I. L. Thomas*
G. McK. Allcock
Physics and Engineering Laboratory
Department of Scientific and Industrial Research
Lower Hutt, New Zealand

Determining the Confidence Level for a Classification

The emphasis is on the use of a maximum likelihood classification system, but the principle may be extended to apply to all classification approaches, be they parametric or non-parametric.

Introduction

During the testing of a maximum-likelihood classifier for land-use classification in New Zealand, the need arose for an account of the application of statistical confidence level assessments in remote sensing. Such an account was necessary to assist the discipline oriented users in evaluating their classifications.

The emphasis here is on the use of a maximum likelihood classification system, but the principle behind deriving a meaningful confidence level may be extended to apply to all classification approaches, be they parametric or non-parametric.

Abstract: The allocation of a confidence level to a classification product is considered to be essential. The acquisition of site-specific data to check the classification is discussed. A statistical approach to the determination of an appropriate confidence level from these check data is presented. Allowance for human assessment and counting errors is included. The approach is directed towards the discipline oriented user of remote sensing data and is illustrated with actual test data.

Classification Methodology

A classification is regarded here as consisting of the following components:

- the acquisition of data;
- a decision by the user as to the level of class separability that is desired and can be attained, being mindful of the spatial averaging of ground cover classes produced by the data acquisition sampling system;
- the selection of training areas which will suit a given computer-based classification software package;
- the effective operation of the package;
- the selection of an appropriate threshold for each class to apply to the likelihood distribution for that class; and
- the creation of an appropriate output product after the thresholds have been applied.

Here we consider Landsat to be the data acquisition system and the IBM Earth Resources MANagement package (ERMAN) as the analysis software (IBM, 1976). This software uses a maximum likelihood classifier.

A data acquisition system produces a set of numbers for each spatial resolution element, or "pixel," on the assumption that each element is homogeneous. However, generally the ground cover within a given element will in fact be heterogeneous to some degree. The analyst must, therefore, decide whether the assumption of homogeneity is acceptable in each case.

Analyst interaction with a software package inserts appropriate training-area characteristics into the classification process. Effective operation of the package requires adequate expertise on the part of the operator as well as accurate software, sufficient mathematical precision, and appropriate delineation of class boundaries within the training data.

Each pixel is classified, using ERMAN, into the

* Presently with General Technology Systems Limited, Brentford, Middlesex TW8 8EQ, United Kingdom.

Photogrammetric Engineering and Remote Sensing,
Vol. 50, No. 10, October 1984, pp. 1491-1496.

0099-1112/84/5010-1491$02.25/0

most likely class type and has an appropriate likelihood associated with it (Swain and Davis, 1978). Obviously, if the number of classes chosen for the classification process does not include all the classes of the area being classified, some pixels will be given an incorrect class association although, hopefully, with a low likelihood. Each class can, however, have associated with it a minimum threshold above which pixels may confidently be expected to be members of that class. The assignment of a specific minimum threshold to each class is a decision that must be made by the analyst.

The creation of an appropriate output product can also include a decision by the analyst. If the computer-produced data are further processed or interpreted to prepare a land-use map, then allowance must be made for the impact of further human decisions.

Consequently, the success of a classification can be influenced by a variety of factors—sensor, software, and human. The derivation of a confidence level for a classification must recognize that it represents such a combination of influences.

Why Have a Confidence Level?

Any computer-derived classification that will lead ultimately to a ground-cover thematic map is based on ground truth data gathered by the user from selected "training" areas. This applies whether unsupervised clustering or supervised classification is employed and whether parametric or non-parametric techniques are used.

The computer may represent classes with similar ground cover by character symbols on a line printer, colored picture elements (pixels) on a television monitor screen, or numbers on computer tape for subsequent transcription to positive transparency film.

The accuracy of the thematic map depends on our ability to extrapolate successfully from the training areas to the whole mapped area. Unless we have some statistical measure of the efficiency of the extrapolation process, we cannot estimate our level of confidence in the classification. Once a confidence level is so quantified, then a user of the classification data can relate it, by means of the probability of correct classification, to actuality over the whole classified area. The classification is thus married to ground actuality by the confidence level. (Here we are presently using the term "ground actuality" to distinguish, and stress, the difference between the set of check data used to evaluate the classification product, from the "ground truth" data used to set up the statistics and, hence, form the classification. The two sets of data must obviously be separate but must equally have the same characteristics of location, type, height, health, etc. Often the ground truth data are from a well controlled and known area whereas the ground actuality data are taken over a wider area, over less pure ground cover pixels, and employ essentially random sampling. Thus, the ground actuality data more closely represent what is actually covering the ground whereas the ground truth data are usually aimed at the purest classes to affect best class separations in the classification process.)

What Is a Confidence Level?

A pixel classified into a particular class can only be either correctly or incorrectly classified. There is no middle ground. That is, if the probability of correct classification of a pixel belonging to a given class is p, and the probability of incorrect classification is q, then

$$p + q = 1. \tag{1}$$

In this case, for one pixel taken at random from the complete set of pixels belonging to the class, the probability P that the pixel is correctly classified is $P(1) = p$. Similarly, the probability of being incorrect is $P(0) = q$.

For a two-pixel sample from the complete set there are four possible combinations, where R indicates the classification has been shown to be correct and W indicates an incorrect result:

$$RR,\ WR,\ RW, \text{ and } WW.$$

Here, the probability of being correct twice is $P(2) = p^2$; the probability of being incorrect twice is $P(0) = q^2$; and the probability of having one correct and one incorrect is $P(1) = 2pq$ (we are not concerned with sequential ordering). Similarly, for a three-pixel sample we could have the combinations

$$RRR,\ WRR,\ RWR,\ RRW,$$
$$WWR,\ RWW,\ WRW, \text{ and } WWW.$$

The probabilities then would be

$$P(3) = p^3,\ P(2) = 3qp^2,$$
$$P(1) = 3q^2p,\ P(0) = q^3.$$

Translating to numerical probabilities, if p = 3/4 and q = 1/4, then

$$P(3) = 27/64,\ P(2) = 27/64,$$
$$P(1) = 9/64,\ P(0) = 1/64.$$

The development of a probability distribution can be noted in the above examples, where the abscissa represents a stipulated number i of correctly classified pixels in a sample of n pixels, and the ordinate represents the probability that the number of correctly classified pixels is found upon examination to be exactly equal to i.

Three other points also emerge:

(a) The probabilities to be associated with 0, 1, 2, . . . , n correctly classified pixels from a sample of n pixels drawn from the complete set may be given by the terms in the binomial expansion of $(p + q)^n$.

(b) As $p + q = 1$, then $(p + q)^n = 1$;
hence $P(n) + \ldots + P(3) + P(2) + P(1) + P(0) = 1$.

(c) The coefficient for $P(i)$ is given by the Binomial Coefficient C_i^n

where
$$C_i^n = \frac{n!}{i!(n - i)!} \tag{2}$$

Consequently, we may represent the probability of finding exactly i pixels correctly classified in an n pixel sample as $P(i)$ where

$$P(i) = C_i^n \, p^i \, q^{n-i} \tag{3}$$

This obviously leads to a distribution of $P(i)$ against i. It is known as the Binomial Distribution. The mean (m) and standard deviation (s) for the Binomial Distribution* are (Moroney, 1956, p. 124)

$$\begin{aligned} m &= np \\ s &= \sqrt{npq} \end{aligned} \tag{4}$$

Usually we are concerned, when checking the efficiency of a classification, with the summation of the probabilities for all stipulated pixels between n and a lower bound, say i. That is, we wish to know the probability that at least i pixels are correctly classified, when a random sample of n pixels is selected.

This probability is called the Confidence Level (CL) for that classification, and is usually expressed as a percentage. Thus, if CL is the integrated probability expressed as a percentage, we can say that we are CL percent confident that the pixels are classified correctly at least i times out of n (or at least $(100i/n)$ percent of the time).

The mathematical evaluation of CL from the Binomial Distribution rapidly becomes tedious. A more convenient approach is sought.

Mood (1950, p. 139) demonstrates that, as the sample size n becomes larger, the discrete Binomial Distribution approaches the continuous Normal Distribution as the limiting case for n tending towards infinity. If the *total population* has both a finite mean and standard deviation, then the *sample* mean and standard deviation may be described by Equation 4, again for an increasing sample size (Mood, 1950; Moroney, 1956). (This is based on the Central Limit Theorem and applies without reference to the form of the population distribution function, provided large samples are involved. By "large", a sample of 50 should be regarded as a minimum (Unthank, 1960) with a sample in the "hundreds" (Mood (1950) suggests 300) being more acceptable.) These conditions would be met by the practical classification tasks we are addressing here.

Consequently, under these conditions, we use the more mathematically tractable Normal Distribution. This is especially useful because when the total area under the curve is normalized to 1.0, the probability we seek is the integrated area between the limits appropriate to n and i. The equation for this unit-area Normal Distribution is

$$\text{Probability Density} = \frac{1}{s\sqrt{2\pi}} \exp\left[-(i - m)^2/2s^2\right] \tag{5}$$

(from Moroney, 1956, p. 117)

Van Genderen *et al.* (1978) show that the number of samples necessary to support the achievement of the desired confidence level in the classification product is a function of that required level. For example, for the attainment of the Anderson *et al.* (1972) suggested level of 90 percent confidence in a classification, Van Genderen *et al.* (1978) conclude that 30 *randomly distributed* samples are necessary, *as a minimum*, to support such an assessment. This is discussed further by Rosenfield *et al.* (1982). (Compare back to the sample sizes felt necessary to permit the Binomial Distribution to be replaced by the Normal Distribution.)

The task is now to redefine the Confidence Level (CL) in terms of Equation 5. Under such a curve the integrated area from three standard deviations below the mean ($m - 3s$) to plus infinity is 0.999. Thus, if we wish to have 99.9 percent confidence in our evaluation of the performance of the classifier, then the lower bound to the number of pixels that must be correctly classified in the check sample is equal to the mean minus three standard deviations.

Ground Actuality Checking of the Classification

Obviously, it is impossible to check every pixel of a classified area. By taking a suitably selected sample of pixels, representative of all conditions of vegetation/soil/climate, etc., that exist over the area, statistical techniques can then lead to a representative confidence level for the classification. Van Genderen *et al.* (1978) outline factors that should be borne in mind when designing any sampling program. They further indicate a simple and acceptable method for establishing a network of sampling sites to support the checking of the required number of samples for the desired level of classification confidence. (However, they do point out that limitations to access may intrude upon the physical implementation of such a sampling program. This, as indicated later, did modify the sam-

* Strictly speaking, the Binomial Distribution requires that, each time we examine whether or not a randomly selected pixel is correctly classified, we should immediately replace the pixel, so that we are always selecting from the complete set (consequently, with a finite chance that the same pixel is selected more than once). In practice, we sample *without* replacement, in which case the Hypergeometric Distribution should be used (Aitken, 1942, pp. 56-58). However, provided that we are dealing with large sample sizes, the properties of the two types of distribution can be assumed to be identical.

pling program used to acquire the test data reported here.)

There is no substitute for field checking the classified dataset against the actual conditions that prevail pixel by pixel on the ground. This is known as the site-specific approach (Mead and Szajgin, 1982).

An alternative, that of checking other classifications or prepared maps, involves another set of human decisions in the process and can only degrade the checking process. Another alternative is to check a multispectral classification using panchromatic air photographs. This also reduces the amount of reliable information that can be applied to the checking process.

Landsat, or any such sampling system, inevitably impresses a sampling grid over the varying ground cover. Allowance must be made for the positioning of this spatial sampling grid when checking the classification against the actual ground cover. The representative ground-cover class for each pixel, or sampling unit, must be determined and used. If the class resolution so imposed is not detailed enough, then a different sampling system, for example an aircraft scanner, should be employed.

The approach used by the New Zealand group is to take site-specific ground truth, distributed over a wide area, by actual on-site inspection. This covers the geographic extent of the classification and includes representative data on different soil types, microclimate, differing cropping cycles, etc. The classification is then set up, in a supervised manner, by using *part* of the ground truth to provide the training areas. The *remainder* of the ground truth can then be used to check the classification accuracy outside of the training areas. A variation of this technique is also used for those areas that have long-lived ground cover. Here, the classification result is taken into the field, in lineprinter format, and individual pixels are checked, and marked off, for accuracy of classification by on-site comparison. The lineprinter product is ideal for this application as it more easily permits pixel location and recording than do the photographic products.

Returning to the degradation in spatial resolution occasioned by the sampling technique: It is obvious that allowance must be made for this in checking a classification product. Field checking must, therefore, be restricted to those areas separated from the road edge or similar clearly non-homogeneous ground cover classes by at least one and preferably two pixels.

If an influence from soil or microclimate is suspected, a subdivision of the check statistics into appropriate soil/microclimate regions is necessary. The computation of individual confidence levels for each class within each of these regions and a comparison of the results then aids the assessment of the level of influence of these factors.

The sampling must also be as representative as possible of the whole classified area. A random distribution of such sampling points over the whole region must be striven for (Van Genderen *et al.*, 1978).

The above were the ground rules used by the New Zealand group when checking their ERMAN computer classification results.

Determination of a Confidence Level

If

N is the number of samples taken,
P is the number of samples that have been correctly classified,
Q is the number of samples that have been incorrectly classified,
m is the (estimated) mean of the distribution,
s is the (estimated) standard deviation of the distribution,
e_m is the standard error of the estimate of the mean,
e_s is the standard error of the estimate of the standard deviation,
e_p is the experimental (human) error in assessing and counting the number of samples that are correctly classified,

and

$$p = P/N \tag{6}$$

$$q = Q/N \tag{7}$$

then

$p + q = 1$ (1) (from previous discussion)

$m = Np$ (4)

$s = \sqrt{Npq}$ (4)

$$e_m = \frac{s}{\sqrt{N}} \quad (8) \text{ (from Moroney, 1956, p. 137)}$$

$$e_s = \frac{s}{\sqrt{2N}} \quad (9) \text{ (from Moroney, 1956, p. 137)}$$

It is assumed that p is greater than 0.1 and that N is greater than 50, so that the Binomial Distribution may be adequately represented by the unit-area Normal Distribution (Moroney, 1956, p. 128).

The error e_p is regarded purely as a human assessment and counting error. "Assessment," in the sense that a field check of a microscopically *heterogeneous* ground-cover pixel must produce a dominant class which is regarded as describing that pixel *homogeneously*. This is a human decision. Similarly, counting techniques will have a human error associated with them. The "assessment" error is minimized by having the same person who set up the ground-truth files, trained the computer classification software, and selected the thresholds, also doing the ground checking over the whole area. An accompanying impartial observer can also be used

to assist in resolving the "yes/no" status of any dubious pixel classifications during the checking process. Errors in ground-cover class interpretation can thus be reduced. This was done by the New Zealand group. Consequently, it was felt that the "assessment" error would be absorbed into the overall classification error, under these conditions. The counting inaccuracy in a test case involving some 25,000 pixels was found to be less than 0.5 percent. This was determined by repeated checks of the same data by different analysts with at least two sets of counts per analyst. e_p was then taken (another human decision) to be 0.5 percent.

The Normal Distribution, normalized to unit area, allows us to determine the number of pixels that would need to be correctly classified to maintain a Confidence Level of 99.9 percent. This is the equivalent of determining the lower acceptable limit for a number of correctly classified pixels as being at the mean of the population distribution minus three standard deviations.

In practice, values for both the mean and the standard deviation are obtained from a restricted (though possibly large) sample drawn from the whole population, and thus are *estimates*, whose standard errors are given by Equations 8 and 9. These equations indicate that e_m and e_s differ from s by numerical factors only, indicating complete correlation among the quantities. Thus, the value of the lower limit, to give a 99.9 percent confidence level, is obtained by taking a value for the mean which is three standard errors *lower* than the estimated mean, and then subtracting three times a standard deviation which is three standard errors *greater* than the estimated standard deviation. That is,

$$99.9\%\ CL = (m - 3e_m) - 3(s + 3e_s). \quad (10)$$

Examination of Equations 4, 8, and 9 will readily show that when m and N are very large, as in the example below, the standard errors are trivial and can be neglected, in which case

$$99.9\%\ CL = m - 3s. \quad (10a)$$

Nevertheless, for the sake of completeness, we have included e_m and e_s in the following illustrative calculations.

As an example of the above approach, we take an actual case of classifying 145.4 km × 117.1 km (2 661 552 Landsat pixels) of the King Country, North Island, New Zealand using the ERMAN package (Benning, 1982).

Because the King Country has highly dissected topography, it was not possible to access, on the ground, such a random sampling network as suggested by Van Genderen *et al.* (1978). Consequently, the site-specific field checking was conducted by driving over most of the road network that existed in the classified land-cover area. The pixels were evaluated at least one pixel away from the road edge and along the adjacent ridge lines. The road network spanned the complete area classified and was believed to thus fulfill reasonably well the random sampling criterion.

25 773 pixels were field checked ($=N$) (0.97 percent of the total area),
24 587 were found to be correctly classified ($=P$), and
1 186 were found to be incorrectly classified ($=Q$).

In this example, we take the complete classification, not by class, and assess an overall probability for the full classification.

From $N = 25\ 773$
$P = 24\ 587$
$Q = 1\ 186$

then $p = 0.9540$
$q = 0.0460$
$m = 24\ 587$
$s = 33.637$
$e_m = 0.210$
$e_s = 0.148$

From equation 10, the lower acceptable limit to give a 99.9 percent confidence level is

$$(m - 3e_m) - 3(s + 3e_s) = 24\ 484 = 95.00 \text{ percent of the sample.}$$

However, the above result does not take account of the counting error e_p, which we have earlier set at 0.5 percent. Thus, 129 pixels (0.5 percent of 25 773) may have been miscounted. To maintain our 99.9 percent confidence in the result, we must therefore reduce the lower acceptable limit by 129; i.e., to 24 355 = 94.50 percent of the sample.

We conclude, with 99.9 percent confidence, that at least 94.50 percent of the pixels in the *whole* area have been correctly classified. That is, if 1000 random samples, each of about 25 000 pixels, were taken from the whole 2.66 million pixels being classified, in only one case would we expect to find a classification accuracy of less than 94.50 percent.

Lesser degrees of confidence may be acceptable in some applications. For instance, if a confidence level of 99 percent was required, all the "threes" in Equation 10 would be replaced by 2.33. To achieve 95 percent confidence, the "threes" in the equation would be replaced by 1.65. Applied to the present example, after taking account of e_p as indicated above, the following results are obtained.

We are
99.9 percent confident that at least 94.50 percent
99 percent confident that at least 94.59 percent
95 percent confident that at least 94.68 percent

of the pixels in the whole area have been correctly classified. The very small spread in these figures is

a direct result of the very large size of the sample taken, which leads to a strongly peaked distribution with a small standard deviation—note that '*s*' is only about 0.1 percent of '*m*'.

As already stated, the above probability of at least 94.50 percent correct classification for a confidence level of 99.9 percent pertains to the full multi-class classification. A similar evaluation should be undertaken to assess applicable confidence levels for individual classes.

Summary

The concept of a confidence level for a classification has been outlined. It is contended that classifications with ascribed confidence levels are to be preferred. However, the derivation technique for these confidence levels should also be indicated and combined with the classification products.

Some of the basic factors that should be considered during field checking of a classification have been discussed.

The statistical background to the derivation of a confidence level based on the Normal Distribution has been presented and illustrated by reference to an actual New Zealand classification exercise using Landsat data processed by means of the IBM ERMAN package.

While this has been developed to support a New Zealand series of projects utilizing a maximum likelihood classifier, we believe the methods to be of general use irrespective of project area, type, or classifier.

Acknowledgments

It is a pleasure to acknowledge helpful discussions on this topic with Miss V. M. Benning (Department of Lands and Survey), Mr. N. P. Ching (New Zealand Forest Service), and Drs. L. J. Fradkin and P. J. Ellis (Physics and Engineering Laboratory, DSIR). Thanks are also due to Miss V. M. Benning for provision of the illustrative data and to IBM New Zealand for their support of the classification aspects of the study through a Joint Research Program Agreement. This work was completed under a United States National Research Council Associateship Program within the Climate and Earth Sciences Laboratory of NOAA. Assistance in preparing this manuscript from Mesdames S. J. Coburn, H. J. McClure of DSIR, and Mrs. O. L. Smith of NOAA/NESDIS is greatly appreciated.

References

Aitken, A. C., 1942. *Statistical Mathematics:* Oliver & Boyd, Edinburgh, U.K.

Anderson, J. R., E. E. Hardy, and J. T. Roach, 1972. *A land use classification system for use with remote sensor data:* U.S. Geological Survey Circular 671.

Benning, V. M., 1982. Land use/cover mapping from Landsat, Ch. 14 of *Computer classification of Landsat and aircraft scanner images—The collected papers of the ERMAN project*, Ed. I. L. Thomas, Physics and Engineering Laboratory Report No. 766, October 1982, DSIR, New Zealand.

IBM, 1976. *Earth Resources—Management II (ERMAN II) User's Guide:* Program No. 5790-ARB, Doc. No. SB11-5008-0, IBM, Brussels, Belgium.

Mead, R. A., and J. Szajgin, 1982. Landsat classification accuracy assessment procedures, *Photogrammetric Engineering and Remote Sensing*, Vol. 48, No. 1, pp. 139-141.

Mood, A. McF., 1950. *Introduction to the Theory of Statistics:* McGraw-Hill, New York, U.S.A.

Moroney, M. J., 1956. *Facts from Figures:* Penguin Books, Middlesex, England.

Rosenfield, G. H., K. Fitzpatrick-Lins, and H. S. Ling, 1982. Sampling for thematic map accuracy testing, *Photogrammetric Engineering and Remote Sensing*, Vol. 48, No. 1, pp. 131-137.

Swain, P. H., and S. M. Davis, 1978. *Remote Sensing: The Quantitative Approach:* McGraw-Hill Book Company, New York, U.S.A.

Unthank, E. L., 1960. *Statistics for Matriculation Mathematics:* Halls, Melbourne, Australia.

Van Genderen, J. L., B. F. Lock, and P. A. Vass, 1978. Remote Sensing: statistical testing of thematic map accuracy, *Remote Sensing of Environment*, Vol. 7, pp. 3-14.

(Received 3 January 1983; revised and accepted 12 February 1984)