EFFECT OF ERRORS IN GROUND TRUTH ON CLASSIFICATION ACCURACY

Mark J. Carlotto[†]
General Dynamics Advanced Information Systems

Abstract

The effect of errors in ground truth on the estimated thematic accuracy of a classifier is considered. A relationship is derived between the true accuracy of a classifier relative to ground truth without errors, the actual accuracy of the ground truth used, and the measured accuracy of the classifier as a function of the number of classes. We show that if the accuracy of the ground truth is known or can be estimated, the true accuracy of a classifier can be estimated from the measured accuracy. In a series of simulations our method is shown to produce unbiased estimates of the true accuracy of the classifier with an uncertainty that depends on the number of samples and the accuracy of the ground truth. A method for determining the relative performance of two or more classifiers over the same area is then discussed. Results indicate that as the number of samples increases one can effectively differentiate between the performance of the classifiers using inaccurate ground truth. It is argued that relative accuracies computed using a large number of inaccurate ground truth points are more representative of the true relative performance of the classifiers since they are being evaluated over a larger portion of the scene. An example is presented that uses this method to evaluate the relative performance of two Landsat classifiers.

1. INTRODUCTION

Accuracy assessment has become an important topic given the increased use of remotely sensed imagery in mapping, environmental monitoring, and other application areas. In order to estimate thematic accuracy, ground truth (i.e., in situ measurements) and/or image truth (in effect, ground truth inferred from imagery) must be collected. Two key questions relate to the number and distribution of samples. The number of samples depends largely on the level of confidence and is well-understood. However, the manner

[†] 8 Milne Way Gloucester MA 01930 (mark.carlotto@gd-ais.com)

in which the samples are selected within the study area remains somewhat of an art due to the complexity of the spatial processes involved.

The accuracy of ground truth is rarely known but is usually assumed to be correct. Ground truth is almost never completely accurate due to differences between the time the imagery was acquired and the ground truth collected, inconsistencies in assigning classes to ground truth, and other factors, many of which are based on human judgment. If ground truth is assumed to be correct but is not, classification errors are blamed on the algorithm or the data, wrongly lowering the classification accuracy (Congalton 1991).

Currently the variability of the land use patterns is considered to exhibit scaling (power-law) behaviour. This suggests that the accuracy of the classifier is of critical importance. The estimation of the scaling exponents of the power-law could be assessed by using the detrended fluctuation analysis, which has already proved its usefulness in several complex systems, like the surface air-pollutants (Varotsos et al 2005), the total ozone content (Varotsos 2005) and the global tropospheric temperature (Varotsos and Kirk-Davidoff 2006).

This paper considers the accuracy of ground truth and its effect on thematic accuracy. In particular the following questions are considered: What is the effect of errors in ground truth on classification accuracy? What is the relation between the number of samples collected and the accuracy of the samples on classification accuracy? In Section 2 we begin by reviewing the relationship between the number of samples and the confidence of the estimated classification accuracy. A brief discussion of random and stratified random sampling is also provided. Our accuracy assessment approach is described in Section 3 which models the effect of errors in ground truth on the estimated accuracy of a classifier. Results of simulations to verify our model are presented. In Section 4 we apply our methodology to estimate the relative performance of two Landsat classifiers. Implications of our approach and areas for future work are discussed in Section 5.

2. BACKGROUND

Thematic accuracy is usually expressed in the form of a confusion matrix *P(k,k')* which summarizes the number of times the true class *k* was assigned class *k*' by the classifier. If the confusion matrix is normalized so that

$$\sum_{k=1}^{K}\sum_{k'=1}^{K}P(k,k')=1, \tag{1}$$

the accuracy of the classifier is equal to the sum of the diagonal terms

$$\alpha = \sum_{k=k'}P(k,k'). \tag{2}$$

Ground truth (i.e., in situ measurements) and/or image truth (in effect, ground truth inferred from imagery) are typically used to estimate of the accuracy of a classifier. Two questions often asked are:

- How many samples are required to obtain a reliable estimate of the classification accuracy?
- Where in the study area should the samples be acquired to obtain an unbiased estimate of the accuracy?

The answer to the first question depends, in part, on the level of confidence required of the estimate. Various approaches for estimating the number of samples required are reviewed by Janssen and van der Wel (1994). Often the binomial distribution is used to model the sampling process. For *N* total samples, the probability that *n* samples are correct is given by

$$P(n)=\binom{N}{n}\alpha^{n}(1-\alpha)^{N-n}. \tag{3}$$

The upper and lower limits of the estimated classification accuracy depends on the number of samples and on the confidence κ. These upper and lower limits α+ and α− are calculated from the binomial distribution

$$\frac{1-\kappa}{2}=\sum_{n=0}^{N-}P(n) \tag{4a}$$

and

$$\frac{1-\kappa}{2} = \sum_{n=N+}^{N} P(n) \tag{4b}$$

respectively where $N- = N\alpha-$ and $N+ = N\alpha+$ are the minimum and maximum values that satisfy the above equations. Fig. 1 plots the upper and lower limits of the classification accuracy as a function of the number of samples required to achieve a confidence of 0.95. Fig. 2 plots the lower limit of the classification accuracy for three levels of confidence. It is often concluded from such curves that several hundred samples are required in most applications to obtain satisfactory estimates.

The second question concerning the manner in which the samples are selected is more complex. Congalton (1988) states that the spatial complexity of a given environment dictates the appropriate sampling scheme to use. Based on a simulation study he concluded that simple random sampling produced results with the least amount of bias and was adequate for most situations. A disadvantage of random sampling is that the number of samples per category is, on average, proportional to the area of the category. This poses a problem in situations where important categories such as roads, small bodies of water, buildings and other isolated structures make up a very small fraction of the image. In such cases, random sampling may fail to generate a single sample for some categories.

Another method known as stratified random sampling generates a given number of samples per stratum, e.g., within each land cover category. Todd et al (1980) used this approach to estimate the accuracy of a classification map derived from Landsat MSS. A disadvantage of this kind of sampling scheme is that it makes it difficult to evaluate multiple classifications over the same area since the sampling is derived from the classifications.

Our work is concerned, in part, with the evaluation of multiple classifiers and data sources over the same area. In particular, we are interested in determining the best classifier based on its performance over a limited number of study areas. Sampling plans are based largely on the diversity and spatial distribution of land cover. However, classification performance depends on other factors as well. For example, the topography modulates the spectral response by a multiplicative factor that is a function of the slope. Variations in moisture and organic content affect the spectral response of exposed soils.

The spectral response of vegetation is affected by moisture stress, disease, and other factors. Different classifiers may be more or less sensitive to these and other effects. Depending on the complexity of the terrain and the environment, more ground truth may be required to insure that these factors are represented in the sample.

3. ACCURACY ASSESSMENT MODEL

Our goal is to begin to understand the effect of errors in ground truth; i.e., how does inaccurate ground truth affect estimates of classification accuracy. Our approach involves the development of an error model that can be used as a basis for simulating the effect of ground truth errors as a function of the error rate, the number of samples, and other parameters.

Ground truth is rarely if ever completely accurate due to differences between the time the imagery was acquired and the ground truth collected, inconsistencies in assigning classes to ground truth, and other factors, many of which are based on human judgment. Let $T = \{t_n\}$ be a series of ground truth samples t_n which take on integer values between one and K, where K is the number of classes. We assume that the t_n cannot be observed directly. Instead we observe another sequence $R = \{r_n\}$ where the r_n are observations of the underlying ground truth t_n. If the t_n are independent identically distributed random variables, the probability that the n-th observation $r_n = k$ given the corresponding ground truth sample $t_n = k'$ is

$$P_{R|T}(r_n = k \mid t_n = k') = \frac{P_{R,T}(r = k, t = k')}{P_T(t = k')} \tag{5}$$

where $P_{R,T}(r = k, t = k')$ is the joint distribution and $P_T(t = k')$ is the prior probability. Assume the K classes are equally likely to occur, i.e., $P_T(t = k') = 1/K$, and the joint distribution has the following simple form:

$$P_{R,T}(r = k, t = k') = \begin{cases} \dfrac{\rho}{K}, \ k = k' \\ \dfrac{1-\rho}{K(K-1)}, \ k \neq k' \end{cases} \tag{6}$$

for $K \geq 2$. This model assumes errors are distributed uniformly over all classes. Since

$$\rho = \sum_{k=k'} P_{R,T}(r = k, t = k'), \tag{7}$$

the observations can be viewed as ground truth with an accuracy of ρ.

Now to determine the effect of errors in ground truth on classification accuracy, we derive the accuracy of a classification relative to *R* and compare this result to the true accuracy measured against *T*. The true accuracy of classifier *A* is

$$\alpha = \sum_{k=k'} P_{A,T}(a = k, t = k') \tag{8}$$

where $P_{A,T}(a = k, t = k')$ is the joint probability (normalized confusion matrix). The accuracy of *A* relative to *R* is similarly defined

$$\gamma = \sum_{k=k'} P_{A,R}(a = k, r = k') \tag{9}$$

where

$$P_{A,R}(a = k, r = k') = \sum_{k''=1}^{K} P_{A,R|T}(a = k, r = k' \mid t = k'') P_T(t = k''). \tag{10}$$

Clearly there are correlations between *A* and *R*, since a classifier uses *R* to produce *A*; however, it can be shown that classifiers can be somewhat tolerant of errors in their training data (Carlotto 1996). We therefore provisionally assume that *A* and *R* can be treated as conditionally-independent processes, and so expand (10) as:

$$P_{A,R}(a = k, r = k') = \sum_{k''=1}^{K} P_{A|T}(a = k \mid t = k'') P_{R|T}(r = k' \mid t = k'') P_T(t = k'') \tag{11}$$

which can be rewritten as

$$P_{A,R}(a = k, r = k') = \sum_{k''=1}^{K} \frac{P_{A,T}(a = k, t = k'') P_{R,T}(r = k', t = k'')}{P_T(t = k'')} \tag{12}$$

for $P_T(t = k'') > 0$. Let us assume the same kind of simple error model for classifier *A* as was assumed for *R* above, namely

$$P_{A,T}(a = k, t = k') = \begin{cases} \dfrac{\alpha}{K}, k = k' \\ \dfrac{1-\alpha}{K(K-1)}, k \neq k' \end{cases} \tag{13}$$

In other words classification errors occur with equal frequency across all classes. $P_{A,R}(a = k, r = k')$, for $k = k'$, is equal to

$$\begin{aligned} &\sum_{k''=1}^{K} \frac{P_{A,T}(a = k, t = k'')P_{R,T}(r = k, t = k'')}{P_T(t = k'')} = \\ &\frac{P_{A,T}(a = k, t = k)P_{R,T}(r = k, t = k)}{P_T(t = k)} + \sum_{k'' \neq k} \frac{P_{A,T}(a = k, t = k'')P_{R,T}(r = k, t = k'')}{P_T(t = k'')} = \\ &\frac{\alpha\rho}{K} + \frac{(1-\alpha)(1-\rho)}{K(K-1)}. \end{aligned} \tag{14}$$

where the last step follows from the substitution of parameters from the confusion matrices. The accuracy of *A* compared to *R* is thus

$$\gamma = \sum_{k=k\copyright} P_{A,R}(a = k, r = k') = K\left(\frac{\rho\alpha}{K} + \frac{(1-\rho)(1-\alpha)}{K(K-1)}\right) = \rho\alpha + \frac{(1-\rho)(1-\alpha)}{(K-1)}. \tag{15}$$

Fig. 3 plots the relative accuracy of a classifier whose true accuracy is $\alpha = 0.8$ against a reference whose accuracy varies over the range $0 \leq \rho \leq 1$. The second term in the above equation is the probability that *A* and *R* both make the same error. Its effect is relatively small (provided α and ρ are not too small) and decreases as the number of classes increases (Fig. 4).

A simulation was performed to verify the correctness of the above model. We start with a uniform random number generator *T* which generates numbers between 1 and *K*. Observations are simulated by another independent random process *R* that alters a given fraction (1-ρ) of the random numbers generated by *T*. The classifier is modeled by a third independent random process *A* that alters a given fraction (1-α) of the random numbers generated by *T*. The relative accuracy γ is estimated from the joint probability distribution $P_{A,R}$ computed over a given number of samples N. Fig. 5 shows excellent agreement between the results of the simulation and the model for N=1000 samples.

Fig. 6 plots the relative accuracy of two classifiers A and B whose true accuracies are α and β. It is noted that at $\rho = 1/K$, the relative accuracies intersect, $\gamma_A = \gamma_B = 1/K$.

We are now prepared to address the relationship between the number of samples collected and the accuracy of the samples on the relative accuracy of a classifier. Let α^+ and α^- be the upper and lower limits of the true accuracy of classifier A at a certain level of confidence for a given number of samples. Using (15), if we assume that the accuracy of R is ρ, the upper and lower accuracies of A relative to R are defined as

$$\gamma^+ = \rho\alpha^+ + \frac{(1-\rho)(1-\alpha^+)}{(K-1)} \text{ and } \qquad \gamma^- = \rho\alpha^- + \frac{(1-\rho)(1-\alpha^-)}{(K-1)} \tag{16}$$

respectively, provided $\rho > 1/K$. In reality we do not know the accuracy of R precisely since it too is estimated from a limited number of samples. If ρ^+ and ρ^- are the upper and lower limits of the accuracy of R relative to T then the corresponding limits of the accuracy of A relative to R are (Fig. 7)

$$\gamma^+ = \rho^+\alpha^+ + \frac{(1-\rho^+)(1-\alpha^+)}{(K-1)} \text{ and } \gamma^- = \rho^-\alpha^- + \frac{(1-\rho^-)(1-\alpha^-)}{(K-1)}. \tag{17}$$

provided $\rho^- > 1/K$ and $\rho^+ > 1/K$. In other words, the uncertainty in γ depends on the uncertainty in α and ρ.

The above effect can be observed in the simulations as the number of samples decreases. Fig. 8 plots the results of the previous simulation using fewer samples (N=100). The computed accuracies tend to scatter within a wedge-shaped region around the line $\gamma = \rho\alpha + (1-\rho)(1-\alpha)/(K-1)$.

Now, to answer the question concerning the relationship between the accuracy of the ground truth and the true accuracy of a classifier, we work backwards, starting with the measured accuracy of A relative to R to infer the true accuracy of A relative to T as a function of the accuracy of R relative to T. Strictly speaking, we cannot invert the observation model to determine T and hence $P_{A,T}$ from $P_{A,R}$ and $P_{R,T}$ since it would lead to a model that is inconsistent with the one presented. Instead we extrapolate the relative accuracy γ_A computed at $\rho = \rho_0$ to estimate what the true accuracy α_{est} would be at $\rho = 1$,

$$\alpha_{\text{est}} = \frac{\gamma_A(K-1) + \rho - 1}{\rho K - 1}. \tag{18}$$

In working backwards, the uncertainty in the accuracy of ρ and γ is amplified in the uncertainty in the estimate of the true accuracy α_{est} as $\rho \rightarrow 1/K$ (compare Fig. 9a and b).

Instead of deriving the exact form of the above distribution, histograms were computed over a large number of trials for different combinations of parameter values. In each trial we generate a ground truth sequence *T*, and from it, produce two sequences. *A* sequence of observations *R* is simulated by altering (1-ρ) of the random numbers generated by *T*. This is treated as ground truth with accuracy ρ. Another sequence which is treated as the output of a classifier *A* with accuracy α is produced by altering (1-α) of the random numbers generated by *T*. For each trial we compute ρ and γ_A from the above sequences and estimate the accuracy of *A* using (18). We note that if *K* is large and neither γ_A nor ρ are too small,

$$\alpha_{\text{est}} \approx \frac{\gamma_A(K-1)}{\rho K} \approx \frac{\gamma_A}{\rho} \tag{19}$$

For a large number of samples, ρ and γ_A can be approximated by normal distributions. The ratio of normal random variables has a Cauchy density (Papoulis 1965).

Although α_{est} is not normally distributed we will use the expected value μ and the standard deviation σ to gain some insight into the behavior of the estimate under different circumstances. First, the effect of different ground truth accuracies and different sample sizes were explored. Statistics are shown in Fig. 10 for $\alpha = 0.8$, based on *N*=50, 100, and 200 samples and *K*=10 classes over the range $0.3 \le \rho_0 \le 1$. The statistics were computed over 1000 trials. As the accuracy of *R* decreases, the estimates become noisy, and the standard deviation increases. However as the number of samples increases, the standard deviation decreases significantly.

Next we compared two classifiers *A* and *B* where $\alpha = 0.8$ and $\beta = 0.7$ for *N*=50 samples and *K*=10 classes over the same range $0.3 \le \rho_0 \le 1$ (Fig. 11). Again, the estimates appear to be unbiased. The standard deviation does not appear to depend on the actual accuracy of the classifier.

Finally the effect of the number of classes are assessed for K=10 and 20 classes based on N=50 samples for $\alpha = 0.8$ over the range $0.3 \leq \rho_0 \leq 1$ (Fig. 12). The number of classes does not appear to affect the mean. In this particular example the standard deviations for K=10 and 20 are about the same until $\rho = 0.6$ at which point the standard deviation appears to increase for the smaller number of classes.

We then considered the relative performance of two classifiers A and B with true accuracies α and β. Define $\Delta\alpha = \alpha^+ - \alpha^-$, $\Delta\beta = \beta^+ - \beta^-$, and $\Delta\rho = \rho^+ - \rho^-$. As N becomes large, $\Delta\alpha$, $\Delta\beta$, and $\Delta\rho$ will approach zero. The uncertainty in the measured accuracies $\Delta\gamma_A$ and $\Delta\gamma_B$ will thus tend to zero. If $\alpha < \beta$, then $\gamma_A < \gamma_B$ provided $\rho > 1/K$. Going the other way, if we observe that $\gamma_A < \gamma_B$ it can be concluded that $\alpha < \beta$. Thus as the number of samples increases we should be able to determine if classifier B is better than A based on their measured accuracies γ_A and γ_B over a wide range of reference accuracies ρ.

Additional simulations were performed to measure the probability (relative frequency) $\gamma_A < \gamma_B$ as a function of the difference between their true accuracies $|\alpha - \beta|$, the reference accuracy ρ, the number of samples N, and the number of classes K. Let the probability that $\gamma_A < \gamma_B$ be denoted $P_{A<B}$. The results from 1000 trials are plotted in Fig. 13 for $\alpha = 0.7$, $\beta = 0.8$, K=10 classes, and N=50, 200, and 800 samples. As N increases the curve tends to a step function at $\rho = 0.1$. Next we increased the difference between α and β. Fig. 14 plots $P_{A<B}$ for $\alpha = 0.7$ and $\beta = 0.9$ as a function of the accuracy ρ based on K=10 classes and N=100 samples. As $|\alpha - \beta|$ increases the ability to differentiate between the classifiers again increases. Fig. 15 shows the effect of the number of classes K. In all plots $P_{A<B} \approx 1/2$ occurs at $1/K$. Based on these results we conjecture that as the number of samples increases, if classifier B is better than A, the probability that $\gamma_A < \gamma_B$ approaches one provided $\rho > 1/K$.

We can thus differentiate between the performance of two or more classifiers at a level of confidence that depends on the accuracy of the ground truth, the difference between their true accuracies, the number of classes, and the number of samples.

4. LANDSAT CLASSIFICATION EXPERIMENT

Often in developing algorithms for exploiting remotely sensed imagery one is interested in the relative performance of one algorithm to another. In this section we apply the above method of assessing relative classification accuracy to determine which of two Landsat classifiers over a given area is better. Earlier we stated an opinion to the effect that a classifier should be tested over a very large number of points within a scene particularly if topographic, environmental, and other factors have not been taken into account. Here we compare two classifiers against a reference classification on a pixel-by-pixel basis. Two classifications of a Landsat TM image over rural Virgina, designated *A* and *B*, are shown in Fig. 16. *A* is a maximum likelihood classification of the area and *B* is a spectral shape classification (Carlotto 1998).

Ground truth data T were obtained at N=77 points within the scene. We used this to classify a 5 m/pixel M7 image that was coregistered to the Landsat to derive a reference classification (Fig. 17). This image was treated as an image of M=33,772 observations R of the underlying ground truth T. The reference classification image R was compared to the ground truth T to produce an estimate of its accuracy ρ based on N samples. The outputs from classifiers A and B were compared to the reference classification to produce estimates of their relative accuracies γ_A and γ_B based on M samples. With our model (19) we estimated the true accuracies α_{est} and β_{est} of A and B, again based on M samples. These values were then compared to their true accuracies α and β measured against the ground truth T from N samples.

ISODATA was used to interactively cluster and classify the M7 data. Image clustering and classification were performed in two phases. First the data were divided into 30 clusters automatically; i.e., without supervision. The clusters were then classified by overlaying the clusters onto different M7 band combinations. Classes were assigned to clusters according to the classification criteria defined in Table 1. For this scene 12 of the 16 possible classes were present (thus K=12). Clusters that could not be assigned a class were further split by masking the clusters, reapplying the ISODATA algorithm, and interactively classifying the resultant sub-clusters. Any remaining sub-clusters that could not be assigned a class were not classified. The labeled clusters were then merged into a final classification image.

Table 2 summarizes the computed accuracy of the reference image *R* and classifiers *A* and *B* compared to ground truth, and the estimated true accuracies of *A* and *B* based on their measured accuracies. The estimated true accuracies α_{est} and β_{est} are within the 99% confidence intervals defined by the actual measured accuracies α and β. From the accuracies computed against the ground truth one might conclude that classifier *A* is better than *B*. However, there is considerable overlap in the ranges of *A* and *B* since the estimates were obtained over only N=77 samples.

Let $P_A(n)$ and $P_B(n)$ be the distributions of the number of correct classifications for the two classifiers. The probability of error (i.e., deciding that classifier *A* is better when *B* is actually better and vice versa) depends on the amount of overlap between the two distributions. If the distributions are approximated as normal with means $\mu_A = \alpha N$ and $\mu_B = \beta N$, and variances $\sigma_A^2 = \alpha(1-\alpha)N$ and $\sigma_B^2 = \beta(1-\beta)N$ respectively, the probability of error is

$$P_{\text{error}} = \frac{\Phi\left(\frac{n_0 - \mu_A}{\sigma_A}\right) + \left[1 - \Phi\left(\frac{n_0 - \mu_B}{\sigma_B}\right)\right]}{2} \qquad (20)$$

where n_0 is the point where the distributions intersect $P_A(n_0) = P_B(n_0)$, and

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du \qquad (21)$$

is the error function. For $\alpha = 0.69$, $\beta = 0.58$ and $N = 77$, $P_{error} \approx 0.15$.

The relative accuracies γ_A and γ_B were computed over the entire image (M=33,772 samples). For such a large number of samples, the uncertainty in γ_A and γ_B is negligable. On the other hand, the uncertainty in the actual accuracies α_{est} and β_{est} estimated from γ_A, γ_B and ρ depends on the uncertainty in ρ which is computed over N samples. However recall that for two classifiers, we were able to demonstrate that $\gamma_B > \gamma_A$ implies $\beta_{est} > \alpha_{est}$ if $\rho > 1/K$. For N=77 samples, $\rho = 0.84$, and K=12 classes, the probability that $\rho < 1/K$ is

$$\Phi\left(\frac{77\times\frac{1}{12}-77\times 0.84}{\sqrt{77\times 0.84\times 0.16}}\right)=\Phi(-18.1) \tag{22}$$

which is an extremely small number. Although there is a significant amount of uncertainty in their absolute accuracies (because of the relatively small number of samples used to estimate ρ), there is virtually no uncertainty in their relative accuracies (because of the relatively large number of samples used to estimate γ_A and γ_B). We thus conclude that classifier *B* is better than *A*.

5. DISCUSSION AND RESULTS

A different view of accuracy assessment has been presented - one that is based on trading-off accuracy for sample size. We have shown that when the ground truth is inaccurate, modeled in this paper by observations *R* of the underlying ground truth *T*, the true accuracy of a classifier *A* compared to *T* can be estimated from the measured accuracy of *A* compared to *R* and the accuracy of *R* compared to *T*. Based on a series of simulations the resultant estimates appear to be unbiased for the model used. The uncertainty in the estimated accuracy depends on the number of samples and the accuracy of the ground truth.

We have gone on to show that as the number of samples increases the ability to determine if one classifier is better than another increases even when the ground truth is inaccurate. This result has implications in efforts that are evaluating the performance of different classification schemes where it is not possible to collect accurate ground truth. Often a basis for comparison is created from coregistered aerial photographs or other sources. In such cases we have shown that it is possible to effectively use this kind of data to determine the relative performance of classifiers even when its true accuracy is in doubt.

Several areas for future work remain. Our model assumed that all classes occur with equal frequency, and that classification errors occur with equal frequency across all classes. In most scenes a few classes tend to dominate. Also, depending on the classifier, errors occur more frequently between spectrally similar classes (evergreen and mixed forests) than between spectrally dissimilar classes (water and bare soil). Different error models need to be evaluated as well as the possibility of using the confusion matrices themselves. We also need to determine if there is a relationship between classification errors and errors in ground truth (i.e., is the conditional independence assumption valid?). Ultimately, a better understanding of the sources of error in ground truth and their effect on classification accuracy is needed.

References

Congalton, R.G., "A comparison of sampling schemes used in generating error matrices for assessing the accuracy of maps generated from remotely sensed data," *Photogrammetric Engineering and Remote Sensing*, Vol. 54, No. 5, May 1988, pp 593-600.

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

Janssen, L.L.V., and van der Wel, F.J.M., "Accuracy assessment of satellite derived land-cover data," *Photogrammetric Engineering and Remote Sensing*, Vol. 60, No. 54 April 1994, pp 419-424.

Papoulis, A. *Probability, Random Variables, and Stochastic Processes*, McGraw Hill, 1965, New York, pp 197-198.

Todd, W.J., Gehring, D.G., and Haman, J.F., "Landsat wildland mapping accuracy," *Photogrammetric Engineering and Remote Sensing*, Vol. 46, No. 4, April 1980, pp 509-520.

Carlotto, M.J., "Using maps to automate the classification of remotely sensed imagery," *Proceedings SPIE*,, Vol. 2758, Orlando, Florida, 1996.

Carlotto, M.J., "Spectral Shape Classification of Landsat Thematic Mapper Imagery," *Photogrammetric Engineering and Remote Sensing*, Vol. 64, No. 9, September 1998.

Varotsos, C., “Power-law correlations in column ozone over Antarctica,” *International Journal of Remote* Sensin, Vol. 26, No 16, pp 3333-3342, 2005.

Varotsos, C, and Kirk-Davidoff, D., “Long-memory processes in ozone and temperature variations at the region 60 degrees S — 60 degrees N,” *Atmospheric Chemistry and Physics*, Vol. 6, pp 4093-4100, 2006.

Varotsos C., Ondov, J. and Efstathiou M., “Scaling properties of air pollution in Athens, “Greece and Baltimore, Maryland,” *Atmospheric Environment*, Vol. 39, pp 4041-4047, 2005.

Table 1 Classes used in Landsat accuracy assessment.

Level 1 Class	Criteria	Level 2 Class	Criteria
Developed	> 50% synthetic	High Intensity	> 80% synthetic
		Low Intensity	50-80% synthetic
		Roads	
Herbaceous Land	> 50% herbaceous	Crops	managed
		Pasture	unmanaged
		Other	
Woody	> 50% woody	Deciduous	> 67% deciduous
		Evergreen	> 67% evergreen
		Mixed	
Barren	< 50% vegetated		
Wetland		Shore	< 50% vegetated
		Emergent	> 50% herbaceous
		Woody	> 50% woody
Water			
Snow/Ice			
Other/Indeterminate			

Table 2 Summary of results from Landsat classification experiment.

Data	Accuracy Measured Against Ground Truth (Based on 77 samples)		Relative Accuracy (Based on 33772 samples)	Estimated Accuracy
Reference Classification, *R*	$\rho = 0.84$	$\rho- = 0.73$ $\rho+ = 0.92$		
Classifier *A*	$\alpha = 0.69$	$\alpha- = 0.58$ $\alpha+ = 0.79$	$\gamma_A = 0.5$	$\alpha_{est} = 0.59$
Classifier *B*	$\beta = 0.58$	$\beta- = 0.47$ $\beta+ = 0.69$	$\gamma_B = 0.56$	$\beta_{est} = 0.66$

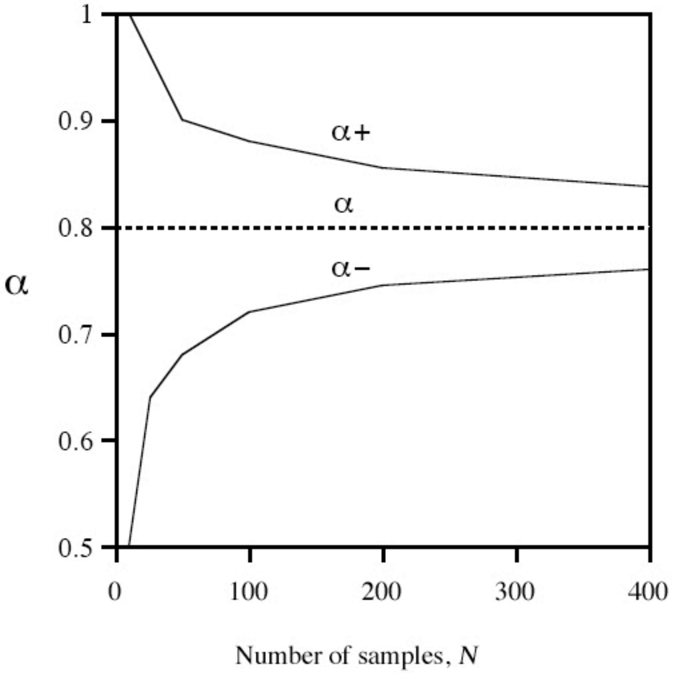


Fig. 1 κ = 0.95 confidence interval for α = 0.8 as a function of the number of samples, *N*.

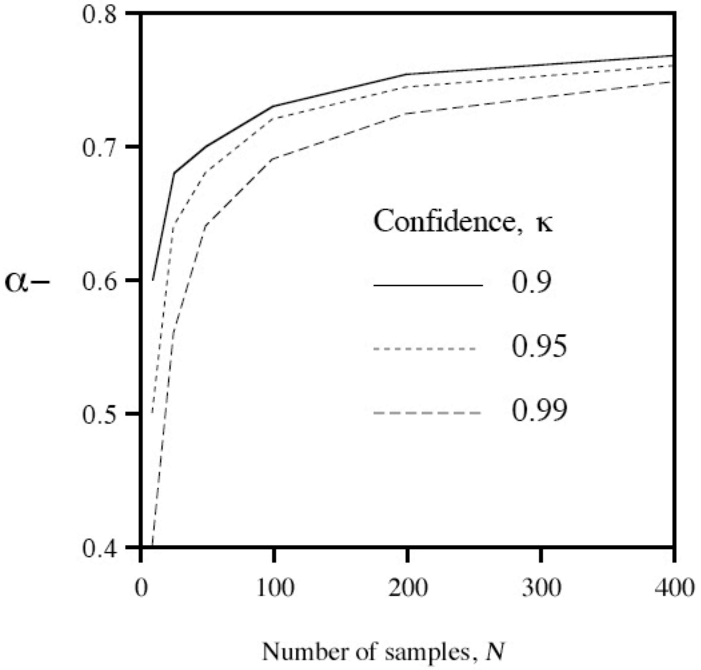


Fig. 2 Lower limits of classification accuracy α− at three confidence levels for α = 0.8.

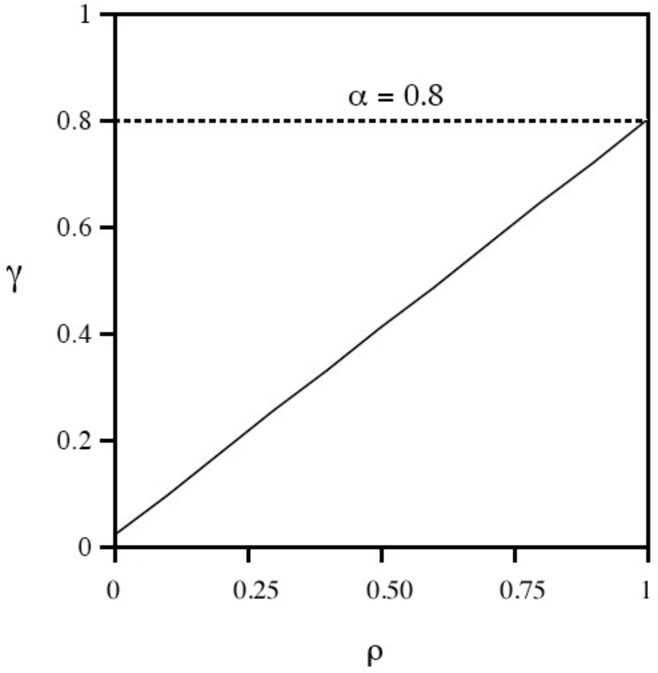


Fig. 3 Relative accuracy γ of a classifier with true accuracy α = 0.8 for *K*=10 classes as a function of the ground truth accuracy ρ.

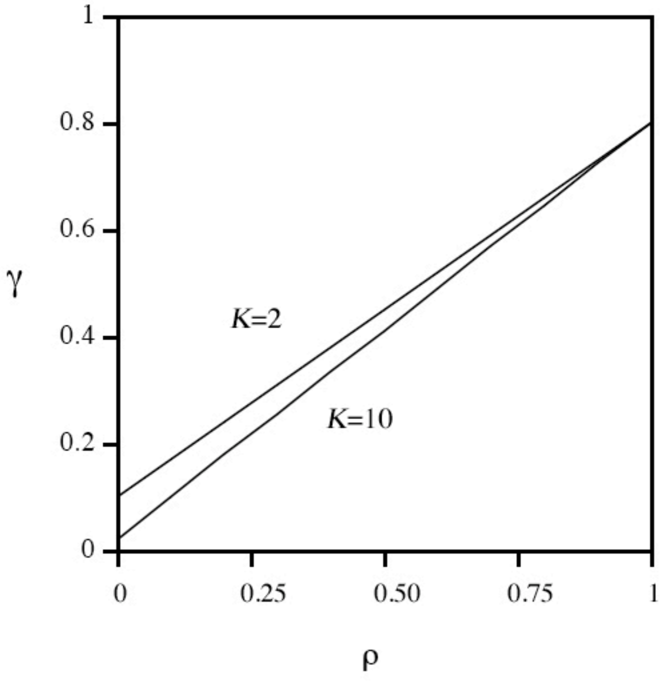


Fig. 4 Effect of the number of classes *K* on the relative accuracy γ of a classifier with true accuracy α = 0.8 as function of the ground truth accuracy ρ.

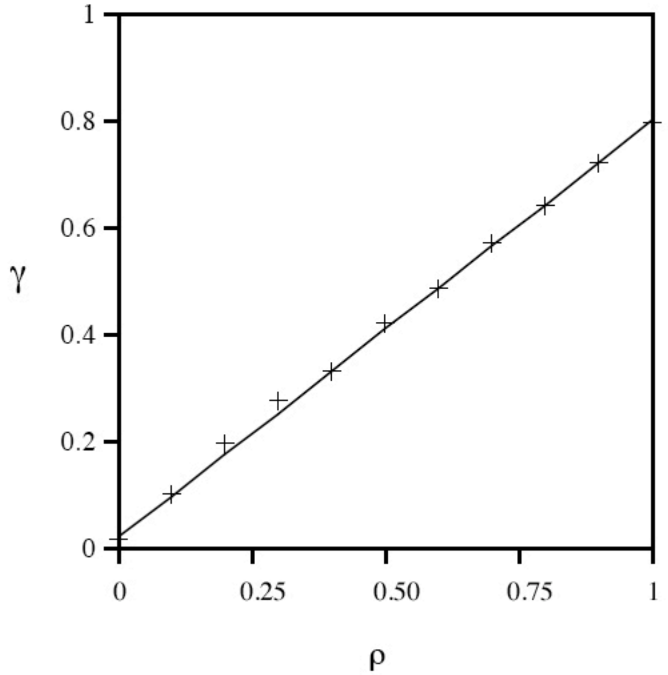


Fig. 5 Simulation results compared to model predictions for α = 0.8, *K*=10, and *N*=1000.

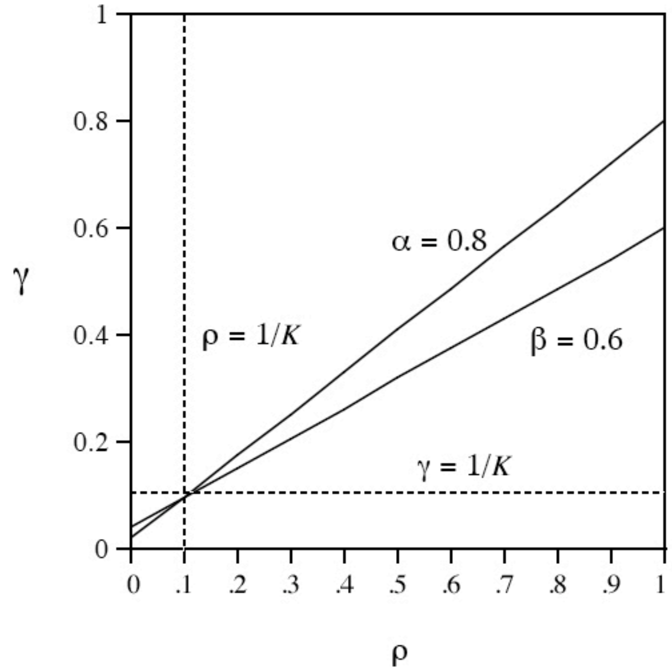


Fig. 6 Relative accuracies of two classifiers having true accuracies α = 0.8 and β = 0.6.

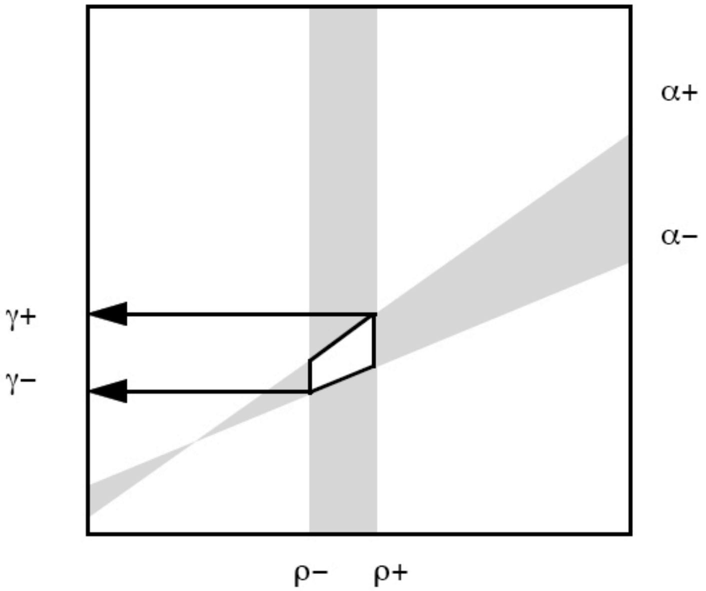


Fig. 7 Uncertainty in estimated classification accuracy depends on the uncertainty in the true accuracy of the classifier and the accuracy of the ground truth.

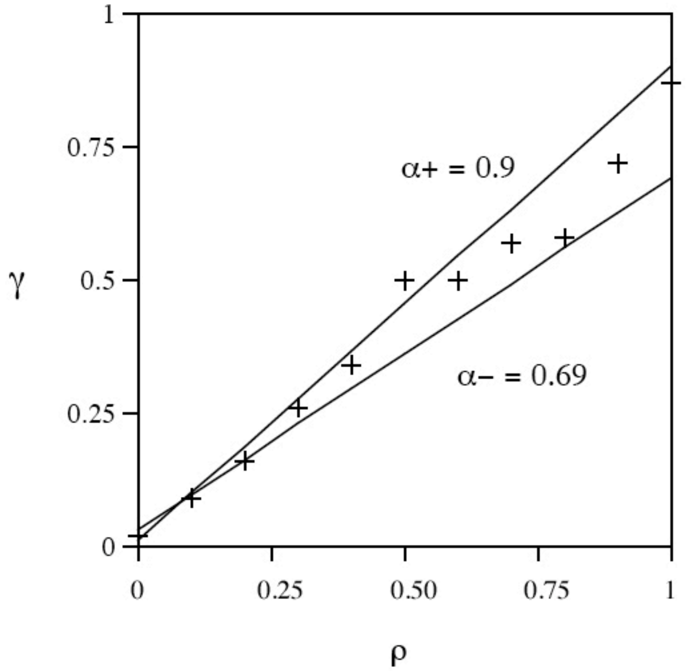
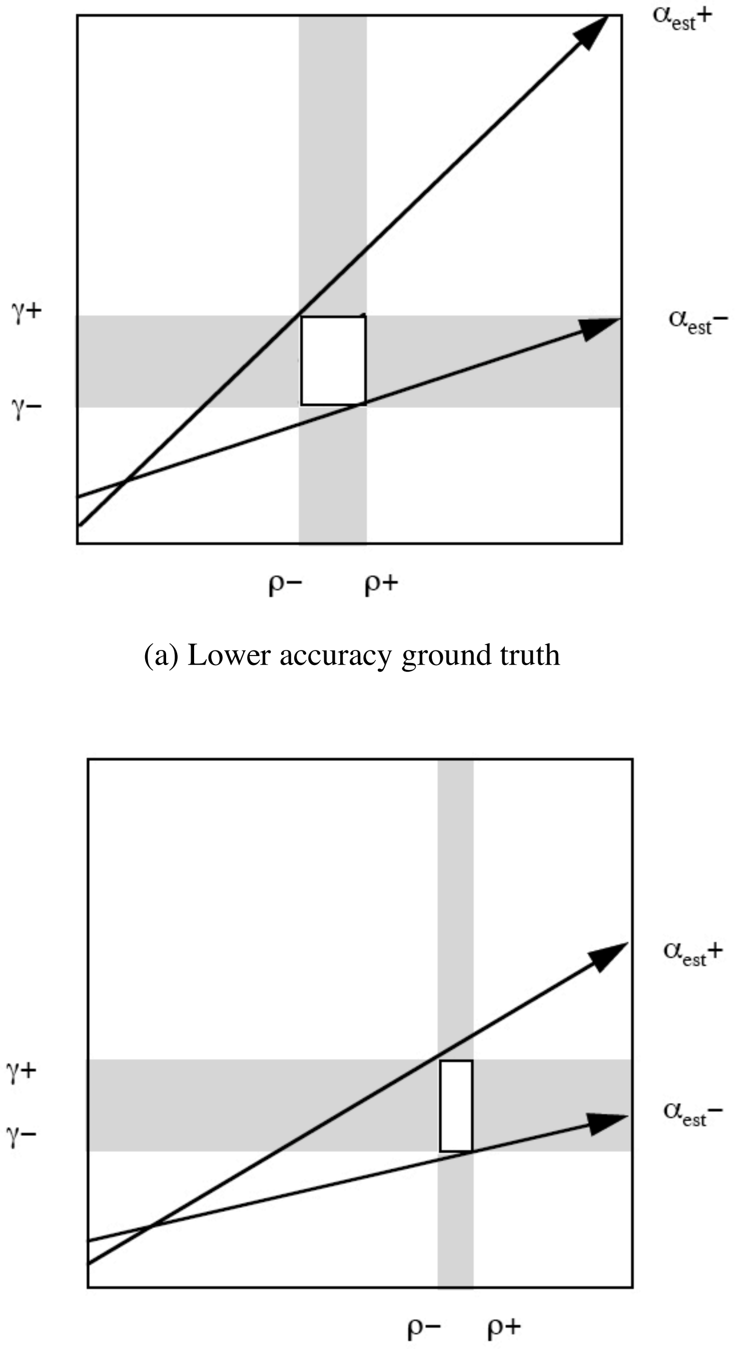


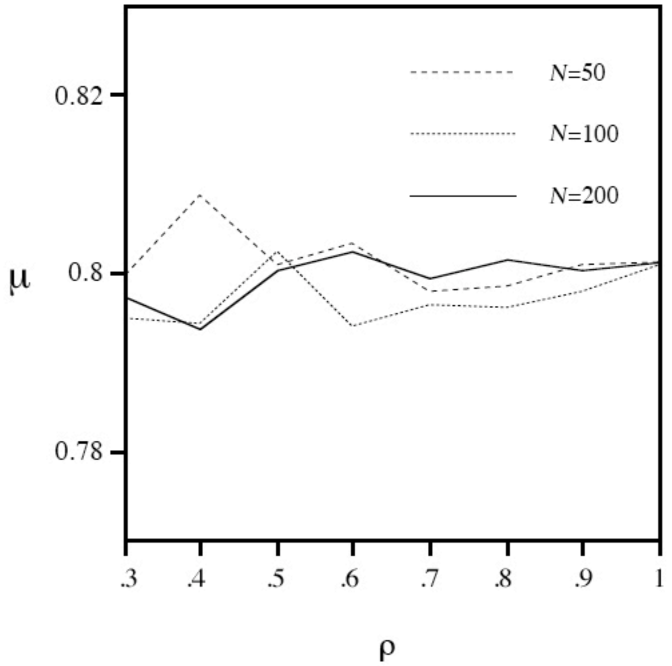
Fig. 8 Simulation results for α = 0.8, *K*=10, and *N*=100. α+ and α- limits correspond to κ=0.99 confidence interval.



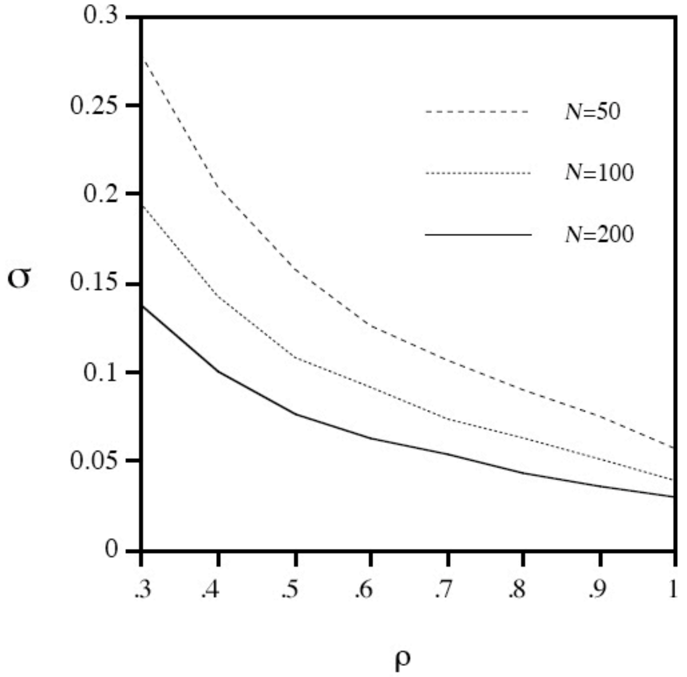
(a) Lower accuracy ground truth

(b) Higher accuracy ground truth

Fig. 9 Estimating the true accuracy of a classifier from its measured accuracy becomes easier as the accuracy of the ground truth increases.

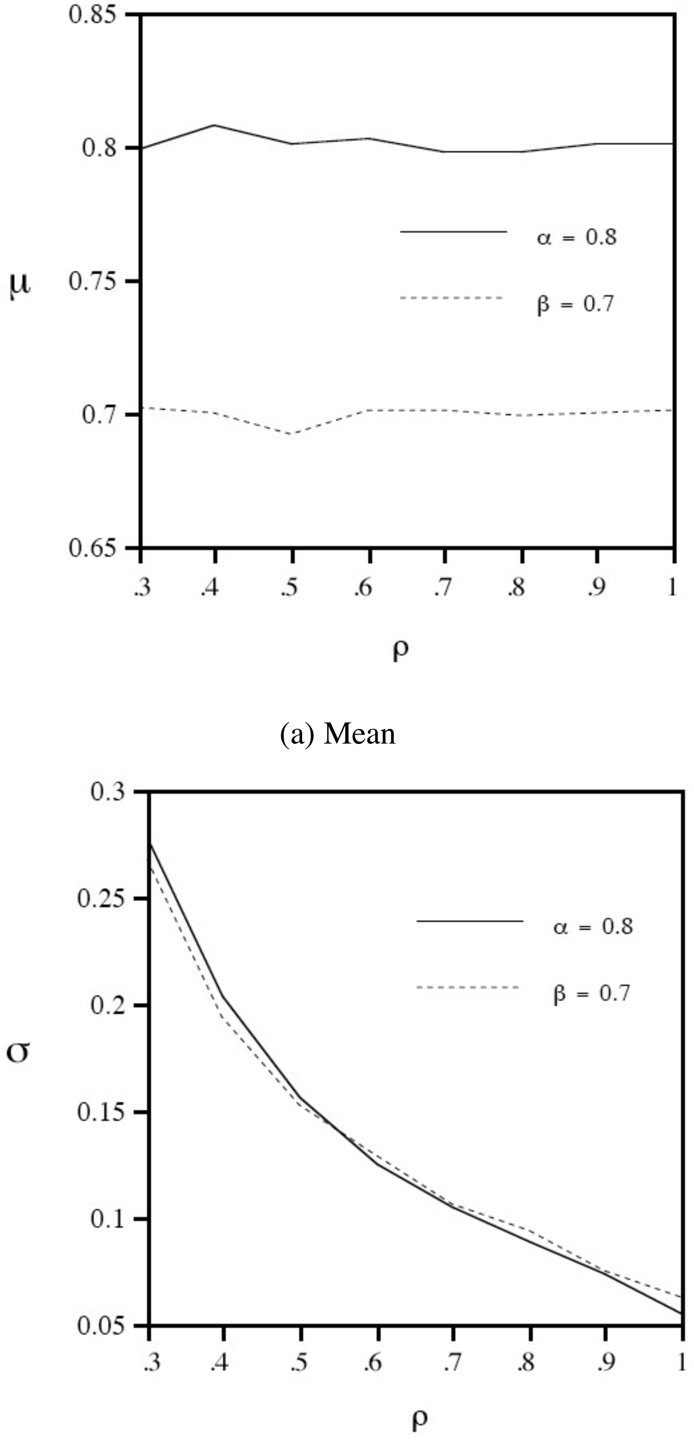


(a) Mean



(b) Standard deviation

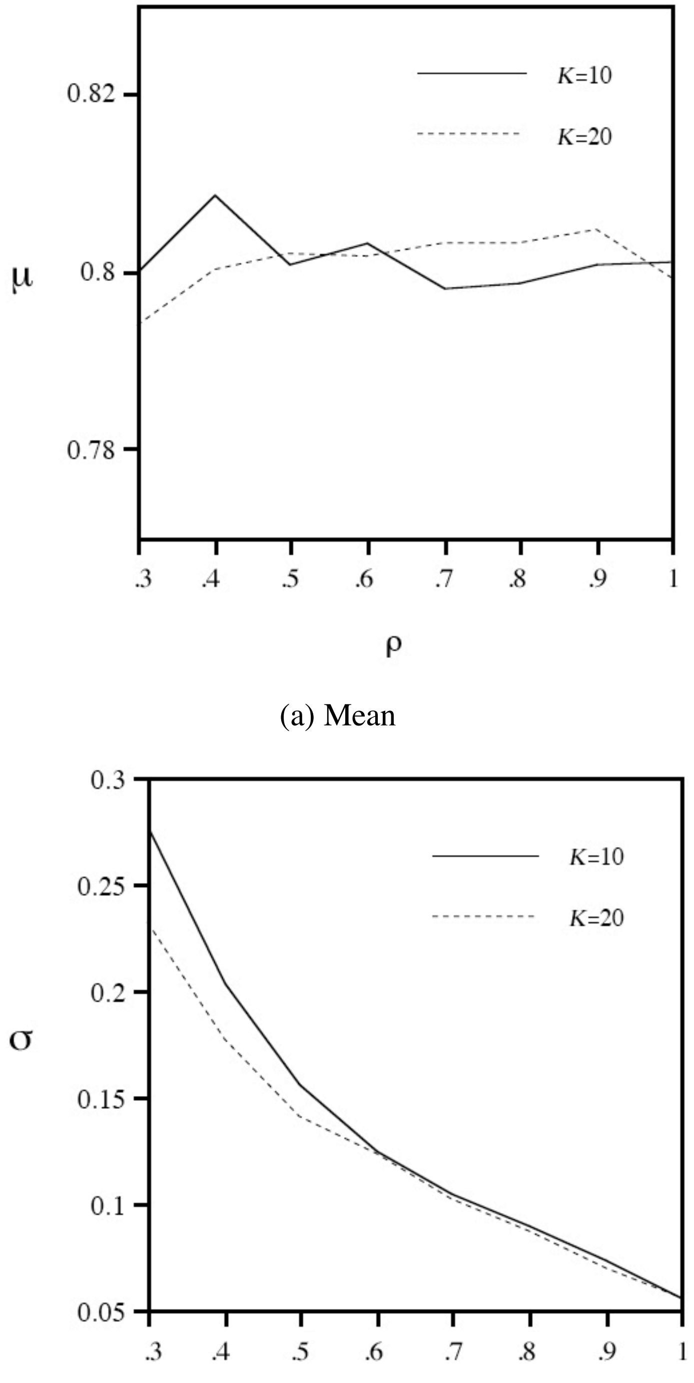
Fig. 10 Mean μ and standard deviation σ of α_{est} as a function of ρ for different sample sizes *N*.



(a) Mean

(b) Standard deviation

Fig. 11 Mean μ and standard deviation σ of α_{est} as a function of ρ for two classifiers ($\alpha = 0.8$, $\beta = 0.7$).



(a) Mean

(b) Standard deviation

Fig. 12 Mean μ and standard deviation σ of α_{est} as a function of ρ for different numbers of classes *K*.

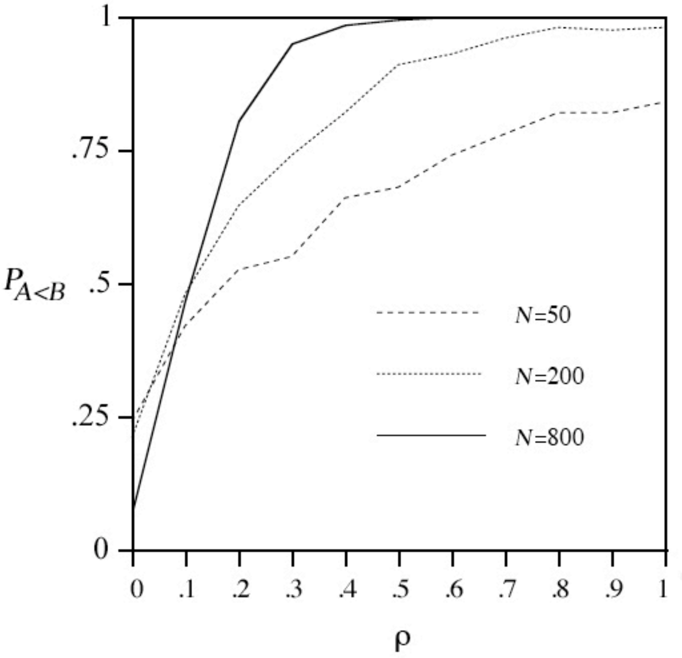


Fig. 13 $P_{A<B}$ as a function of ρ of a pair of classifiers (α = 0.7, β = 0.8) for different sample sizes.

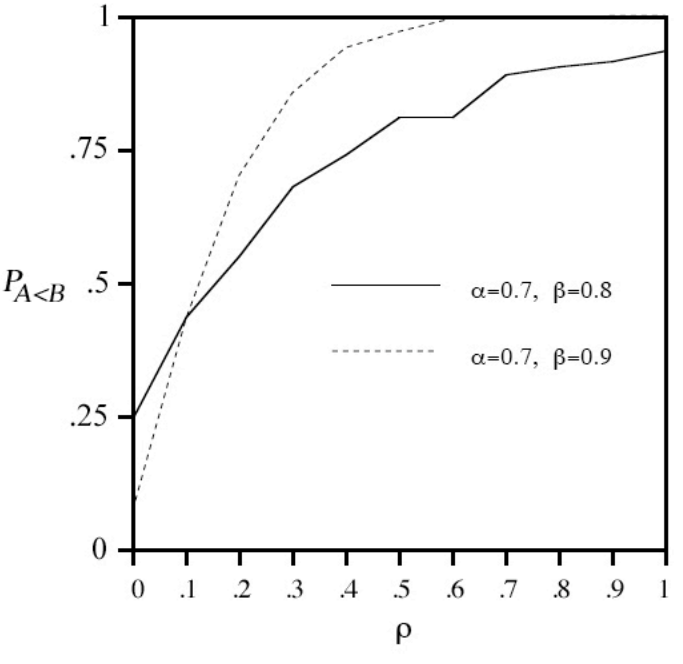


Fig. 14 $P_{A<B}$ as a function of ρ for two different pairs of classifiers (*N*=100 samples).

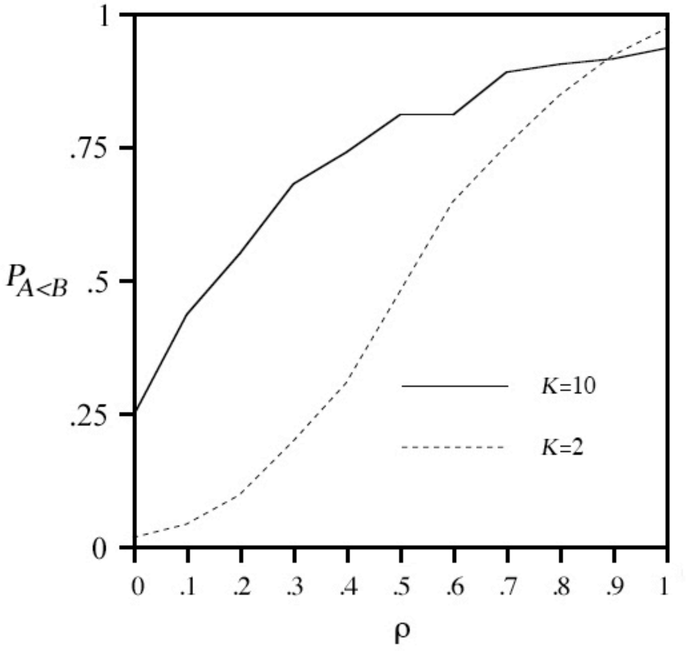
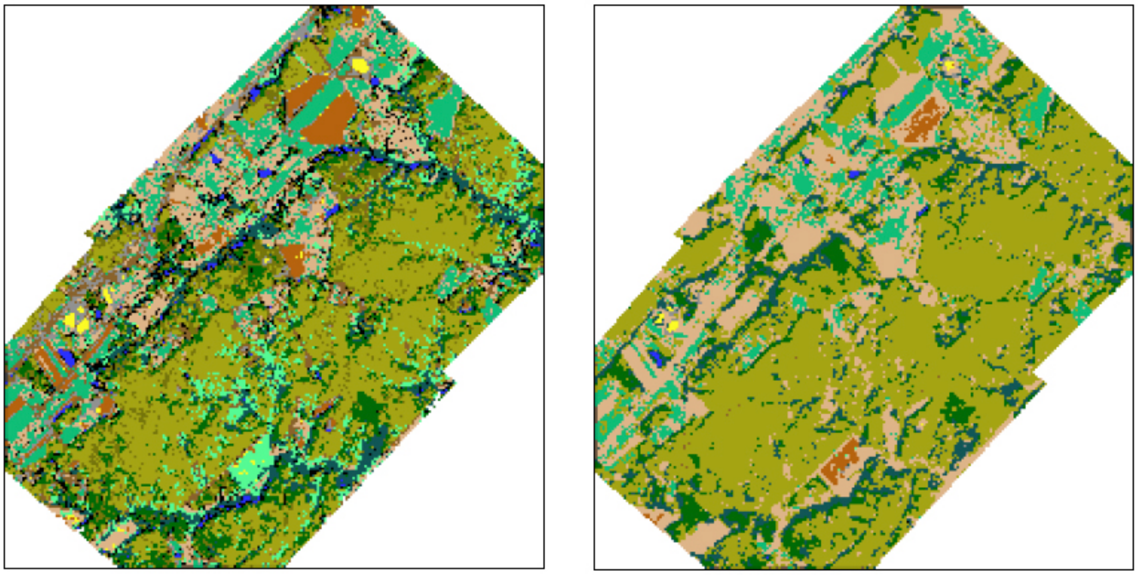


Fig. 15 $P_{A<B}$ as a function of ρ for different numbers of classes.



Classifier *A* output

Classifier *B* output

Fig. 16 Outputs from two classifiers: *A* was produced by a maximum likelihood classifier and *B* from a spectral shape classifier.

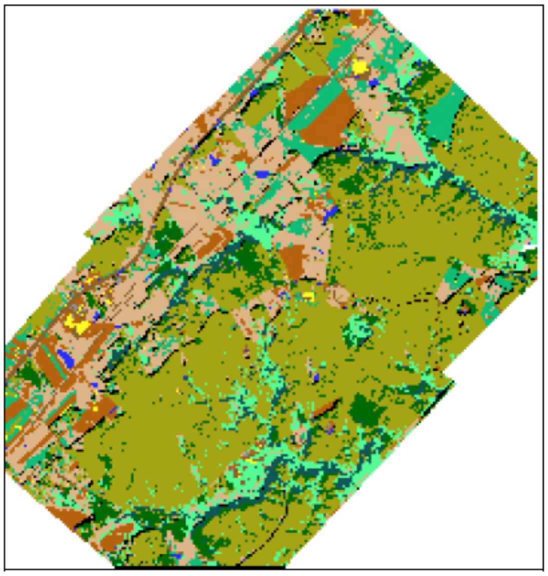


Fig. 17 Reference classification image R derived from M7 imagery using interactive clustering and labelling. Classes defined in Table 1 were assigned to clusters using 5 m M7 data as image truth.