Approach II – the training-sample-based evaluation approach

If each of the sample datasets $\{S_1, S_2, \cdots, S_m\}$ was used as the training sample in an LULC classification, it would yield *m* sets of training-sample confusion matrix. This evaluation approach aims to estimate the classifier global accuracy, i.e. $p_i(\Omega, \Omega)$, by using the training-sample accuracy, i.e. $p_i(S_\ell, S_\ell)$, as the estimator,

$$\hat{p}_i(\Omega, \Omega) = p_i(S_\ell, S_\ell). \tag{10}$$

Suppose that all possible samples of a fixed sample size, i.e. the ensemble of samples, are available. Then, as the number of training samples increases, the mean of the training-sample accuracy approaches to the classifier global accuracy i.e.,

$$\frac{1}{m}\sum_{\ell=1}^{m} p_i(S_\ell, S_\ell) \xrightarrow[m\to+\infty]{} p_i(\Omega, \Omega). \tag{11}$$

The above equation indicates that the ensemble mean ($m \to +\infty$) of the training-sample accuracy equals the classifier global accuracy. In real practice of LULC classification, we have only one set of training sample (m = 1) and thus the only training-sample accuracy is used as an estimate of the classifier global accuracy and the training-sample-based evaluation is subject to training data uncertainty.

Approach III – the bootstrap-sample-based evaluation approach

Both the reference-sample-based and the training-sample-based evaluation approaches are subject to training data uncertainty. In this third approach, we aim to estimate the classifier global accuracy by providing a confidence interval of the classifier global accuracy. This is achieved by bootstrap resampling from the only training sample.

Given a training dataset S_ℓ, suppose that a large number (for example, M = 1000) of bootstrap samples, $S_1^B, S_2^B, \cdots, S_M^B$, were generated from the training dataset. We then conduct LULC classification using each of these bootstrap samples as the training sample, and M sets of *bootstrap-sample* accuracy, i.e. $p_{i\ell}(S_j^B, S_j^B),\ j = 1, 2, \cdots, M;\ i = 1, 2, \cdots k$, are obtained. Note that the subscript ℓ indicates that bootstrap samples are generated from the training dataset S_ℓ and the bootstrap-sample accuracy is dependent on the training dataset. Details of bootstrap resampling and its application for LULC classification can be found in Horowitz (2001) and Hsiao and Cheng (2016).

Let q_1^B and q_2^B respectively represent the 0.025 and 0.975 sample quantiles of $p_{i\ell}(S_j^B, S_j^B),\ j = 1, 2, \cdots, M$, then $[q_1^B, q_2^B]$ forms a 95% confidence interval of $p_i(\Omega, \Omega)$, i.e.,

$$P[q_1^B \le p_i(\Omega, \Omega) \le q_2^B] = 0.95 \tag{12}$$

It is worthy to note that, with increasing number of bootstrap samples, the mean of bootstrap-sample accuracy approaches to the training-sample accuracy, i.e.,

$$\frac{1}{M}\sum_{j=1}^{M} p_{i\ell}(S_j^B, S_j^B) \xrightarrow[m\to+\infty]{} p_i(S_\ell, S_\ell). \tag{13}$$

Combining Equations (11) and (13), it yields

$$\frac{1}{m}\sum_{\ell=1}^{m}\left(\frac{1}{M}\sum_{j=1}^{M} p_{i\ell}(S_j^B, S_j^B)\right) \xrightarrow[M\to+\infty]{m\to+\infty,} p_i(\Omega, \Omega). \tag{14}$$

If only one set of training sample is available (m = 1), then the mean and sample quantile range $[q_1^B, q_2^B]$ of the bootstrap-sample accuracy are a point estimate and 95% confidence interval of the classifier global accuracy, respectively. To validate the above relationships

between various accuracy measures of LULC classification and to demonstrate the advantage of using the classifier global accuracy for assessment of LULC classification results, we carried out stochastic simulation for a simple case (two-class and two-feature, 2C2F) and a more complicated case (four-class and three-feature, 4C3F) of LULC classification.

3. Stochastic Simulation of LULC Classification

In this section, we consider a special case of LULC classification with two land-cover classes (C_1 and C_2) and two classification features (X_1 and X_2). For each land-cover class, the two classification features form a bivariate Gaussian distribution. The mean vector, covariance matrix of classification features and *a priori* probabilities of C_1 and C_2 are listed in Table 4. The two classification features are negatively correlated ($\rho = -0.75$) for C_1 and positively correlated ($\rho = 0.65$) for C_2.

Table 4. Parameters of the bivariate Gaussian distributions of classes C_1 and C_2 (2C2F case).

Parameters	Class 1	Class 2
Mean vector	$\begin{bmatrix} 80 \\ 120 \end{bmatrix}$	$\begin{bmatrix} 140 \\ 150 \end{bmatrix}$
Covariance matrix	$\begin{bmatrix} 1225 & -525 \\ -525 & 400 \end{bmatrix}$	$\begin{bmatrix} 900 & 390 \\ 390 & 400 \end{bmatrix}$
A priori probability	0.4	0.6

For a k-class, p-feature LULC classification using multispectral remote sensing images, a pixel can be characterized by a feature vector $X^T = (x_1, x_2, \cdots, x_p)$ and the probability density function of the i-th class can be expressed by

$$f(X|C_i) = \frac{1}{\sqrt{2\pi}^p} \exp\left[-\frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i)\right], i = 1,2,\cdots,k. \quad (15)$$

where μ_i and Σ_i are the mean vector and covariance matrix, respectively. For our simulation, the Bayes classification method which considers the *a priori* probabilities of individual LULC classes was chosen as the classifier. The class-specific discriminant functions of the Bayes classification are

$$d_i(X) = \ln p(C_i) - \frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(X-\mu_i)^T \Sigma_i^{-1}(X-\mu_i), i = 1,2,\cdots,k. \quad (16)$$

where $p(C_i)$ represents the a priori probability of the i-th class. A pixel with feature vector X is assigned to the class having the highest value of discriminant function, i.e.,

$$\text{Assign } X \text{ to } C_i \;\; if \; d_i(X) > d_j(X), \;\; j = 1,2,\cdots,k; \; j \neq i. \tag{17}$$

Simulation settings and details of the three evaluation approaches are described below.

 # Simulation, classification and evaluation of the reference-sample-based approach

Simulate 1001 sample datasets of the two-class bivariate Gaussian distribution, i.e., $\boldsymbol{S} = (\boldsymbol{S_1}, \boldsymbol{S_2}, \cdots, \boldsymbol{S_{1001}})$ $(\boldsymbol{m = 1001})$,

For $\boldsymbol{\ell}$ from 1 to m { **# a total of m simulation runs**

Consider $\boldsymbol{S_\ell}$ as the training dataset (S_ℓ^1 for C_1 and S_ℓ^2 for C_2) and calculate the sample parameters of the bivariate Gaussian distributions of C_1 and C_2,

Determine the discriminant functions of C_1 and C_2 by using the above sample parameters,

Apply the above discriminant functions to $\boldsymbol{S_\Omega}$ (described in Section 3.1) and calculate the *training-sample-based global* (producer's, user's, and overall) accuracies of C_1 and C_2, i.e., $\boldsymbol{p_i(S_\ell, \Omega)}$,

For each of the remaining dataset $\boldsymbol{S_j, j = 1, 2, \cdots, m; j \neq \ell}$ {

Conduct classification on $\boldsymbol{S_j}$ by using the discriminant functions established by $\boldsymbol{S_\ell}$,

Calculate the *reference-sample* producer's, user's, and overall accuracies of C_1 and C_2,

}

Calculate the mean of the m-1 reference-sample producer's, user's, and overall accuracy, respectively.

Compare the training-sample-based global accuracy and the mean reference-sample accuracy.

}