A NEW THINKING OF LULC CLASSIFICATION ACCURACY ASSESSMENT

K. S. Cheng [1, 2, *], J.Y. Ling [1], T. W. Lin [1], Y. T. Liu [1], Y. C. Shen [1], Y. Kono [3]

[1] Dept. of Bioenvironmental Systems Engineering, National Taiwan University, Taiwan, R.O.C. – (rslab, r05622005, b03602057, b03602007, b03602023)@ntu.edu.tw
[2] Master Program in Statistics, National Taiwan University, Taiwan, R.O.C.
[3] Center for Southeast Asian Studies, Kyoto University, Kyoto, Japan - kono@cseas.kyoto-u.ac.jp

Commission VI, WG VI/4

KEY WORDS: Land-Use/Land-Cover (LULC), Accuracy Assessment, Confusion Matrix, Confidence Interval, Bootstrap Resampling

ABSTRACT:

A majority of studies involving remote sensing LULC classification conducted classification accuracy assessment without consideration of the training data uncertainty. In this study we present new concepts of LULC classification accuracies, namely the training-sample-based global accuracy and the classifier global accuracy, and a general expression of different measures of classification accuracy in terms of the sample dataset for classifier training and the sample dataset for evaluation of classification results. Through stochastic simulation of a two-feature and two-class case, we demonstrate that the training-sample confusion matrix should replace the commonly adopted reference-sample confusion matrix for evaluation of LULC classification results. We then propose a bootstrap-simulation approach for establishing 95% confidence intervals of classifier global accuracies.

1. INTRODUCTION

1.1 General Instructions

When conducting a supervised LULC classification using remote sensing images, a set of multi-class ground-truth training samples is collected and used to establish classification rules and multi-class boundaries in the feature space. This is referred to as the training stage of LULC classification. In the validation stage, the classification rules established by the training data are applied to an independent set of multi-class ground-truth reference samples. Different measures of classification accuracy of the reference samples are then summarized in a confusion matrix (or error matrix) for assessment of the classification accuracies and performance of the LULC classification. A critical assumption of the classification accuracy analysis is that the confusion matrix is truly representative of the classification results of the entire study area. Class-specific producer's and user's accuracy summarized in a confusion matrix can be considered as *sample accuracy* and are only estimates of the true, yet unknown, *global accuracy* (or *population accuracy*) concerning the entire study area (Hay, 1988; Stehman and Czaplewski, 1998). These accuracies or errors are inherently associated with uncertainties due to variability or uncertainty in selection of training and reference samples (Weber and Langille, 2007).

Many classification accuracy assessments were conducted using the same dataset as was used to train the classifier. Such training and validating on the same dataset could result in overestimation of classification accuracy (Congalton, 1991). In this paper, the confusion matrix established by using the same dataset for training and validation is referred to as the *training-sample confusion matrix* and the confusion matrix established by using an independent dataset of reference samples is called the *reference-sample confusion matrix*. Although assessing the reference-sample confusion matrix has become a common practice for evaluation of LULC classification results, accuracies presented in the reference-sample confusion are dependent on the training samples through the classification rules established in the training stage of LULC classification. Thus, a thorough assessment of classification accuracies needs to take the uncertainty in training data selection into account. Therefore, the objectives of this study are two-fold: (1) to investigate the effect of training and reference data selection on classification accuracy and (2) to propose an approach for a quantitative assessment of the uncertainty in LULC classification results.

2. THEORETICAL BASES

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 1.

* Corresponding author

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

Table 1. Different measures of LULC classification accuracy.

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.

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 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. 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.

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. 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.

2.1 The reference-sample-based evaluation approach

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 reference sample sets ($S_j, j = 1,2,\cdots,m; j \neq \ell$). 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

$$\hat{p}_i(S_\ell, \Omega) = p_i(S_\ell, S_j);\ \ell \neq j. \tag{1}$$

Using a large number of reference samples, 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.,

$$\frac{1}{(m-1)}\sum_{\substack{j=1,\\ j\neq\ell}}^{m} p_i(S_\ell, S_j) \xrightarrow[m\to+\infty]{} p_i(S_\ell, \Omega). \tag{2}$$

In practice 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.

2.2 The training-sample-based evaluation approach

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,

$$\hat{p}_i(\Omega, \Omega) = p_i(S_\ell, S_\ell). \tag{3}$$

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.,

$$\frac{1}{m}\sum_{\ell=1}^{m} p_i(S_\ell, S_\ell) \xrightarrow[m\to+\infty]{} p_i(\Omega, \Omega). \tag{4}$$

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.

2.3 The bootstrap-sample-based evaluation approach

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).

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.,

$$P[q_1^B \le p_i(\Omega,\Omega) \le q_2^B] = 0.95 \quad (5)$$

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.,

$$\frac{1}{M}\sum_{j=1}^{M} p_{i\ell}(S_j^B, S_j^B) \xrightarrow[m\to+\infty]{} p_i(S_\ell, S_\ell). \quad (6)$$

Combining Equations (4) and (6), it yields

$$\frac{1}{m}\sum_{\ell=1}^{m}\left(\frac{1}{M}\sum_{j=1}^{M} p_{i\ell}(S_j^B, S_j^B)\right) \xrightarrow[M\to+\infty]{m\to+\infty,} p_i(\Omega,\Omega). \quad (7)$$

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.

3. STOCHASTIC SIMULATION OF LULC CLASSIFICATION

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). Let the two classification features form a bivariate Gaussian distribution and the mean vector, covariance matrix of classification features and *a priori* probabilities of C_1 and C_2 are listed in Table 2. The two classification features are negatively correlated ($\rho = -0.75$) for C_1 and positively correlated ($\rho = 0.65$) for C_2.

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

Table 2. Parameters of the bivariate Gaussian distribution.

For our simulation, the Bayes classification method which considers the *a priori* probabilities of individual LULC classes was chosen as the classifier. For the above setting, the classifier global accuracies, i.e., $\boldsymbol{p_i(\Omega,\Omega)}, i = 1, 2$, are shown in Table 3.

Global Accuracy	Class 1	Class 2
Producer's	0.94385	0.91146
User's	0.87664	0.96055
Overall	0.92442	

Table 3. Classifier global accuracies by the Bayes classifier.

The objective of our simulation is to demonstrate that a 95% confidence interval of the classifier global accuracy can be established by using bootstrap samples. We generated a large number (1000) of training datasets from the two-class bivariate Gaussian distribution. Then, for each training dataset, 1000 sets of bootstrap samples were generated and used to establish a confidence interval of the classifier global accuracy. Finally, we evaluated the proportion of these confidence intervals covering the classifier global accuracy.

4. RESULTS AND DISCUSSION

4.1 Two-class, two-feature case

For convenience of explanation, abbreviations PA, UA and OA represent the producer's, user's, and overall accuracies, respectively. A number affixed to PA and UA indicates the Land-cover class. For example, PA1 represents the producer's accuracy of C_1.

Figure 1 demonstrates the 95% bootstrap confidence intervals and the mean bootstrap-sample accuracies for 100 (301 – 400) sets of training samples. For every training sample set, the mean of 1000 bootstrap-sample accuracies falls very close to the training-sample accuracy. The covering rates of classifier-global-accuracy were 0.977, 0.945, 0.947, 0.977, and 0.953 for PA1, PA2, UA1, UA2, and OA, respectively. These covering rates are slightly higher than (for PA1 and UA2) or nearly equal to 0.95 (for PA2, UA1, and OA), indicating the practical applicability of the bootstrap confidence interval proposed in this study.

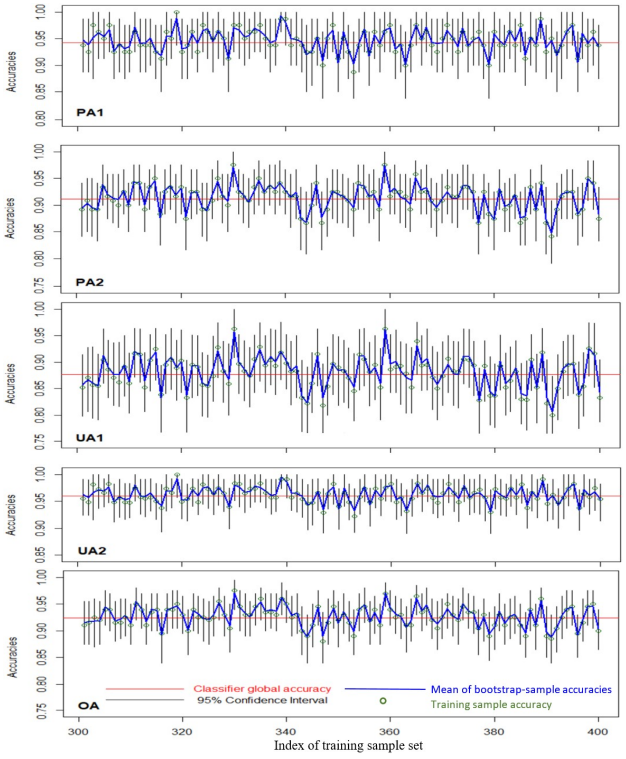

Figure 1. Illustration of 95% bootstrap confidence intervals for the 2-class, 2-feature case.

We also investigated how the number of bootstrap samples affects the classifier-global-accuracy covering rates of the 95% bootstrap confidence interval. Table 4 shows that covering rates of all accuracy measures become more stabilized for 100 or more bootstrap samples. Thus, we recommend using at least 100 bootstrap samples to construct the 95% bootstrap confidence intervals of different classifier global accuracies.

4.2 Four-class, three-feature case

In order to demonstrate the capability of the proposed bootstrap-sample-based evaluation approach for more complicated LULC classification applications, we conducted

similar stochastic simulation for a 4-class, 3-feature LULC classification case. We assume the three classification features of individual land-cover classes form a trivariate Gaussian distribution. The mean vector, covariance matrix of classification features of individual land-cover classes are listed in Table 5. The *a priori* probabilities of individual land-cover classes (Class 1 through 4) are 0.2, 0.4, 0.25 and 0.15, respectively. An exemplar demonstration of training samples is shown in Figure 2.

Number of bootstrap samples	PA1	PA2	UA1	UA2	OA
30	0.940	0.936	0.934	0.939	0.924
40	0.953	0.937	0.936	0.953	0.937
50	0.971	0.960	0.959	0.970	0.951
100	0.966	0.945	0.947	0.968	0.953
200	0.973	0.953	0.950	0.973	0.945
300	0.970	0.972	0.973	0.971	0.968
400	0.971	0.946	0.942	0.971	0.936
500	0.973	0.959	0.958	0.974	0.945
1000	0.977	0.945	0.947	0.977	0.953

Table 4. Classifier-global-accuracy covering rates of 95% bootstrap-sample confidence intervals with respect to different number of bootstrap samples.

	Mean vector	Covariance matrix
Class 1	$\begin{bmatrix} 87.96 \\ 61.85 \\ 118.42 \end{bmatrix}$	$\begin{bmatrix} 66.65 & 62.86 & 5.78 \\ 62.86 & 77.46 & -8.41 \\ 5.78 & -8.41 & 140.11 \end{bmatrix}$
Class 2	$\begin{bmatrix} 127.69 \\ 116.18 \\ 80.31 \end{bmatrix}$	$\begin{bmatrix} 161.54 & 53.49 & 39.35 \\ 53.49 & 177.16 & 64.00 \\ 39.35 & 64.00 & 159.26 \end{bmatrix}$
Class 3	$\begin{bmatrix} 74.90 \\ 49.92 \\ 92.98 \end{bmatrix}$	$\begin{bmatrix} 29.93 & 27.92 & 12.57 \\ 27.92 & 35.09 & 1.90 \\ 12.57 & 1.90 & 137.73 \end{bmatrix}$
Class 4	$\begin{bmatrix} 104.90 \\ 86.42 \\ 89.73 \end{bmatrix}$	$\begin{bmatrix} 66.23 & 42.80 & 14.08 \\ 42.80 & 106.03 & -9.52 \\ 14.08 & -9.52 & 175.32 \end{bmatrix}$

Table 5. Parameters of the trivariate Gaussian distribution of the 4-class, 3-feature case.

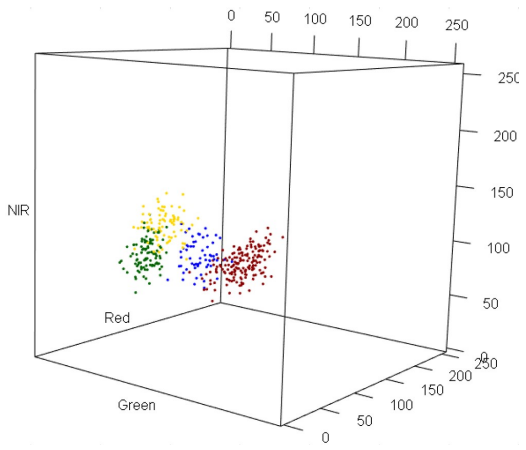

Figure 2. An exemplar demonstration of training data for the 4-class, 3-feature case.

The 95% bootstrap confidence intervals and the mean bootstrap-sample accuracies for 100 (301 – 400) sets of training samples for the 4-class 3-feature case is shown in Figure 3. For every training sample set, the mean of 1000 bootstrap-sample accuracies falls very close to the training-sample accuracy. The covering rates of classifier-global-accuracy were 0.96, 0.976, 0.949, 0.955, 0.958, 0.955, 0.963, 0.958 and 0.936 for PA1 to PA4, UA1 to UA4, and OA, respectively.

5. SUMMARY AND CONCLUSIONS

In This study we present new concepts of LULC classification accuracies, namely the training-sample-based global accuracy and the classifier global accuracy, and a general expression of different measures of classification accuracy in terms of the sample dataset for classifier training and the sample dataset for evaluation of classification results. We also conducted stochastic simulations for a two-feature two-class LULC classification case to demonstrate the practical applicability of the proposed bootstrap simulation approach for establishing 95% confidence intervals of classifier global accuracies. The conclusions are as follows:

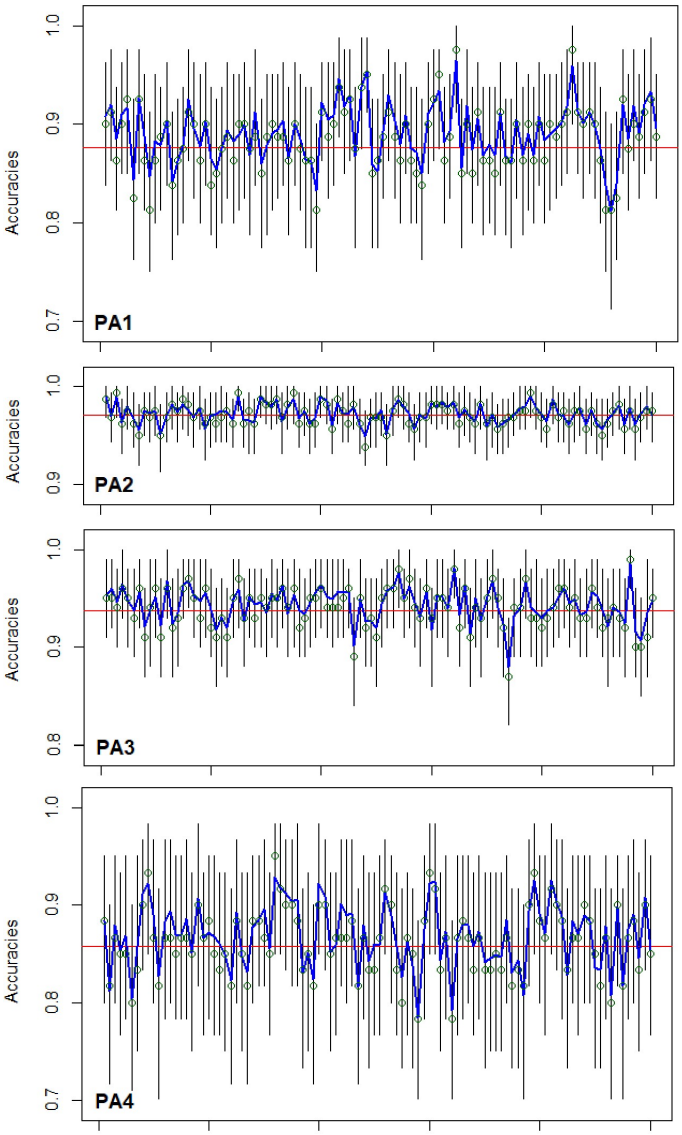

Figure 3. Illustration of 95% bootstrap confidence intervals for the 4-class, 3-feature case (continued).

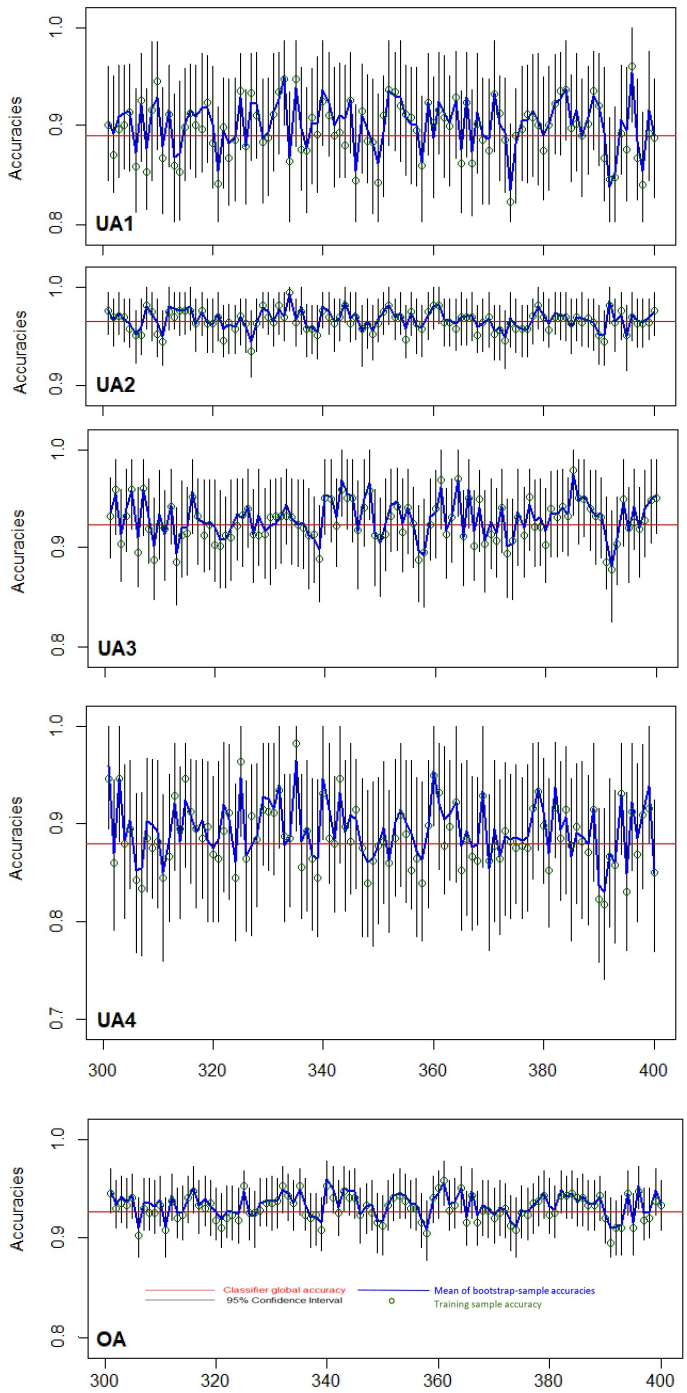

Figure 3. Illustration of 95% bootstrap confidence intervals for the 4-class, 3-feature case.

(1) The commonly adopted reference-sample classification accuracies are subject to uncertainties in the training and reference data. At its best, the reference-sample accuracy can only provide a good estimate of the *global* accuracy achieved by a specific training sample. It does not provide information about the *global* accuracy that can be achieved the classifier.
(2) Through rigorous stochastic simulations, we demonstrated the practical applicability of the proposed bootstrap confidence interval. We recommend using at least 100 bootstrap samples to construct the 95% bootstrap confidence intervals of different classifier global accuracies.

ACKNOWLEDGEMENTS

We acknowledge the financial support from the Ministry of Science and Technology of Taiwan through a project grant (MOST-104-2918-I-002-013). The corresponding author is grateful to the Center for Southeast Asian Studies (CSEAS) of the Kyoto University, Japan for hosting his sabbatical leave and providing excellent research environment and facilities.

REFERENCES

Hay, A.M., 1988. The derivation of global estimates from a confusion matrix. *International Journal of Remote Sensing*, 9, pp. 1395–1398.

Stehman, S.V., Czaplewski, R.L., 1998. Design and Analysis for Thematic Map Accuracy Assessment: Fundamental Principles. *Remote Sensing of Environment*, 64, pp. 331–344.

Weber, K.T., Langille, J., 2007. Improving classification accuracy assessments with statistical bootstrap resampling techniques. *GIScience & Remote Sensing*, 44, pp. 237–250.

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

Horowitz, J.L., 2001. The bootstrap. In *Handbook of Econometrics*; Heckman, J.J., Leamer, E., Eds.; North Holland Publishing Company: New York, NY, USA, Volume 5, pp. 3160–3228.